Zhenyu **Li**

Postgraduate in HIT · CV Researcher

92 Xidazhi Street, Nangang District, Harbin City, Heilongjiang, 150001, China

(+86) 188-0041-9432 | vgumypxt@gmail.com | https://zhyever.github.io/homepage | https://github.com/zhyever

"I am currently looking for a worldwide Ph.D. position with a full scholarship."

Education

Harbin Institute of Technology

Harbin, China

M.S. in Computer Science and Technology

Sep. 2021 - Present

- Got funding for four accepted papers

Harbin Institute of Technology

Harbin, China

B.S. in Computer Science and Technology (89.58/100, top 25%)

Sep. 2017 - Jul. 2021

- Got exemption from examination and scholarship for postgraduate study in Harbin Institute of Technology

Internship Experience

SenseTime Research

Shanghai, China

Research Intern

Jan. 2022 - July. 2022

- Researched unsupervised domain adaptation algorithms for monocular 3D object detection. Model performance surpasses the Oracle on the target domain. One paper as the first author is accepted of ECCV 2022.
- Deployed the aforementioned unsupervised domain adaptation algorithm in industrial project with GAC Group. Achieved excepted goals.
- Researched multi-view monocular 3D object detection algorithms. One paper as the second author is accepted ACM MM 2022.
- Researched domain generalization and unsupervised domain adaptation algorithms for monocular 3D object detection. One paper as the first author is under review

SenseTime Research

Shanghai, China

Perception Algorithm Development Intern

Mar. 2021 - Sep. 2021

- Built up a ReID dataset based on the ground-truth system, utilized the Fast-ReID framework to train a ReID model, and developed the ReID model to the ADAS system.
- Built and deployed the DeepSort multi-object tracking algorithm in the ADAS system (C++), including importing appearance representation from the ReID model and adopting the cascade association strategy. The algorithm formed a patent for SenseTime.
- Researched multi-object tracking algorithms.
- Researched multi-modal contrastive learning algorithms for spatial-aware visual representations to benefit 3D-related downstream tasks. Supervised by Ang Li, Hongyang Li, Bolei Zhou, and Hang Zhao, one paper as the first author was accepted to AAAI 2022.
- Researched multi-modal 3D object detection algorithms. One paper as the second author was accepted to IJCAI 2022. Another paper as the second author is accepted ECCV 2022.

Research Experience

Monocular Depth Estimation Toolbox

Major contributor, Codebase, 250+ stars

- Monocular-Depth-Estimation-Toolbox is an open source monocular depth estimation toolbox based on PyTorch and mmSegmentation v0.16.0.
- Project website: `https://github.com/zhyever/Monocular-Depth-Estimation-Toolbox`.

Unsupervised Domain Adaptation for Monocular 3D Object Detection via Self-Training

First Author, ECCV 2022

- Monocular 3D object detection (Mono3D) has achieved unprecedented success with the advent of deep learning techniques and emerging large-scale autonomous driving datasets. However, drastic performance degradation remains an unwell-studied challenge for practical cross-domain deployment as the lack of labels on the target domain. In this paper, we first comprehensively investigate the significant underlying factor of the domain gap in Mono3D, where the critical observation is a depth-shift issue caused by the geometric misalignment of domains. Then, we propose STMono3D, a new self-teaching framework for unsupervised domain adaptation on Mono3D. To mitigate the depth-shift, we introduce the geometry-aligned multi-scale training strategy to disentangle the camera parameters and guarantee the geometry consistency of domains. Based on this, we develop a teacher-student paradigm to generate adaptive pseudo labels on the target domain. Benefiting from the end-to-end framework that provides richer information of the pseudo labels, we propose the quality-aware supervision strategy to take instancelevel pseudo confidences into account and improve the effectiveness of the target-domain training process. Moreover, the positive focusing training strategy and dynamic threshold are proposed to handle tremendous FN and FP pseudo samples. STMono3D achieves remarkable performance on all evaluated datasets and even surpasses fully supervised results on the KITTI 3D object detection dataset. To the best of our knowledge, this is the first study to explore effective UDA methods for Mono3D.

SimIPU: Simple 2D Image and 3D Point Cloud Unsupervised Pre-Training for Spatial-Aware Visual Representations

First Author, AAAI 2022

- Pre-training has become a standard paradigm in many computer vision tasks. However, most of the methods are generally designed on the RGB image domain. Due to the discrepancy between the two-dimensional image plane and the three-dimensional space, such pre-trained models fail to perceive spatial information and serve as sub-optimal solutions for 3D-related tasks. To bridge this gap, we aim to learn a spatial-aware visual representation that can describe the three-dimensional space and is more suitable and effective for these tasks. To leverage point clouds, which are much more superior in providing spatial information compared to images, we propose a simple yet effective 2D Image and 3D Point cloud Unsupervised pre-training strategy, called SimIPU. Specifically, we develop a multi-modal contrastive learning framework that consists of an intra-modal spatial perception module to learn a spatial-aware representation from point clouds and an inter-modal feature interaction module to transfer the capability of perceiving spatial information from the point cloud encoder to the image encoder, respectively. Positive pairs for contrastive losses are established by the matching algorithm and the projection matrix. The whole framework is trained in an unsupervised end-to-end fashion. To the best of our knowledge, this is the first study to explore contrastive learning pre-training strategies for outdoor multi-modal datasets, containing paired camera images and LIDAR point clouds.

Towards Model Generalization for Monocular 3D Object Detection

First Author, Arxiv

- Monocular 3D object detection (Mono3D) has achieved tremendous improvements with emerging large-scale autonomous driving datasets and the rapid development of deep learning techniques. However, caused by severe domain gaps (e.g., the field of view (FOV), pixel size, and object size among datasets), Mono3D detectors have difficulty in generalization, leading to drastic performance degradation on unseen domains. To solve these issues, we combine the position-invariant transform and multi-scale training with the pixel-size depth strategy to construct an effective unified camera-generalized paradigm (CGP). It fully considers discrepancies in the FOV and pixel size of images captured by different cameras. Moreover, we further investigate the obstacle in quantitative metrics when cross-dataset inference through an exhaustive systematic study. We discern that the size bias of prediction leads to a colossal failure. Hence, we propose the 2D-3D geometry-consistent object scaling strategy (GCOS) to bridge the gap via an instance-level augment. Our method called DGMono3D achieves remarkable performance on all evaluated datasets and surpasses the SoTA unsupervised domain adaptation scheme even without utilizing data on the target domain.

BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation

First Author, Arxiv

- Monocular depth estimation is a fundamental task in computer vision and has drawn increasing attention. Recently, some methods reformulate it as a classification-regression task to boost the model performance, where continuous depth is estimated via a linear combination of predicted probability distributions and discrete bins. In this paper, we present a novel framework called BinsFormer, tailored for the classification-regression-based depth estimation. It mainly focuses on two crucial components in the specific task: 1) proper generation of adaptive bins and 2) sufficient interaction between probability distribution and bins predictions. To specify, we employ the Transformer decoder to generate bins, novelly viewing it as a direct set-to-set prediction problem. We further integrate a multi-scale decoder structure to achieve a comprehensive understanding of spatial geometry information and estimate depth maps in a coarse-to-fine manner. Moreover, an extra scene understanding query is proposed to improve the estimation accuracy, which turns out that models can implicitly learn useful information from an auxiliary environment classification task. Extensive experiments on the KITTI, NYU, and SUN RGB-D datasets demonstrate that BinsFormer surpasses state-of-the-art monocular depth estimation methods with prominent margins.

DepthFormer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation

First Author, Arxiv

- This paper aims to address the problem of supervised monocular depth estimation. We start with a meticulous pilot study to demonstrate that the long-range correlation is essential for accurate depth estimation. Therefore, we propose to leverage the Transformer to model this global context with an effective attention mechanism. We also adopt an additional convolution branch to preserve the local information as the Transformer lacks the spatial inductive bias in modeling such contents. However, independent branches lead to a shortage of connections between features. To bridge this gap, we design a hierarchical aggregation and heterogeneous interaction module to enhance the Transformer features via element-wise interaction and model the affinity between the Transformer and the CNN features in a set-to-set translation manner. Due to the unbearable memory cost caused by global attention on high-resolution feature maps, we introduce the deformable scheme to reduce the complexity. Extensive experiments on the KITTI, NYU, and SUN RGB-D datasets demonstrate that our proposed model, termed DepthFormer, surpasses state-of-the-art monocular depth estimation methods with prominent margins. Notably, it achieves the most competitive result on the highly competitive KITTI depth estimation benchmark.

Self-supervised Monocular Depth Estimation via Discrete Strategy and Uncertainty

First Author, Letter for IEEE/CAA Journal of Automatica Sinica

- This letter is concerned with self-supervised monocular depth estimation. To estimate uncertainty simultaneously, we propose a simple yet effective strategy to learn the uncertainty for self-supervised monocular depth estimation with the discrete strategy that explicitly associates the prediction and the uncertainty to train the networks. Furthermore, we propose the uncertainty-guided feature fusion module to fully utilize the uncertainty information.

Deformable Feature Aggregation for Dynamic Multi-Modal 3D Object Detection

Second Author, ECCV 2022

- Point clouds and RGB images are two general perceptional sources in autonomous driving. Each has its own strength: point clouds can provide accurate localization of objects, and images are more dense and rich in semantic information. Recently, AutoAlign presents a learnable paradigm in fusing these two modalities for 3D object detection. However, it suffers from the large computational cost introduced by global-wise attention. In this work, we propose deformable crossattention feature aggregation module. It attends to sparse learnable sampling points for cross-modal relational modeling, which greatly speeds up the multi-level feature aggregation and enhances the tolerance to calibration error. By carrying out a novel image-level dropout training scheme, our model is able to infer in a dynamic manner. Moreover, complex GT-AUG under multi-modal setting is also a barrier in achieving high performance detectors. To overcome this issue, we present a simple yet effective cross-modal augmentation strategy on convex combination of image patches given their depth information. To this end, we propose AutoAlignV2, a faster and stronger multi-modal 3D detection framework, built on top of AutoAlign. Extensive experiments on nuScenes benchmark demonstrate the effectiveness and efficiency of AutoAlignV2. Notably, our best model reaches 72.4 NDS on nuScenes test leaderboard, achieving new state-of-the-art results among all published multi-modal 3D object detectors.

Graph-DETR3D: Rethinking Overlapping Regions for Multi-View 3D Object Detection

Second Author, ACMM MM 2022

- 3D object detection from multiple image views is a fundamental and challenging task for visual scene understanding. Due to its low cost and high efficiency, multi-view 3D object detection has demonstrated promising application prospects. However, accurately detecting objects through perspective views in the 3D space is extremely difficult due to the lack of depth information. Recently, DETR3D introduces a novel 3D-2D query paradigm in aggregating multi-view images for 3D object detection and achieves state-of-the-art performance. In this paper, with intensive pilot experiments, we quantify the objects located at different regions and find that the "truncated instances" (i.e., at the border regions of each image) are the main bottleneck hindering the performance of DETR3D. Although it merges multiple features from two adjacent views in the overlapping regions, DETR3D still suffers from insufficient feature aggregation, thus missing the chance to fully boost the detection performance. In an effort to tackle the problem, we propose Graph-DETR3D to automatically aggregate multi-view imagery information through graph structure learning (GSL). It constructs a dynamic 3D graph between each object query and 2D feature maps to enhance the object representations, especially at the border regions. Besides, Graph-DETR3D benefits from a novel depth-invariant multi-scale training strategy, which maintains the visual depth consistency by simultaneously scaling the image size and the object depth. Extensive experiments on the nuScenes dataset demonstrate the effectiveness and efficiency of our Graph-DETR3D. Notably, our best model achieves 49.5 NDS on the nuScenes test leaderboard, achieving new state-of-the-art in comparison with various published image-view 3D object detectors.

AutoAlign: Pixel-Instance Feature Aggregation for Multi-Modal 3D Object Detection

Second Author, IJCAI 2022

- Object detection through either RGB images or the LiDAR point clouds has been extensively explored in autonomous driving. However, it remains challenging to make these two data sources complementary and beneficial to each other. In this paper, we propose AutoAlign, an automatic feature fusion strategy for 3D object detection. Instead of establishing deterministic correspondence with camera projection matrix, we model the mapping relationship between the image and point clouds with a learnable alignment map. This map enables our model to automate the alignment of non-homogenous features in a dynamic and data-driven manner. Specifically, a cross-attention feature alignment module is devised to adaptively aggregate pixel-level image features for each voxel. To enhance the semantic consistency during feature alignment, we also design a self-supervised cross-modal feature interaction module, through which the model can learn feature aggregation with instance-level feature guidance. Extensive experimental results show that our approach can lead to 2.3 mAP and 7.0 mAP improvements on the KITTI and nuScenes datasets, respectively. Notably, our best model reaches 70.9 NDS on the nuScenes testing leaderboard, achieving competitive performance among various state-of-the-arts.

Deep Learning based Monocular Depth Estimation: A Survey

Second Author, Chinese Journal of Computers

- This paper provides an overview of the-state-of-the-art on monocular depth estimation methods. First, the definition of monocular depth estimation, standard datasets used in training, evaluation metrics, and applications are introduced. Then, we review some representative methods according to different training manners: supervised, unsupervised and semi-super-vised. The existing methods based on different learning manners are divided into several types. We summarize supervised methods and classify them into framework enhancement, introducing auxiliary information, improving loss function, classification-based methods, methods applying conditional random field, methods applying genera-tive adversarial network, and methods based on partial depth labels. As for unsupervised methods, we classify them into models trained by image pairs and monocular videos. Improving methods are divided into mask-based methods, applying visual odometry, methods applying generative adversarial network, and methods forward fast and light runtime performance. In terms of semi-supervised methods, they are similarly trained on image pairs or monocular videos. The proposed methods are classified into methods that apply a generative adversarial network and introduce semantic information. The ideas, advantages, and disadvantages of each type of method are analyzed in detail. Finally, we sort out future development trends and key technologies of monocular depth estimation methods based on deep learning.

Presentation

IEEE/CAA JAS Symposium Series II, Intelligent Visual Perception in Smart City

Virtual

Presenter for our recent work

Apr. 2012

- Introduced our work: "Enhancing Self-supervised Monocular Depth Estimation via Discrete Disparity and Uncertainty", "SimIPU: Simple 2D Image and 3D Point Cloud Unsupervised Pre-Training for Spatial-Aware Visual Representations", "DepthFormer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation", and "BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation".

Honors & Awards

International

Mar. 2022	**1th Place**, The KITTI Vision Benchmark Suite: Monocular Depth Estimation (BinsFormer)	*Online*
Sep. 2021	**1th Place**, The KITTI Vision Benchmark Suite: Monocular Depth Estimation (DepthFormer)	*Online*
Mar. 2021	**6th Place**, Mobile AI 2021 Monocular Depth Estimation Challenge (CVPR2021 Workshop)	*Online*

Domestic

2022	**FinalList**, Chunhui Scholarship for Harbin Institute of Technology	*China*
2021	**First-Class**, Scholarship for Postgraduate Students	*China*
2020	**Second-Class**, Scholarship for Undergraduate Students	*China*
2019	**Second-Class**, Scholarship for Undergraduate Students	*China*
2018	**Third-Class**, Scholarship for Undergraduate Students	*China*

Skills

Interest-Area	Scene Perception and Understanding, 3D Reconstruction, Multi-Modal Learning
Research-Topic	Monocular Depth Estimation, Monocular 3D Object Detection, Multi-Modal Pre-training for ADAS
Framework	mmDetection3D, mmSegmentation, Depth-Estimation-Toolbox
Languages	Chinese, English (CET-4, CET-6, TOFEL)
Programming	Python (PyTorch), JAVA, C, C++
Daily-Hobbies	Basketball, Breaking Dance, Cooking, Working Out