Research of 3D Perception Algorithm based on Multi-sensor Fusion

Application in Target Tracking Tasks for Unmanned Surface Vehicles (USVs)

Yiheng XUE

Student ID: [redacted]

School of Advanced Technology

Xi'an Jiaotong-Liverpool University

Huzhou Institute of Zhejiang University

December 13, 2023

Table of Contents

Introduction

USV Tracker: A Robust Tracking System Based on Multi-Sensor Fusion and Elastic Planning

Developed a simulator for algorithm optimization and a multi-sensor perception platform for Unmanned Surface Vehicles (USVs). This system enables the estimation of target 3D coordinates and trajectory prediction independently of target communication. Integrated the predicted trajectory with planning algorithms for comprehensive validation in both simulated and real-world environments.

1. Implemented YOLO for efficient **target detection** and sequential tracking.
2. Developed sparse grid-map formatted **obstacle mapping** in dynamic environments.
3. Employed **3D reconstruction** for image dataset creation.
4. Utilized Extended Kalman Filter with linear classifiers for **trajectory prediction**, achieving accuracy comparable to LSTM network methods.

Key Considerations

Edge computing, Modularity, Deep learning limitations, Graphical efficiency (simulated), Sensor customization benefits (physical)

Problem Statement

General Question

This study delves into the distinct advantages of various sensors in 3D perception tasks, investigating how their integration can lead to more stable and accurate perception systems across a broad range of applications.

1. **Fusing Varied-Frequency Sensor Data**: Effective methods in motion scenarios.
2. **Image Detection in Complex Environments**: Advantages over point clouds for dynamic, obstacle-rich settings.
3. **Stable Tracking in Intense Motion**: Ensuring consistent multi-sensor target tracking.
4. **Perception-Decision Connectivity**: Establishing robust links between perception and decision-making in robotics.

Figure: Perception Sensors from 1-D (Left) to 3-D (Right)

Problem Statement

Specific Question

The study aims to realize stable and cost-effective 3D target localization and trajectory prediction for USV target tracking on mobile platforms, ensuring consistent target pursuit with effective obstacle avoidance.

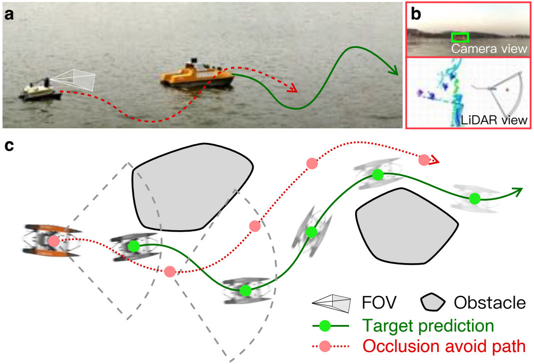


Figure: USV Tracker Sketch in Obstacle Map

Related Work

3D Perception Tasks/Datasets

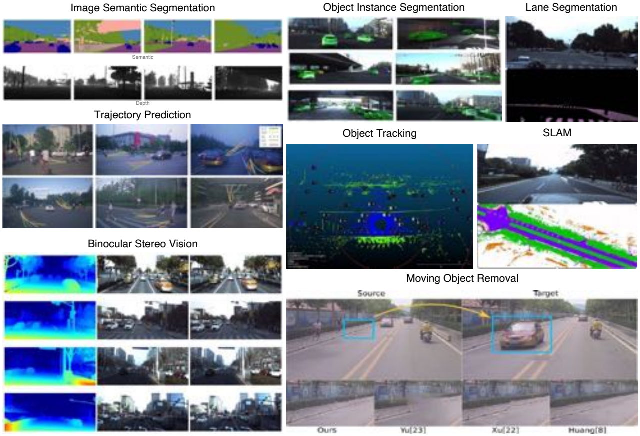


Figure: Examples of 3D Perception Tasks/Datasets

Top row: Semantic segmentation images from Waymo[1]. Middle row: Trajectory prediction using [2]. Bottom row, right: Moving object removal by [3]. All other tasks and data derived from sub-tasks in the Apollo dataset [4].

Potential applications for USV tracking:

Semantic Segmentation: Elimination of lens stains and segmentation of bright reflections on water surfaces.

Trajectory Prediction: Prediction of target trajectories during tracking.

Stereo Vision: Estimation of target depth.

SLAM: Construction of obstacle maps using sparse features.

Object Removal: Building obstacle maps without interference from targets.

Related Work

Perception Algorithms and Common Sensors

2D Image Algorithms

Detection: Faster-RCNN[5], Yolo[6], DETR[7]
Segmentation: U-Net[8], DeepLab[9]
Tracking: SiamFC[10], MDNet[11], DeepSORT[12]
...

3D Point Cloud Algorithms

Detection: VoteNet[13], PV-RCNN[14], VoxelNet[15]
Segmentation: PointNet[16], PointNet++[17], 3D U-Net[18]
Tracking: PointTrackNet[19], 3D-siamRPN[20]
...

Multi-sensor Fusion

Detection: MV3D[21], AVOD[22]
Segmentatio: FusionSeg[23], PointFusion[24]
BEV: LSS[25], BEVFormer[26], BEVFusion[27]
...

Sensor	Dimension	Freqency	Accuracy Level	Spatial Information
IMU	1D, 6/9DoF	100-1000Hz	B, A w/ loop	Full-body motion
GPS	1D, 3D pos	1-10Hz	B, A w/ DGPS	Outdoor position
Bio-sensor	1D, Pressure	0.02-400Hz	A+	3D contact pressure
Camera	2D	10-150Hz	A	Dense 2D pixels
LiDAR	3D	$\sim$10Hz rotary	B	Sparse 3D point cloud

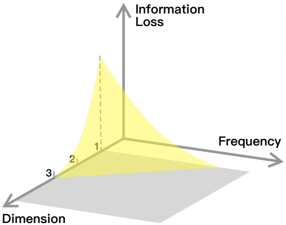
Table: Sensor Comparison

Multi-Sensor Selection Criteria

In an unstructured dynamic motion environment, the selection of an appropriate sensor combination is crucial. GPS and IMU are commonly employed to furnish motion correction parameters for perception sensors.

Hypothesis - Information Conservation

Higher-dimensional sensors operate at lower frequencies, leading to more information loss in dynamic environments.



1. 1-D sensors are precise but limited in capturing 3D details.
2. 2-D sensors like cameras provide richer features suitable for dense 3D reconstruction.
3. 3-D sensors, such as LiDAR, offer spatial accuracy but are less effective in dynamic settings due to lower frequencies.

This highlights a trade-off in sensor system design for dynamic robotic applications.

Experiment Statement

Combination I - Signal Classification

1D signal + IMU

dynamic motion, **real-time** feedback

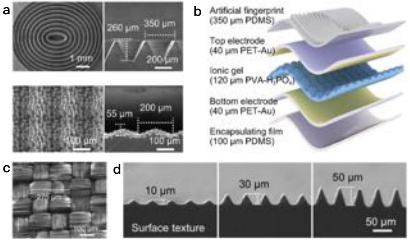


Figure: High-frequency Sensor

Combination II - 3D Reconstruction

2D image + IMU + GPS

rich features, **limited-view**

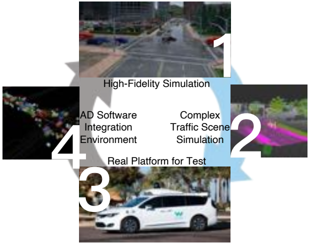


Figure: Sim-real-platform Pipeline

Challenges in USV Target Tracking

Spare vision **features**, **dynamic** water surface, complex control, **real-time** processing, **limited-view**, environmental variability, multi-sensor integration, precision navigation, ...

Experiment Setup

Goal: design an effective multi-sensor platform for USV in dynamic environment.

Combination I - Signal Classification

1D signal + IMU

1. Low dimensional high frequency sensor can obtain the 3D structure information.
2. Dynamic motion posture correction achieved with 100Hz IMU data.

Combination II - 3D Reconstruction

2D image + IMU + GPS

1. 2D image target detection stable, algorithmic depth prediction in 3D unreliable.
2. Stable detection achievable in dynamic motion with a 30fps camera.

Multi-Sensor Combination for USV Target Tracking System

Spare vision **features**, **dynamic** water surface, complex control, **real-time** processing, **limited-view**, environmental variability, multi-sensor integration, precision navigation, ...

3D point cloud + 2D image + IMU + GPS

USV Tracker: 3-D Multi-sensor Fusion for Target Tracking

LiDAR (10Hz) + Camera (30Hz) + IMU (100Hz) + GPS (10Hz)

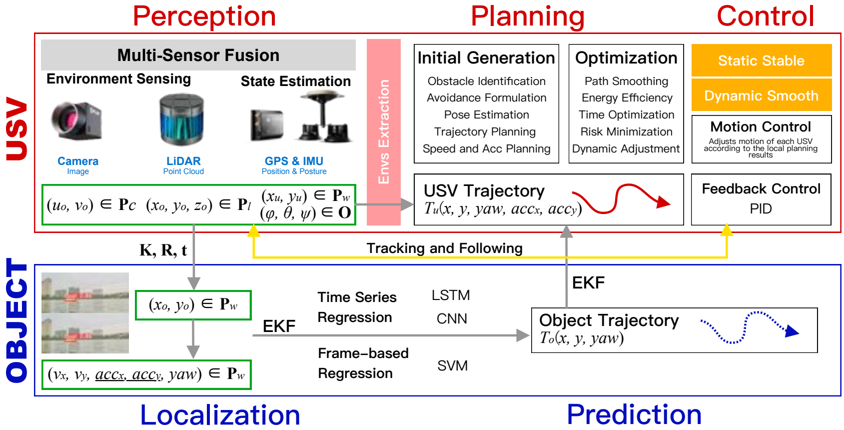


Figure: Diagram of the USV Target Tracking System

Coordinate Transformation Formulas

Computing the Transformation Matrix in Simulator and Calibrating with Checkerboard in Physical Experiments

$$f = \frac{I_w}{2\tan\left(\frac{FOV}{2}\right)} \quad (1)$$

$$R = I + \sin(\theta)K + (1 - \cos(\theta))K^2 \quad (2)$$

$$\mathbf{p}_{\mathrm{norm}} = K^{-1}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (3)$$

$$\mathbf{p}_{\mathrm{usv}} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}\mathbf{p}_{\mathrm{norm}} \quad (4)$$

$$\mathbf{p}_{\mathrm{point\_cloud}} = \mathrm{SE}(3)\mathbf{p}_{\mathrm{usv}} \quad (5)$$

$$\mathbf{p}_{\mathrm{target\_global}} = R_\theta \mathbf{p}_{\mathrm{usv}} + \mathbf{p}_{\mathrm{usv\_global}} \quad (6)$$

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (7)$$

- f: Camera focal length.
- I_w: Width of the image.
- FOV: Field of view of the camera, in radians.
- R: Rotation matrix, computed using Rodrigues' rotation formula in simulator.
- θ: Magnitude of the rotation vector.
- K: Skew-symmetric matrix derived from the unit rotation vector.
- $\mathbf{p}_{\mathrm{norm}}$: Normalized coordinates of image points in 3D space.
- $\mathbf{p}_{\mathrm{usv}}$: Coordinates transformed to the USV.
- $\mathbf{p}_{\mathrm{point\_cloud}}$: USV coordinates mapped to the LiDAR space.
- $\mathbf{p}_{\mathrm{target\_global}}$: Target's global coordinates, computed using the observer's orientation and global position of the USV.
- R_θ: Rotation matrix representing the observer's orientation.

USV Platform in Simulation

LiDAR (10Hz) + Camera (30Hz) + IMU (100Hz) + GPS (10Hz)

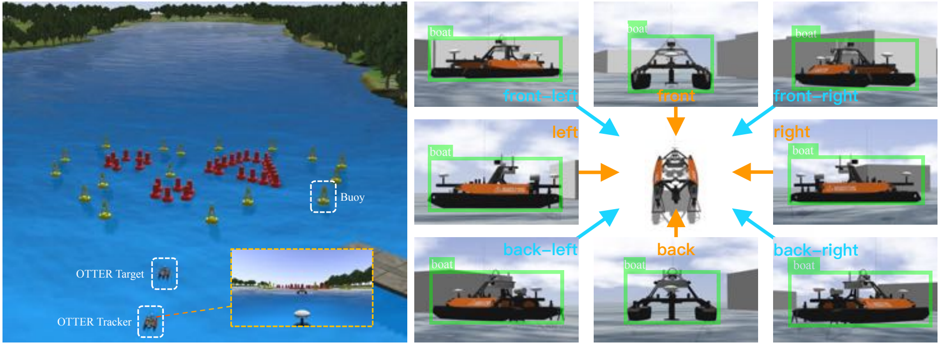


Figure: USV Simulator in Gazebo Featuring Obstacle Targets and Camera Captured Images (Left), and Image-Based Orientation Prediction Through 8 Viewpoints (Right)

- **LiDAR**: 32-beam, FOV $360^\circ \times -15^\circ \times 15^\circ$, 160k pts/s
- **Camera**: 1241×376, 30fps, FOV $80^\circ \times 60^\circ$

USV Platform in Real-world

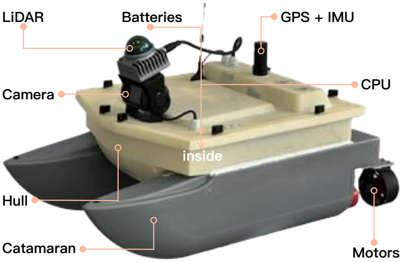


Figure: USV Hardware and Sensor Layout (Left), BEV of Huzhou Experiment Site, Zhejiang (Right)

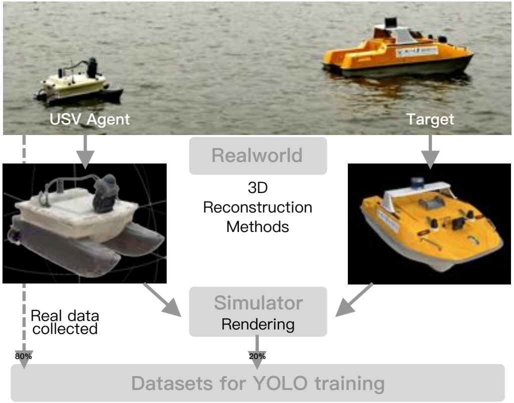
- **LiDAR**: Livox-mid360 (waterproof), FOV $360^\circ \times -7^\circ \times 52^\circ$, 200k pts/s
- **Camera**: USB camera (waterproof), 1920×1080, 30fps, FOV $80^\circ \times 60^\circ$
- **CPU**: Intel i7-1165G7@4.7GHz
- **Motors**: 2 brushless motors, 180W
- **GPS**: Ublox-zedf9p, rtk
- **IMU**: Witmotion-hwt905
- **Controller**: PX4
- **Max Speed**: 2.7m/s
- **Weight**: $\sim$ 5kg
- **Duration**: $\sim$ 35mins
- **Distance**: $\leq$ 500m
- **Tracking Range**: $\sim$ 7m

USV Platform in Real-world

Enhanced YOLO Dataset via 3D Reconstruction Techniques

1. 3D visual reconstruction of targets
2. Integration of 2 models into simulator
3. Generation of an expanded dataset through simulation
4. Enhanced training for YOLO object detection using this enriched dataset

Implemented 3D visual reconstruction in practical experiments for modeling targets and USVs. Integrated these models into a simulator to create an **enriched dataset**. This blend of simulated and real-world data **enhances training efficiency** for YOLO object detection.



3D Reconstruction Results

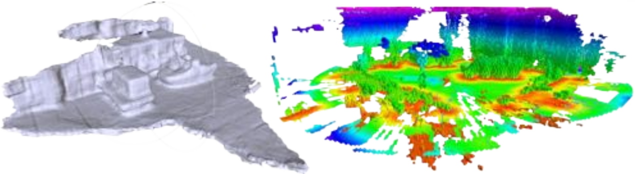


Figure: 3D Object Modeling via **3D Point Clouds + IMU** and Obstacle Map Construction

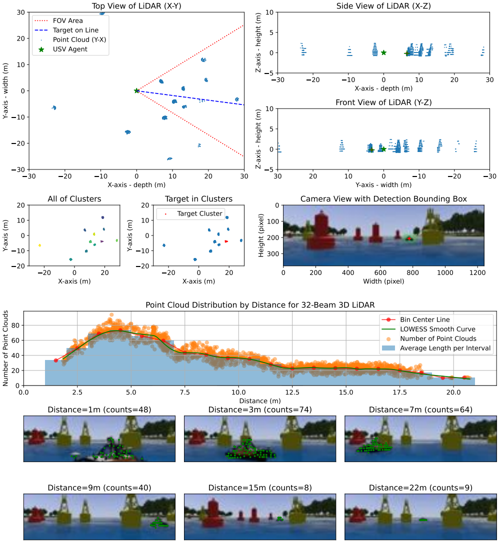
Sparse point clouds meet obstacle mapping accuracy needs, but are inadequate for dense 3D feature capture due to resolution limits above 0.2m in voxel or grid maps. Precise 3D reconstruction requires **2D images + IMU** integration for detailed feature capture.

Figure: Video Results

Figure: Video Results

Experiment Results

3D Localization via Visual Target Detection and Point Cloud Clustering in BEV Space



Pipeline of 3D Target Localization

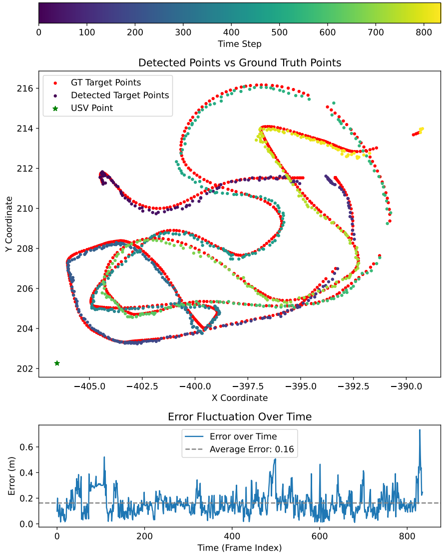
1. Object detection via images
2. Camera-LiDAR coordinate transformation
3. Motion distortion elimination in point clouds
4. Clustering of point clouds
5. Matching and labeling clusters with image detections

Sparse Point Clouds vs. Dense Pixels

In the simulation, a 32-channel LiDAR, akin to Livox, with a 30-degree vertical FOV and 10Hz frequency is utilized. Increased target distance leads to fewer point cloud reflections and shape recognition challenges. Beyond 7 meters, distinguishing targets from cylindrical obstacles becomes difficult, establishing **7m as the optimal tracking distance** within hardware constraints.

Experiment Results

Continuous Target Localization and Global Position Estimation in Obstacle-Rich Spaces



Stability and Accuracy

Sustains continuous monitoring with an average error of 16 cm, well below the 2m boat length.

Robustness

Operates effectively in dense obstacle settings and maintains detection on turbulent water surfaces.

Efficiency

Performs target localization and trajectory prediction on CPU with over 5 fps, optimizing computational resources.

Experiment Results

Trajectory Prediction

Target direction prediction is segmented into eight angles, achieving over 90% accuracy, highlighting the importance of directional accuracy in planning and tracking tasks.

Accurate Nonlinear Trajectory Prediction

1. USV yaw aligns with trajectory tangent, foundational knowledge.
2. Predictive planning in obstructed areas improves robustness.
3. Target behavior prediction crucial for planning, providing key input.
4. Due to limited accuracy in image classification and real-world challenges, combined with finite computational resources, the method pivots from image-based 8-direction prediction to trajectory analysis, utilizing prior knowledge for simpler, trajectory-focused orientation prediction.

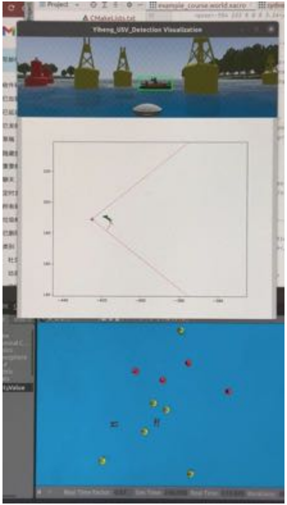
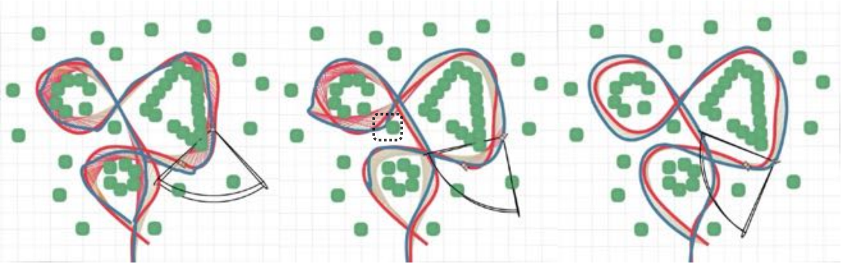


Figure: Video Results

Experiment Results

Perception and Planning: Collaborative Integration for Precise Target Localization and Direction Prediction with Focus on Maintaining Target within Visual Field

Our approach integrates precise perception and strategic planning, excelling in maintaining the target at the center of the FOV. This synergy allows for effective tracking of predicted trajectories, especially in intricate environments, facilitating uninterrupted pursuit even when the target is temporarily lost.



(a) baseline (b) elastic tracker (c) our method

Target USV's Trace Tracking USV's Trace Buoy

Figure: Comparative Experimental Results in Complex Aquatic Environment

Red lines between trajectories signify target occlusion, while the black dashed box highlights the collision site of the USV with an obstacle.

Target Tracking Demo in Real-World

Figure: Video Results

Discussion and Future Work

This experiment employed a hybrid approach combining simulation with physical trials, rapidly validating the feasibility of algorithms and engineering code in a simulated environment. The tasks were divided into perception and planning, each independently executable in the simulation. Limitations included the vessel's unstable dynamics; its small size and unstable center of gravity and buoyancy significantly affected perception during extensive movements, occasionally causing the target to exceed the Field of View's (FOV) vertical edges.

Future Work

1. Redesign the vessel to increase ballast and enhance pitch stability.
2. Add hardware sensors to detect water flow, allowing feedback adjustment of the control unit to stabilize the course angle.
3. Encapsulate the algorithm in an end-to-end neural network using PyTorch, streamlining the system and exploring reinforcement learning approaches.

Conclusions

Hypothesis - Information Conservation

Higher-dimensional sensors operate at lower frequencies, leading to more information loss in dynamic environments.

- **Enhanced stability through multi-sensor fusion**: In target tracking, image-based solutions more effectively handle environmental interferences due to their rich information content, allowing for successful target detection even amidst disturbances. Despite their sparsity, point clouds provide precise 3D depth information, **necessitating multi-sensor fusion for a stable system**.
- **Sensor selection for specific tasks**: For 3D modeling, dense feature sensors like **cameras are essential for detail**, while depth-reliable 3D sensors, despite feature sparsity, are **crucial for accurate spatial** obstacle localization in planning tasks.

References I

[1] Pei Sun et al. "Scalability in Perception for Autonomous Driving: Waymo Open Dataset". In: **arXiv preprint arXiv:1912.04838** (2019).

[2] Yuexin Ma et al. "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents". In: **Proceedings of the AAAI Conference on Artificial Intelligence**.

[3] Miao Liao et al. "DVI: Depth Guided Video Inpainting for Autonomous Driving". In: **arXiv preprint arXiv:2007.08854** (2020).

[4] Xinyu Huang et al. "The ApolloScape Open Dataset for Autonomous Driving and its Application". In: **arXiv preprint arXiv:1803.06184** (2018).

[5] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: **Advances in neural information processing systems** 28 (2015).

[6] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016, pp. 779–788.

[7] Nicolas Carion et al. "End-to-end object detection with transformers". In: **European conference on computer vision**. Springer. 2020, pp. 213–229.

References II

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: (2015), pp. 234–241.

[9] Liang-Chieh Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: **IEEE transactions on pattern analysis and machine intelligence** 40.4 (2017), pp. 834–848.

[10] Luca Bertinetto et al. “Fully-Convolutional Siamese Networks for Object Tracking”. In: **arXiv preprint arXiv:1606.09549** (2016).

[11] Hyeonseob Nam and Bohyung Han. “Learning Multi-Domain Convolutional Neural Networks for Visual Tracking”. In: **arXiv preprint arXiv:1510.07945** (2015).

[12] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. “Simple Online and Realtime Tracking with a Deep Association Metric”. In: **arXiv preprint arXiv:1703.07402** (2017).

[13] Charles R Qi et al. “Deep hough voting for 3d object detection in point clouds”. In: **proceedings of the IEEE/CVF International Conference on Computer Vision**. 2019, pp. 9277–9286.

[14] Shaoshuai Shi et al. “Pv-rcnn: Point-voxel feature set abstraction for 3d object detection”. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. 2020, pp. 10529–10538.

References III

[15] Yin Zhou and Oncel Tuzel. “Voxelnet: End-to-end learning for point cloud based 3d object detection”. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2018, pp. 4490–4499.

[16] Charles R Qi et al. “Pointnet: Deep learning on point sets for 3d classification and segmentation”. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2017, pp. 652–660.

[17] Charles Ruizhongtai Qi et al. “Pointnet++: Deep hierarchical feature learning on point sets in a metric space”. In: **Advances in neural information processing systems** 30 (2017).

[18] Özgün Çiçek et al. “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation”. In: (2016), pp. 424–432.

[19] Sukai Wang, Yuxiang Sun, and Chengju et al. Liu. “PointTrackNet: An End-to-End Network for 3-D Object Detection and Tracking from Point Clouds”. In: **IEEE Robotics and Automation Letters** 5.2 (2020), pp. 3206–3212.

[20] Zheng Fang et al. “3D-SiamRPN: An End-to-End Learning Method for Real-Time 3D Single Object Tracking Using Raw Point Cloud”. In: **arXiv preprint arXiv:2108.05630** (2021).

References IV

[21] Xiaozhi Chen et al. "Multi-view 3d object detection network for autonomous driving". In: **Proceedings of the IEEE conference on Computer Vision and Pattern Recognition**. 2017, pp. 1907–1915.

[22] Jason Ku et al. "Joint 3d proposal generation and object detection from view aggregation". In: **2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**. IEEE. 2018, pp. 1–8.

[23] Suyog Jain, Bo Xiong, and Kristen Grauman. "FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos". In: **CVPR** (2017).

[24] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. "PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation". In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. 2018, pp. 244–253.

[25] Jonah Philion and Sanja Fidler. "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d". In: **Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16**. Springer. 2020, pp. 194–210.

References V

[26] Zhiqi Li et al. "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers". In: **European conference on computer vision**. Springer. 2022, pp. 1–18.

[27] Zhijian Liu et al. "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation". In: **2023 IEEE International Conference on Robotics and Automation (ICRA)**. IEEE. 2023, pp. 2774–2781.

Acknowledgement

- [redacted]
- [redacted]
- [redacted]
- [redacted]

Thank You !

Email: [redacted]