

Using semi-supervised learning and video point tracking for automatic monitoring of land use changes in regional transportation axes

Mohammad Mohammadi  

1. Master's degree in Computer Engineering - Artificial Intelligence and Robotics, Khajeh Nasiruddin Toosi University of Technology, Tehran, Iran. Email: mohammadmohamadibi79@gmail.com

ARTICLE INFO

Article type:
Research Paper

Article history:

Received 10 September 2025

Received in revised form 10 June 2026

Accepted 21 September 2025

Available online 23 September 2026

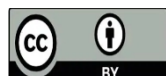
Keywords:

Video semantic segmentation semi-supervised learning land use regional planning

Abstract

Road semantic segmentation plays a key role in scene understanding for autonomous and assisted driving systems, but accurate pixel-level labeling of each frame of road videos is expensive and time-consuming. In this paper, a semi-supervised learning framework for semantic segmentation of road videos is introduced, which relies on exploiting motion constraints extracted from point tracking. In this approach, only the first frame of each 15-frame sequence is manually pixel-labeled, and for subsequent frames, pseudo-label masks are generated using point tracking by the CoTracker3 model. These pseudo-labels are then used to semi-supervisedly train a lightweight LR-ASPP-based segmentation network with a MobileNetV3 backbone. The combined error function consists of a supervised component on labeled frames and a pseudo-supervised component considering tracked pixels. The proposed framework is evaluated on the KITTI-STEP dataset, which is an extended version of KITTI for segmentation and tracking of each pixel in video. The results show that the proposed semi-supervised model SSL-Tracker achieves an mIoU of 27.67% and a pixel accuracy of 72.58%, recording a 2.54% improvement in mIoU over the fully supervised base model trained only with labeled frames (25.13%). A more detailed analysis of the classes shows that the use of point tracking provides a significant advantage, especially for dynamic classes such as “car” and “bicycle”. These results indicate that a significant improvement in semantic road segmentation can be achieved by exploiting the motion information embedded in the video without increasing the number of manual labels.

How to cite: Mohammadi, M. (2026). Using semi-supervised learning and video point tracking for automatic monitoring of land use changes in regional transportation axes. 476-492. *Geography and Regional Planning*, 16(3), <https://doi.org/10.22034/jgeoq.2025.546328.4338>



© Author(s) retain the copyright and full publishing rights

DOI: <https://doi.org/10.22034/jgeoq.2025.546328.4338>

Publisher: Qeshm Institute of Higher Education

Introduction

Semantic road segmentation is a fundamental task in computer vision, particularly for intelligent transportation systems, autonomous vehicles, and advanced driver-assistance systems. It involves assigning a semantic label to each image pixel, identifying categories such as roads, sidewalks, vehicles, pedestrians, buildings, and the sky. Although deep learning methods have significantly improved segmentation accuracy, their effectiveness generally depends on large datasets with pixel-level annotations, which are expensive and time-consuming to produce. Semi-supervised learning offers a promising alternative by combining a limited number of labeled images with a larger collection of unlabeled video frames. Road videos also contain valuable temporal and motion information that can support consistent semantic predictions across consecutive frames. This study proposes a semi-supervised framework that uses CoTracker3 to track points across video sequences and propagate semantic labels from an annotated frame to subsequent unlabeled frames. The generated pseudo-labels are then used to train a lightweight semantic segmentation network based on LR-ASPP and MobileNetV3. The proposed approach is evaluated on the KITTI-STEP dataset to assess its effectiveness in improving segmentation performance while reducing the need for manual pixel-level annotation.

Methodology

This study employs an experimental approach using the KITTI-STEP dataset, divided into non-overlapping sequences of 15 video frames. Only the first frame of each sequence is manually annotated. CoTracker3 tracks sampled points across subsequent frames to generate pseudo-labels. These labels train a semi-supervised LR-ASPP segmentation network with a MobileNetV3 backbone, combining supervised and pseudo-supervised cross-entropy losses. Performance is evaluated using pixel accuracy and mean Intersection over Union (mIoU).

Results and Discussion

The findings indicate that the proposed SSL-Tracker framework improves semantic segmentation by incorporating point-tracking information into semi-supervised learning. Experiments on the KITTI-STEP dataset

show that the proposed approach achieves a pixel accuracy of 72.58% and an mIoU of 27.67% in the reported sampling-grid experiment, compared with 70.47% pixel accuracy and 25.13% mIoU for the baseline model. The results demonstrate that tracking-based pseudo-labels can enhance segmentation performance without requiring additional manual annotations. Class-specific analysis reveals particularly meaningful results for dynamic objects, including cars and bicycles. The reported intersection over union reaches 38.01% for cars and 11.69% for bicycles, whereas improvements for static classes, such as buildings and fences, are more limited. Temporal analysis indicates that pseudo-label quality generally declines as the distance from the annotated frame increases. However, this reduction remains relatively gradual across the 15-frame sequences. The proportion of successfully tracked points also decreases in later frames because of occlusion and objects moving outside the camera's field of view. Additional experiments show that a 300×300 sampling grid achieves the best performance among the tested configurations. Moreover, setting the pseudo-label loss coefficient to 1 produces the highest reported mIoU. Overall, the findings confirm the potential of point tracking to improve semi-supervised road-scene segmentation.

Conclusion

Neoliberalism This study proposed a semi-supervised framework for semantic segmentation of road scenes using point tracking with CoTracker3. The primary objective was to reduce dependence on expensive pixel-level annotations while improving segmentation performance across video sequences. In the proposed approach, only the first frame of each 15-frame sequence requires manual annotation, while pseudo-labels for subsequent frames are generated by tracking points and transferring their semantic labels. These pseudo-labels are then used to train a lightweight segmentation network based on LR-ASPP and MobileNetV3. Experimental results on the KITTI-STEP dataset demonstrate the potential of this approach to improve segmentation performance without increasing manual annotation requirements. The reported experiments show that the proposed method achieves a pixel accuracy of 72.58% and an

mIoU of 27.67%, compared with 70.47% pixel accuracy and 25.13% mIoU for the baseline model. Class-specific analysis indicates that point tracking is particularly beneficial for dynamic objects, including cars and bicycles, whereas improvements for static categories are more limited. Temporal analysis also reveals a gradual decline in pseudo-label quality in later frames, mainly because of occlusion, reduced visibility, and objects leaving the camera's field of view. These findings highlight the importance of reliable tracking and filtering uncertain pixels during pseudo-label generation. Several directions can improve the proposed framework in future research. First, combining point tracking with depth

information, geometric constraints, and camera motion estimation may improve segmentation across both static and dynamic classes. Second, bidirectional temporal consistency and uncertainty estimation could help refine pseudo-labels and reduce error accumulation. Third, integrating more advanced tracking models or multiple complementary trackers may improve label reliability. Overall, the study demonstrates that motion information embedded in road videos provides a valuable resource for semi-supervised semantic segmentation and can support more efficient development of intelligent transportation and autonomous driving systems.

Ethical considerations

Following the principles of research ethics

The authors have observed the principles of ethics in conducting and publishing this scientific research, and this is confirmed by all of them.

Data Availability Statement

Data available on request from the authors.

Acknowledgements

First author: Preparation of samples, conducting experiments and collecting data, performing calculations, statistical analysis of data, analysis and interpretation of information and results, preparing a draft of the article

Ethical Considerations

The authors affirm that they have adhered to ethical research practices, avoiding plagiarism, misconduct, data fabrication or falsification, and have provided their consent for this article's publication.

Funding

This research was conducted without any financial support from Payam Noor University.

Conflict of Interest

The authors declare no conflict of interest

استفاده از یادگیری نیمه نظارتی و ردیابی نقاط ویدیویی برای پایش خودکار تغییرات کاربری اراضی در محورهای مواصلاتی منطقه‌ای

محمد محمدی

۱. فارغ التحصیل کارشناسی ارشد مهندسی کامپیوتر - گرایش هوش مصنوعی و ریاتیکز، دانشگاه صنعتی خواجه نصیرالدین طوسی، رایانامه:

mohamadmohamadibi79@gmail.com

چکیده	اطلاعات مقاله
قطعه‌بندی معنایی جاده نقشی کلیدی در درک صحنه برای سامانه‌های رانندگی خودکار و کمک‌راننده دارد، اما برچسب‌گذاری دقیق پیکسلی برای هر فریم از ویدیوهای جاده‌ای پرهزینه و زمان‌بر است. در این مقاله، یک چارچوب یادگیری نیمه نظارتی برای قطعه‌بندی معنایی ویدیوهای جاده‌ای معرفی می‌شود که بر بهره‌گیری از قیود حرکتی استخراج شده از ردیابی نقاط متکی است. در این رویکرد، تنها فریم اول هر دنباله ۱۵ فریمی دارای برچسب پیکسلی دستی است و برای فریم‌های بعدی، ماسک‌های شبه برچسب با استفاده از ردیابی نقاط توسط مدل CoTracker3 تولید می‌شود. این شبه برچسب‌ها سپس برای آموزش نیمه نظارتی یک شبکه قطعه‌بندی سبک مبتنی بر LR-ASPP با ستون فقرات MobileNetV3 به کار می‌روند. تابع خطای ترکیبی شامل جزء نظارت شده روی فریم‌های دارای برچسب و جزء شبه نظارتی با لحاظ کردن پیکسل‌های ردیابی شده طراحی شده است. چارچوب پیشنهادی بر روی مجموعه داده KITTI-STEP، که نسخه‌ای گسترش یافته از KITTI برای قطعه‌بندی و ردیابی هر پیکسل در ویدیو است، ارزیابی شده است. نتایج نشان می‌دهد که مدل نیمه نظارتی پیشنهادی SSL-Tracker به mIoU برابر ۲۷.۶۷٪ و دقت پیکسلی ۷۲.۵۸٪ دست می‌یابد و نسبت به مدل پایه کاملاً نظارت شده که فقط با فریم‌های دارای برچسب آموزش دیده است (۲۵.۱۳٪)، بهبود ۲.۵۴ درصدی در mIoU را ثبت می‌کند. تحلیل دقیق تر کلاس‌ها نشان می‌دهد که استفاده از ردیابی نقاط، به‌ویژه برای کلاس‌های پویا مانند «خودرو» و «دوچرخه» مزیت قابل توجهی ایجاد می‌کند. این نتایج نشان می‌دهد که می‌توان با بهره‌گیری از اطلاعات حرکتی نهفته در ویدیو و بدون افزایش تعداد برچسب‌های دستی، بهبود قابل ملاحظه‌ای در قطعه‌بندی معنایی جاده به دست آورد.	نوع مقاله: مقاله پژوهشی تاریخ دریافت: ۱۴۰۵/۰۴/۱۹ تاریخ بازنگری: ۱۴۰۵/۰۵/۲۰ تاریخ پذیرش: ۱۴۰۵/۰۶/۳۰ تاریخ انتشار: ۱۴۰۵/۰۷/۰۱ کلیدواژه‌ها: قطعه‌بندی معنایی ویدیویی یادگیری نیمه نظارتی کاربری اراضی برنامه ریزی منطقه ای

استناد: محمدی، محمد. (۱۴۰۵). استفاده از یادگیری نیمه نظارتی و ردیابی نقاط ویدیویی برای پایش خودکار تغییرات کاربری اراضی در محورهای مواصلاتی منطقه‌ای. جغرافیا و برنامه ریزی منطقه‌ای، ۱۶(۳)، ۴۷۶-۴۹۲. DOI: 10.22034/ 10.22034/jgeoq.2025.546328.4338



© نویسندگان.

ناشر: موسسه آموزش عالی قشم

مقدمه

قطعه‌بندی معنایی^۱ جاده یکی از وظایف بنیادی در بینایی کامپیوتری برای سامانه‌های هوشمند حمل‌ونقل، خودروهای خودران و سامانه‌های کمک‌راننده پیشرفته است. در این مسئله، هدف آن است که برای هر پیکسل از تصویر، یک برچسب معنایی از میان کلاس‌هایی مانند جاده، پیاده‌رو، خودرو، عابر پیاده، آسمان و غیره تخصیص داده شود. این سطح از درک صحنه امکان تصمیم‌گیری ایمن‌تر و قابل اتکاتر را برای سیستم‌های کنترل فراهم می‌کند و در نهایت به کاهش تصادفات و بهبود مدیریت ترافیک منجر می‌شود. رویکردهای مبتنی بر یادگیری عمیق^۲ در سال‌های اخیر پیشرفت چشمگیری در قطعه‌بندی معنایی تصویر داشته‌اند و معماری‌های متنوعی از شبکه‌های کاملاً کانولوشنی، SegNet، U-Net و DeepLab برای این منظور معرفی شده‌اند (Long, 2015) (Ronneberger, 2015) (Badrinarayanan, 2017) (Chen L.-C. a., 2018) (Chen L.-C. a., 2017). با این حال، تقریباً تمام این روش‌ها متکی بر مجموعه داده‌های بزرگ کاملاً برچسب‌خورده هستند که تولید آن‌ها، به‌خصوص در سطح پیکسل، هزینه‌ای بسیار بالا تحمیل می‌کند. برچسب‌گذاری پیکسلی ویدیوهای جاده‌ای، علاوه بر زمان زیاد، به تخصص و دقت بالا نیاز دارد و در عمل، گسترش سامانه‌های رانندگی خودکار را با محدودیت منابع مواجه می‌کند.

یادگیری نیمه‌نظارتی^۳ در این زمینه به‌عنوان راهکاری جذاب مطرح شده است؛ زیرا امکان استفاده هم‌زمان از تعداد کم داده برچسب‌دار و حجم زیاد داده بدون برچسب را برای بهبود عملکرد مدل فراهم می‌کند (Nasim, 2017) (Wei-Chih, 2018) (Yassine, 2020) (Yunchao, 2022) (Yuyuan, 2022). در حوزه قطعه‌بندی معنایی، روش‌های نیمه‌نظارتی معمولاً از دو ایده کلیدی بهره می‌برند: تنظیم سازگاری که سازگاری خروجی مدل را تحت اغتشاش‌های ورودی یا فضای ویژگی تحمیل می‌کند، و شبه‌برچسب‌گذاری^۴، که در آن پیش‌بینی‌های مدل روی داده‌های بدون برچسب به‌عنوان برچسب کاذب برای مرحله آموزش مجدد استفاده می‌شود (Yassine, 2020) (Yunchao, 2022) (Yuyuan, 2022) (Yuan, 2022). در ویدیوهای جاده‌ای، علاوه بر اطلاعات ظاهری، قیود هندسی^۵ و حرکتی ناشی از حرکت دوربین و اشیاء در صحنه نیز وجود دارد. این قیود در بسیاری از کارهای کلاس یک مانند برآورد جریان نوری^۶ و ساختار از حرکت مورد استفاده قرار گرفته‌اند، اما استفاده سیستماتیک از آن‌ها در یادگیری نیمه‌نظارتی قطعه‌بندی معنایی هنوز یک موضوع فعال تحقیقاتی است (Horn, 1981) (Dosovitskiy A. a., 2015) (Teed, 2020). در این مقاله، بر این ایده تکیه می‌شود که حرکت اشیاء در فضای تصویر را می‌توان از طریق ردیابی مجموعه‌ای از نقاط ویژگی در طول زمان مدل کرد و این اطلاعات حرکتی را می‌توان به‌طور مؤثری برای انتشار برچسب از یک فریم برچسب‌دار به فریم‌های بدون برچسب در همان دنباله به کار گرفت.

برای تحقق این هدف، از مدل CoTracker3 استفاده می‌شود که یک ردیاب مبتنی بر مبدل^۷ (Vaswani, 2017) (Dosovitskiy A. a., 2021) برای ردیابی نقطه‌ها در ویدیو است (Nikita, 2024) (Kotar, 2023). ترکیب چنین ردیابی دقیقی با یک فریم برچسب‌دار در ابتدای دنباله، امکان تولید شبه‌برچسب‌های پیکسلی برای فریم‌های بعدی را فراهم می‌کند، بدون آن‌که نیازی به مداخله انسانی باشد. این شبه‌برچسب‌ها سپس برای آموزش نیمه‌نظارتی یک شبکه قطعه‌بندی سبک استفاده می‌شوند تا دقت کلی مدل بهبود یابد. در این مقاله، ابتدا وضعیت موجود و کارهای مرتبط در زمینه ردیابی نقاط^۸ و یادگیری نیمه‌نظارتی قطعه‌بندی معنایی مرور می‌شود. سپس، چارچوب پیشنهادی که شامل مرحله تولید شبه‌برچسب مبتنی بر ردیابی نقاط و مرحله آموزش نیمه‌نظارتی شبکه قطعه‌بندی است، تشریح می‌گردد. در ادامه، جزئیات پیاده‌سازی و نتایج تجربی روی مجموعه داده KITTI-STEP ارائه می‌شود و در نهایت، جمع‌بندی و پیشنهادهایی برای پژوهش‌های آینده مطرح می‌گردد.

¹ Semantic Segmentation

² Deep Learning

³ Semi-supervised Learning

⁴ Pseudo-label

⁵ Geometric Constraints

⁶ Optical Flow

⁷ Transformer

⁸ Point Tracking

پیشینه ی پژوهش

در این بخش، دو محور اصلی کارهای مرتبط بررسی می‌شود: ردیابی نقاط در ویدیو و یادگیری نیمه نظارتی قطعه بندی معنایی. همچنین به طور خلاصه به کارهایی اشاره می‌شود که از قیود هندسی در قطعه بندی استفاده کرده اند. در حوزه ردیابی نقاط، روش های کلاسیک بر پایه ویژگی های موضعی و جریان نوری شکل گرفته اند. (Lowe, 2004) (Bay, 2006) ردیاب هایی چون Lucas-Kanade و KLT از تقریب حرکت محلی و فرض ثبات شدت روشنایی برای ردیابی نقاط استفاده می کنند، اما در مواجهه با انسداد، تغییرات شدید ظاهری و صحنه های کم بافت دچار خطا می شوند (Lucas, 1981) (Shi, 1994). با ظهور یادگیری عمیق، شبکه هایی برای برآورد جریان نوری متراکم^۱ مانند FlowNet و RAFT معرفی شدند که دقت بالاتری در حرکت های پیچیده فراهم می کنند (Teed, 2020) (Dosovitskiy A. a., 2015).

نسل جدیدی از روش ها، ردیابی نقاط را به صورت «ردیابی هر نقطه» با استفاده از معماری های مبدل مدل می کنند. CoTracker یک نمونه شاخص از این رویکرد است که با استفاده از یک شبکه مبدل و محاسبه حجم همبستگی چهاربعدی^۲ بین فریم ها، امکان ردیابی مشترک^۳ هزاران نقطه را به طور هم زمان فراهم می کند. نسخه جدیدتر آن، CoTracker3، با بهره گیری از برچسب های کاذب تولید شده توسط ردیاب های دیگر روی ویدیوهای واقعی، از طریق یک فرآیند شبه برچسب گذاری و یادگیری در دنیای واقعی، دقت و پایداری خود را بهبود داده است. در کنار این کارها، روش هایی مانند SAM-PT، ترکیب مدل های قطعه بندی عمومی مانند Segment Anything با ردیاب های نقطه ای را برای قطعه بندی ویدیو پیشنهاد کرده اند.

در حوزه یادگیری نیمه نظارتی برای قطعه بندی معنایی، دو خط اصلی پژوهش وجود دارد. نخست، روش های مبتنی بر تنظیم سازگاری^۴ هستند که در آن ها مدل ملزم می شود پیش بینی های سازگاری تحت اغتشاش های مختلف در ورودی یا فضای ویژگی ها داشته باشد. نمونه هایی مانند Cross-Consistency Training با استفاده از چند کدگشا و اعمال اغتشاش بر روی بازنمایی^۵ ها، مدل را وادار به تولید خروجی های یکسان می کنند و از این طریق از داده های بدون برچسب بهره می برند. دوم، روش های مبتنی بر برچسب های کاذب هستند که ابتدا یک مدل اولیه با داده های برچسب دار آموزش می دهند و سپس با استفاده از پیش بینی های آن روی داده های بدون برچسب، شبه برچسب تولید کرده و مدل را مجدداً آموزش می دهند. آثار جدیدتر با اضافه کردن برآورد عدم قطعیت، حافظه های رده بندی شده و توابع خطای تمایزی، کیفیت شبه برچسب ها و پایداری آموزش را بهبود داده اند.

استفاده از قیود هندسی و اطلاعات سه بعدی در قطعه بندی نیز مورد توجه قرار گرفته است. برای مثال، برخی پژوهش ها نشان داده اند که ترکیب عمق و محدودیت های هندسی با شبکه های قطعه بندی می تواند کیفیت مرزها و سازگاری فضایی را بهبود دهد. همچنین در زمینه ویدیو، کارهایی به استخراج ویژگی های پایدار در زمان و استفاده از جریان نوری برای پالایش نقشه های قطعه بندی پرداخته اند. با وجود این پیشرفت ها، استفاده مستقیم از ردیابی نقاط، آن هم در قالب ردیابی مشترک نقطه ها با معماری های مبدل، برای تولید شبه برچسب های ویدیویی در قطعه بندی معنایی هنوز به طور محدود مورد بررسی قرار گرفته است. این مقاله این شکاف را هدف قرار می دهد و نشان می دهد که بهره گیری از CoTracker3 برای انتشار ماسک های معنایی در دنباله های ویدیویی جاده ای می تواند به طور مؤثر در یادگیری نیمه نظارتی مورد استفاده قرار گیرد.

روش پیشنهادی مبتنی بر ردیابی نقاط

چارچوب پیشنهادی شامل دو مؤلفه اصلی است: تولید شبه برچسب برای فریم های بدون برچسب در هر دنباله ویدیویی با استفاده از CoTracker3، و سپس آموزش نیمه نظارتی یک شبکه قطعه بندی معنایی با استفاده از این شبه برچسب ها. در این بخش، هر یک از این دو مؤلفه به تفصیل تشریح می شود.

¹ Dense

² 4D Correlation

³ Joint Tracking

⁴ Consistency Regularization

⁵ Representation

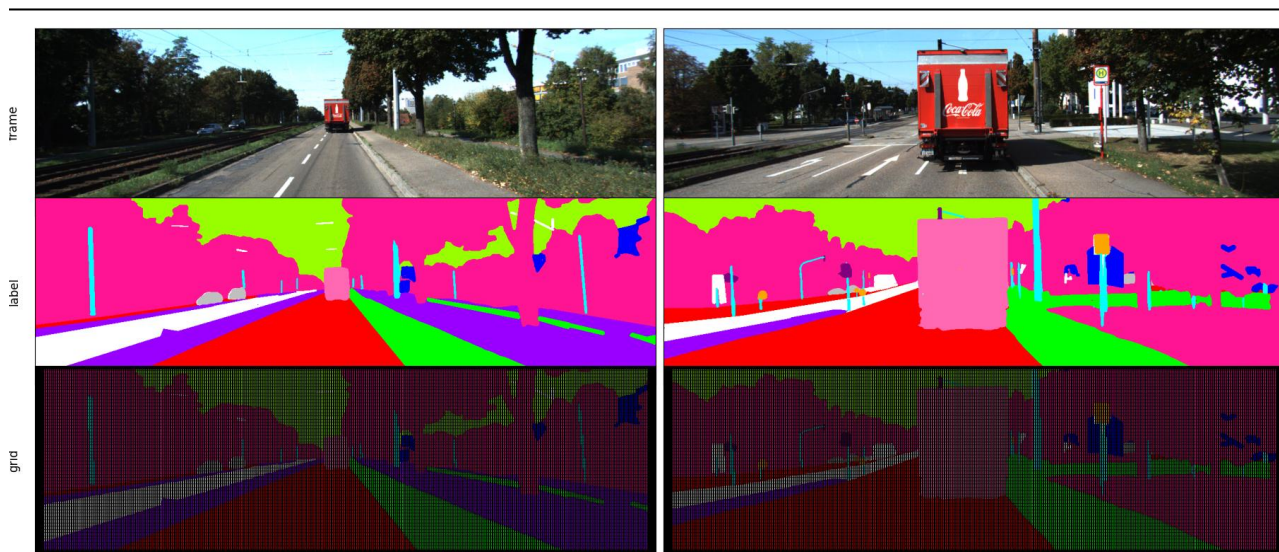
⁶ Loss Functions

ساختار دنباله‌های ویدیویی و داده برچسب‌دار

مجموعه داده مورد استفاده، KITTI-STEP (Geiger, 2012) است که بر پایه KITTI Tracking ساخته شده و شامل دنباله‌های ویدیویی از صحنه‌های رانندگی در محیط‌های شهری و جاده‌ای، همراه با ماسک‌های قطعه‌بندی^۱ معنایی در سطح پیکسل برای برخی فریم‌ها است. این مجموعه داده از نظر تنوع شرایط نوری، حضور اشیاء پویا و تنوع صحنه، چالشی و مناسب برای ارزیابی روش‌های ویدیویی است. ویدیوها در این کار به دنباله‌های غیرهمپوشان با طول ۱۵ فریم تقسیم شده‌اند. تنها فریم اول هر دنباله به عنوان فریم برچسب‌دار در نظر گرفته شده و ماسک معنایی پیکسلی آن به عنوان داده مرجع واقعی به کار می‌رود. چهارده فریم بعدی در هر دنباله، بدون برچسب انسانی هستند و هدف، تولید شبه برچسب برای این فریم‌ها است. این تنظیم مسئله، سناریویی واقع‌گرایانه را شبیه‌سازی می‌کند که در آن اپراتور انسانی تنها بخشی از فریم‌ها را برچسب‌گذاری کرده است.

انتخاب نقاط اولیه برای ردیابی

شکل ۱: شبکه نمونه‌برداری شده فریم اول



برای استفاده از ردیابی نقاط، ابتدا لازم است مجموعه‌ای از نقاط اولیه روی فریم اول هر دنباله انتخاب شود. این نقاط باید به گونه‌ای توزیع شوند که تا حد امکان سطح تصویر را پوشش دهند و در عین حال، قابلیت ردیابی مناسبی داشته باشند. در این کار، از یک شبکه ۳۰۰ در ۳۰۰ برای نمونه‌برداری نقاط استفاده شده است. بدین ترتیب، نقاط روی یک گرید منظم در مختصات تصویر^۲ قرار می‌گیرند و هر نقطه، مختصات پیکسلی متناظری در فریم اول دارد. این راهبرد موجب می‌شود که پوشش فضایی مناسب در تمام نواحی تصویر صورت گیرد، به ویژه در نواحی مربوط به جاده، خودرو و سایر اشیاء مهم. استفاده از گرید منظم، علاوه بر سادگی، امکان تحلیل آماری دقیق نسبت نقاط ردیابی شده در فریم‌های مختلف را نیز فراهم می‌آورد. نمونه‌های از شبکه نقاط نمونه‌برداری شده روی فریم اول در شکل **Error! Reference source not found.** نمایش داده شده است.

ردیابی مشترک نقاط با CoTracker3

پس از انتخاب نقاط اولیه، دنباله ۱۵ فریمی و مجموعه نقاط اولیه به عنوان ورودی به مدل CoTracker3 (Nikita, 2024). CoTracker3 یک مدل مبتنی بر مبدل (Vaswani, 2017) است که با دریافت فریم‌های ویدیو و مختصات نقاط در

¹ Segmentation Masks

² Image Coordinates

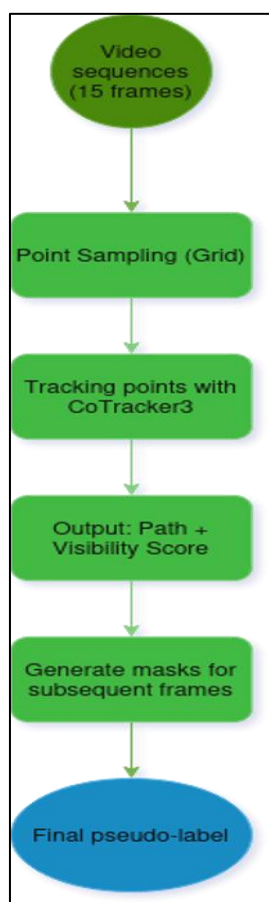
یک یا چند فریم، مسیر هر نقطه را در تمام فریم‌ها تخمین می‌زند و علاوه بر مختصات، امتیازی برای قابلیت رؤیت^۱ هر نقطه در هر فریم نیز برمی‌گرداند.

در سطح مفهومی، CoTracker3 ابتدا ویژگی‌های مکانی-زمانی را برای هر فریم با یک کدگذار کانولوشنی استخراج می‌کند. سپس یک حجم همبستگی چهاربعدی بین ویژگی‌های فریم‌هایی که نقطه‌ها در آن‌ها تعریف شده و سایر فریم‌ها محاسبه می‌کند تا شباهت مکانی بین نواحی مختلف را به دست آورد. این حجم همبستگی به همراه توصیف‌گرهای نقطه‌ها به یک شبکه مبدل داده می‌شود که در چندین تکرار، مختصات نقاط، امتیاز اطمینان^۲ و قابلیت رؤیت را به‌روز می‌کند. یادگیری این مدل با ترکیبی از خطای هابر بر روی مختصات، و آنتروپی متقاطع برای قابلیت رؤیت و اطمینان انجام شده است. خروجی CoTracker3 برای هر دنباله ۱۵ فریمی، شامل مختصات دوبعدی هر نقطه در تمام فریم‌ها، و نیز امتیاز قابلیت رؤیت در بازه است. این اطلاعات پایه تولید شبه‌برچسب‌های پیکسلی را تشکیل می‌دهد.

تبدیل مسیر نقاط به شبه‌برچسب پیکسلی

برای تولید ماسک معنایی در فریم‌های بدون برچسب، از مسیر نقاط ردیابی شده استفاده می‌شود. ابتدا، برای فریم اول دنباله، کلاس معنایی هر نقطه به‌سادگی از روی ماسک برچسب‌دار خوانده می‌شود؛ بدین صورت که مختصات نقطه در ماسک حقیقی جست‌وجو شده و مقدار کلاس در آن پیکسل به‌عنوان برچسب نقطه ثبت می‌شود. سپس برای هر یک از ۱۴ فریم بعدی، نقاطی که قابلیت رؤیت بالاتر از یک آستانه مشخص دارند، فعال فرض شده و مختصات آن‌ها روی تصویر مقصد به‌عنوان حامل برچسب کلاس متناظر ثبت می‌شود. نمای کلی فرآیند تولید شبه‌برچسب مبتنی بر ردیابی نقاط در شکل **Error! Reference source not found.** نشان داده شده است.

^۱ Visibility
^۲ Descriptor
^۳ Confidence



شکل ۲: فرآیند تولید شبه‌برچسب با ردیابی نقاط

شبکه قطعه‌بندی معنایی و تابع خطا

برای قطعه‌بندی معنایی، از معماری LR-ASPP با ستون فقرات MobileNetV3 (Howard, 2019) استفاده شده است. این معماری به گونه‌ای طراحی شده که تعادل مناسبی بین دقت و کارایی محاسباتی برقرار کند و برای کاربردهای بلادرنگ مانند رانندگی خودکار مناسب است. ستون فقرات MobileNetV3 با استفاده از بلوک‌های باقی‌مانده معکوس و کانولوشن‌های تفکیک‌پذیر عمقی^۱، ویژگی‌های غنی و فشرده‌ای از تصویر استخراج می‌کند و مازول LR-ASPP با تجميع اطلاعات چندمقیاسی، نقشه ویژگی مناسب برای قطعه‌بندی پیکسل‌به‌پیکسل را تولید می‌نماید.

در مرحله آموزش، در هر گام، یک دسته^۲ شامل ترکیبی از فریم‌های دارای برچسب واقعی^۳ (فریم‌های اول دنباله‌ها) و فریم‌های دارای شبه‌برچسب (فریم‌های دوم تا پانزدهم) ساخته می‌شود. تابع خطای کلی برای به‌روزرسانی پارامترهای شبکه، ترکیبی از خطای نظارت‌شده و خطای شبه‌نظارتی است. جزء نظارت‌شده بر اساس آنتروپی متقاطع پیکسلی^۴ بین خروجی شبکه و برچسب‌های حقیقی تعریف می‌شود. جزء شبه‌نظارتی نیز از آنتروپی متقاطع صرفاً برای نقاط ردیابی شده استفاده می‌کند، به این معنا که خطای سایر پیکسل‌ها در محاسبه تابع هدف لحاظ نمی‌شود.

در نهایت، تابع خطای کلی به صورت زیر تعریف می‌شود:

$$L_{\text{total}} = L_{\text{supervised}} + \lambda L_{\text{pseudo-supervised}} \quad (1)$$

که در آن λ ضریبی برای تنظیم میزان اثر شبه‌برچسب‌ها در فرآیند آموزش است.

¹ Depthwise Separable Convolutions

² Batch

³ Ground Truth

⁴ Pixel-wise Cross-Entropy

در آزمایش‌های انجام‌شده، مقادیر مختلف λ بررسی شده و مقدار λ برابر ۱ بهترین تعادل را بین یادگیری از داده واقعی و شبه‌برچسب‌ها نشان داده است.

تنظیمات آزمایش و معیارهای ارزیابی

پایه‌سازی مدل‌ها با استفاده از کتابخانه PyTorch انجام شده و آموزش بر روی یک پردازنده گرافیکی RTX۳۰۸۰Ti صورت گرفته است؛ تنها مرحله ردیابی نقاط با گرید ۳۰۰ در ۳۰۰، به دلیل بار محاسباتی بالا، روی پردازنده گرافیکی Tesla۲۰۰H اجرا شده است. تصاویر ورودی تا حد لازم به ابعاد ثابت تغییر اندازه داده شده‌اند و افزایش داده^۱ متعارف شامل برش تصادفی اعمال شده است. بهینه‌سازی با استفاده از الگوریتم SGD (Robbins, 1951) و نرخ یادگیری اولیه ۰/۰۰۰۱ انجام شده و نرخ یادگیری به کمک زمان‌بند MultiStepLR در طول آموزش کاهش یافته است. تعداد دورهای آموزش ۱۰۰ در نظر گرفته شده است. برای ارزیابی عملکرد، از دو معیار استاندارد قطعه‌بندی معنایی استفاده شده است. نخست، دقت پیکسلی^۲ که نسبت پیکسل‌های به‌درستی طبقه‌بندی شده به کل پیکسل‌ها را نشان می‌دهد. دوم، همپوشانی بر اساس اجتماع و اشتراک (IoU) که برای هر کلاس به صورت نسبت اشتراک پیش‌بینی و برچسب واقعی بر اجتماع آن‌ها تعریف می‌شود و میانگین آن روی کلاس‌ها به‌عنوان mIoU گزارش می‌شود. این معیار در ارزیابی‌های بین‌المللی مانند KITTI و Cityscapes نیز استفاده می‌شود.

نتایج و بحث

در این بخش، نتایج کمی و کیفی مدل SSL-Tracker ارائه و تحلیل می‌شود و با مدل پایه کاملاً نظارت‌شده مقایسه می‌گردد.

نتایج کلی

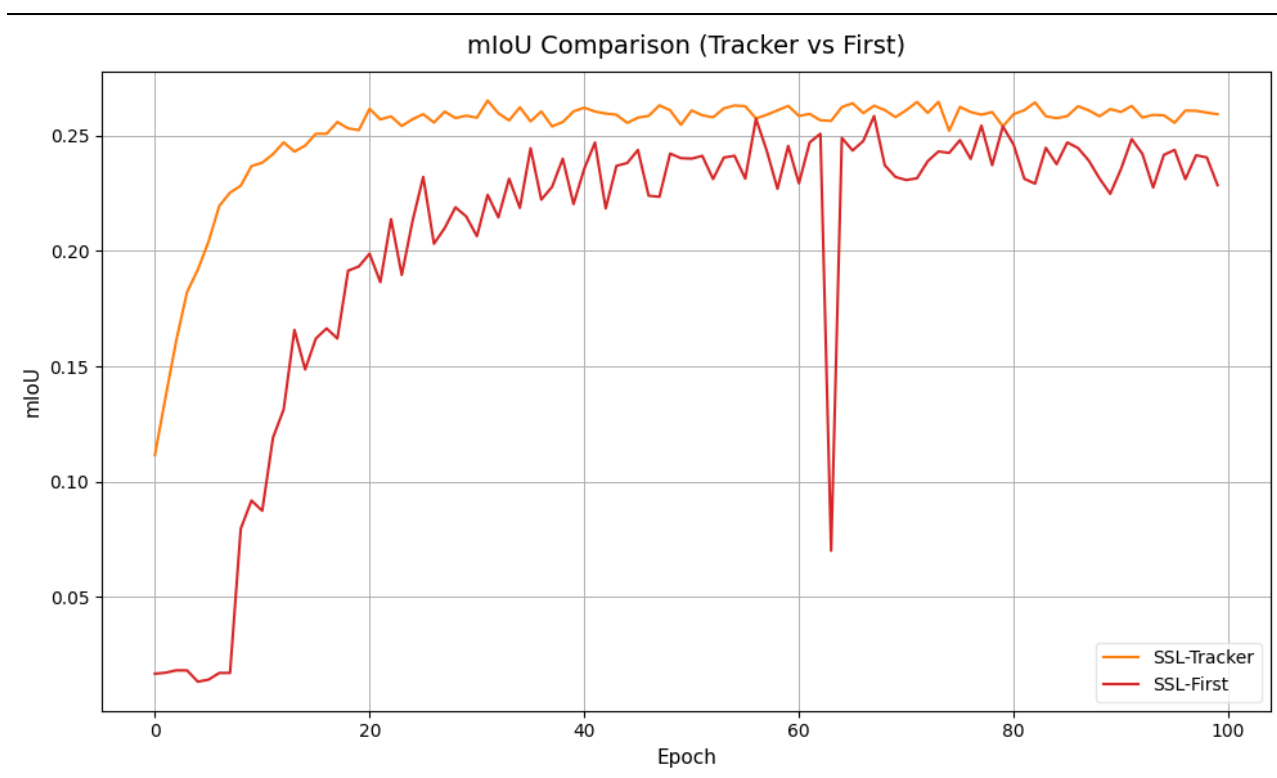
جدول **Error! Reference source not found.** نتایج مدل پایه نظارت‌شده و مدل نیمه‌نظارتی پیشنهادی را روی مجموعه داده ارزیابی KITTI-STEP نشان می‌دهد. در مدل پایه نظارت‌شده SSL-First شبکه تنها با استفاده از فریم‌های اول دنباله‌ها که دارای برچسب واقعی هستند، آموزش داده شده است. نمودار مقایسه mIoU بین مدل پایه و مدل پیشنهادی در شکل **Error! Reference source not found.** نمایش داده شده است.

جدول ۱: مقایسه عملکرد روش‌ها

روش	mIoU	Accuracy
روش پایه	25.13	70.47
روش پیشنهادی	26.67	72.58

¹ Data Augmentation

² Pixel Accuracy



شکل ۳: نمودار مقایسه mIoU

همان‌گونه که مشاهده می‌شود، استفاده از شبه‌برچسب‌های تولیدشده توسط ردیابی نقاط باعث افزایش 2.54 درصدی در mIoU و حدود 2.1 درصدی در دقت پیکسلی می‌شود. این بهبود، با توجه به این‌که هیچ برچسب انسانی اضافی در فریم‌های میانی دنباله‌ها وارد نشده است، نشان‌دهنده اثربخشی رویکرد پیشنهادی در بهره‌گیری از داده‌های بدون برچسب است.

تحلیل به تفکیک کلاس

برای بررسی دقیق‌تر اثر روش پیشنهادی، مقدار IoU برای کلاس‌های مختلف گزارش شده است. نتایج نشان می‌دهد که بیشترین بهبود در کلاس‌های پویا مانند «خودرو» و «دوچرخه» به دست آمده است. برای مثال، IoU کلاس «خودرو» در مدل پایه به‌طور چشمگیری پایین‌تر از مقدار متناظر در مدل SSL-Tracker بوده و در مدل پیشنهادی به حدود 38 درصد رسیده است. همچنین کلاس «دوچرخه» که عموماً به دلیل اندازه کوچک و ظهور محدود در صحنه، یک کلاس دشوار محسوب می‌شود، در مدل SSL-Tracker به IoU حدود 11.69 درصد دست یافته است. در مقابل، در برخی کلاس‌های ایستا مانند «ساختمان» یا «حصار»، بهبود کمتر بوده یا در حد ناچیزی باقی مانده است. این رفتار با ماهیت روش ردیابی نقطه‌ای قابل تبیین است؛ زیرا نقاط ردیابی شده روی اشیاء متحرک با تغییرات قابل‌توجه مکان در طول فریم‌ها حرکت می‌کنند و اطلاعات بیشتری در مورد تغییر شکل مرز و ناحیه این اشیاء فراهم می‌آورند، در حالی که برای نواحی بزرگ و تقریباً ایستا، فواید نسبی کمتری از ردیابی نقاط حاصل می‌شود. مقادیر IoU به تفکیک کلاس‌ها در جدول Error! Reference source not found. ارائه شده است.

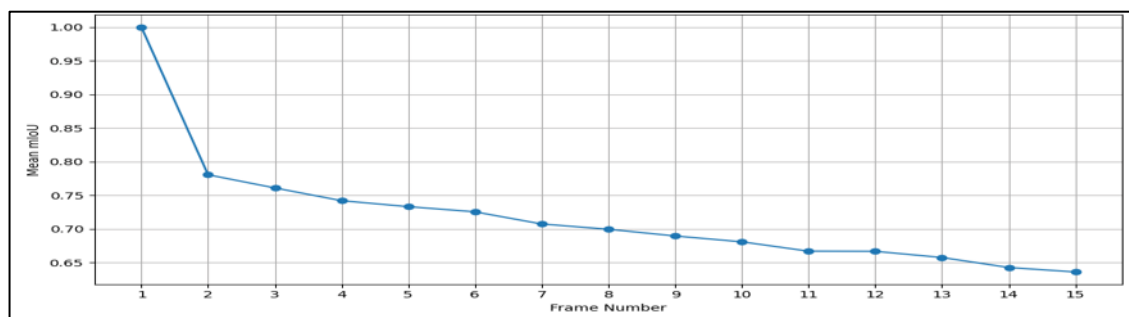
جدول ۲: مقدار IoU به تفکیک کلاس

کلاس	SSL-Tracker
جاده	66.41
پیاده‌رو	10.47

13.84	ساختمان
0	دیوار
3.28	حصار
8.62	تیر چراغ برق
0.04	چراغ راهنمایی
7.58	علامت راهنمایی و رانندگی
74.47	پوشش گیاهی
40.81	زمین
71.40	آسمان
1.13	عابر پیاده
0.2	راننده
38.01	خوردو
0.11	کامیون
0.18	اتوبوس
0	قطار
0	موتور سیکلت
11.69	دوچرخه

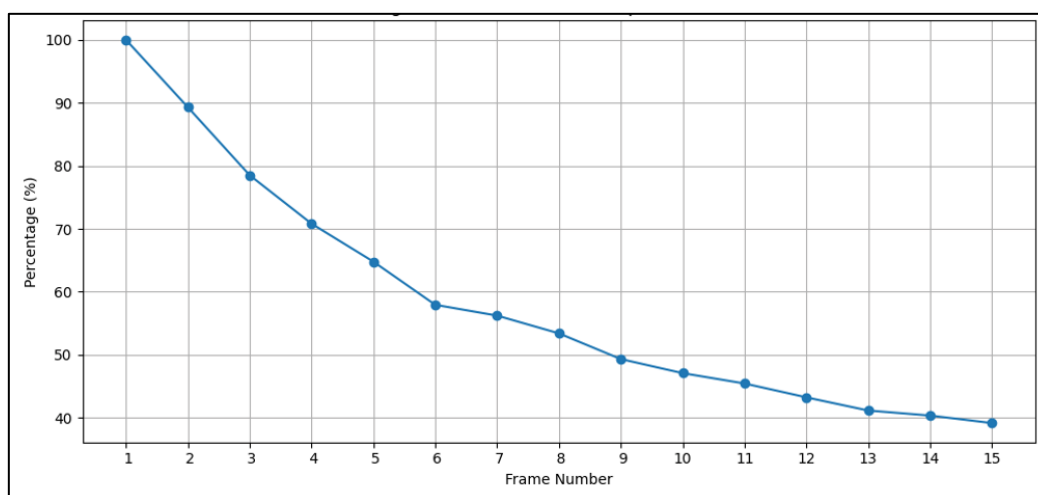
پایداری در طول فریم‌ها و مسئله انباشت خطا

یکی از نگرانی‌های محتمل در روش‌های مبتنی بر انتشار در طول دنباله، انباشت خطا^۱ در فریم‌های دورتر از فریم برچسب‌دار است. برای بررسی این مسئله، mIoU حاصل از شبه‌برچسب‌ها به‌عنوان تابعی از شماره فریم در دنباله تحلیل شده است. نمودار تغییرات mIoU شبه‌برچسب‌ها نسبت به شماره فریم در شکل **Error! Reference source not found.** نمایش داده شده است.



شکل ۴: میانگین mIoU به ازای هر فریم برای روش SSL-Tracker

شکل ۵: درصد نقاط قابل ردیابی در فریم‌های مختلف

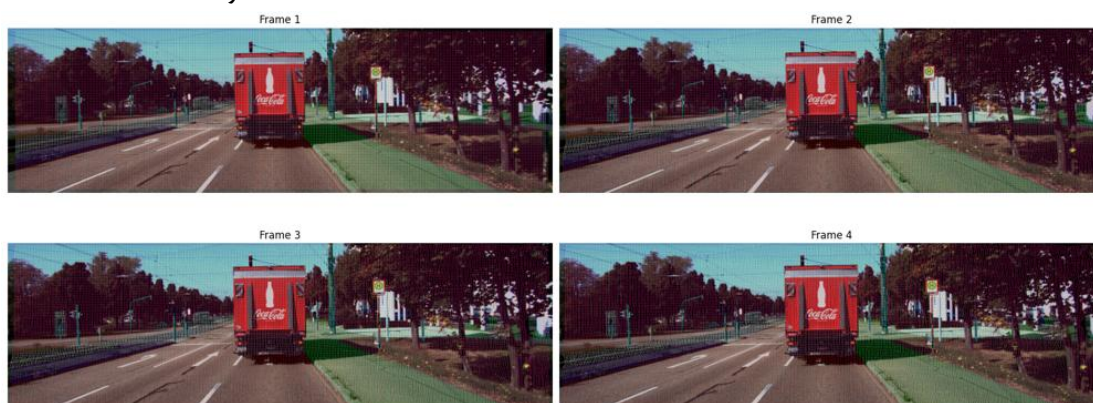


نتایج نشان می‌دهد که به‌طور کلی، mIoU شبه‌برچسب‌ها در فریم‌های ابتدایی دنباله بالاتر بوده و با فاصله گرفتن از فریم اول، به‌تدریج کاهش می‌یابد، اما این کاهش در طول 15 فریم بررسی‌شده نسبتی ملایم دارد و همچنان در محدوده قابل‌قبولی باقی می‌ماند. درصد نقاطی که در فریم‌های مختلف نسبت به فریم اول قابل ردیابی باقی مانده‌اند در شکل **Error! Reference source not found.** نشان داده شده است.

نتایج حاکی از آن است که در فریم‌های میانی دنباله، بخش عمده‌ای از نقاط هنوز قابل ردیابی‌اند، اما در فریم‌های انتهایی، درصد قابل توجهی از نقاط به علت خروج از میدان دید یا انسداد از دست می‌روند. این پدیده تأییدی بر اهمیت طراحی آستانه قابلیت رؤیت و کنار گذاشتن پیکسل‌های نامطمئن^۱ در تولید شبه‌برچسب‌ها است.

تحلیل کیفی

در سطح کیفی، بر روی نمونه‌هایی از دنباله‌های ویدیویی، ماسک‌های تولیدشده توسط مدل نیمه‌نظارتی با برچسب‌های حقیقی مقایسه شده‌اند. این بررسی بصری نشان می‌دهد که مدل SSL-Tracker توانسته است مرز خودروها، نواحی جاده و پوشش گیاهی را به‌طور نسبتاً دقیق دنبال کند. در بسیاری از موارد، مدل حتی در حضور انسدادهای جزئی یا تغییرات نوری متوسط، توانسته است شکل کلی اشیاء را در طول فریم‌ها حفظ کند. نمونه‌هایی از ماسک‌های تولیدشده در بخش‌های مختلف دنباله در شکل‌های **Error! Reference source not found.** تا **Error! Reference source not found.** ارائه شده‌اند.

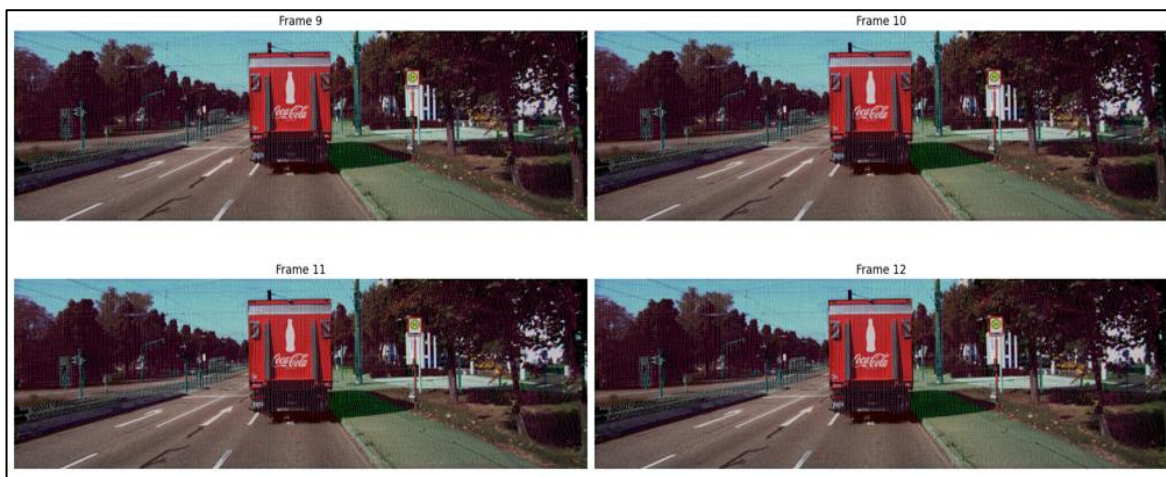


شکل ۶: ماسک‌های پیش‌بینی شده با استفاده از ردیابی نقاط بخش

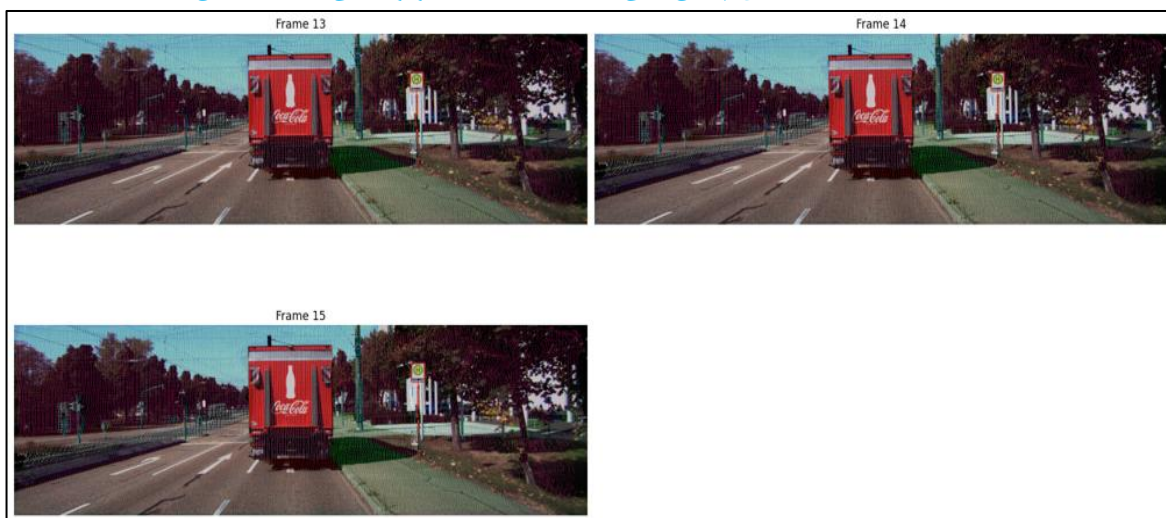
شکل ۷: ماسک‌های پیش‌بینی شده با استفاده از ردیابی نقاط بخش



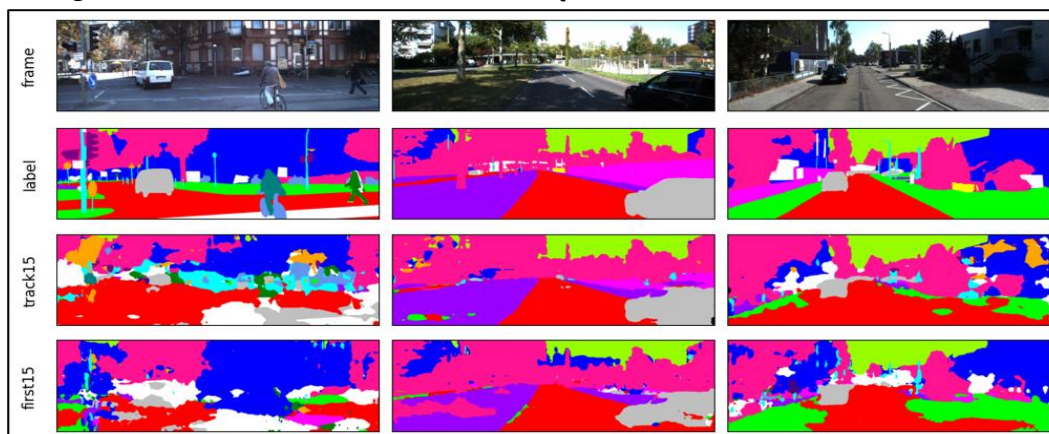
شکل ۸: ماسک‌های پیش‌بینی شده با استفاده از ردیابی نقاط بخش



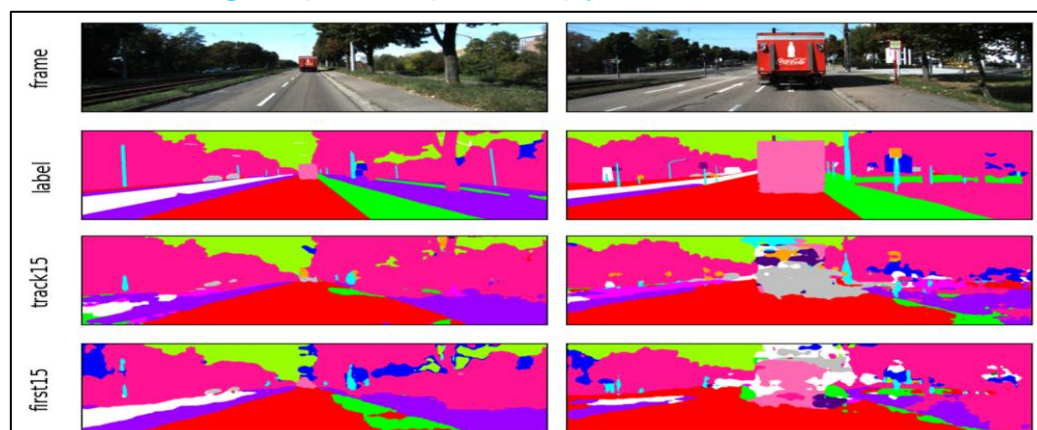
شکل ۹: ماسک‌های پیش‌بینی شده با استفاده از ردیابی نقاط بخش



در عین حال، در برخی صحنه‌ها، به‌ویژه زمانی که اشیاء بسیار کوچک، دوردست یا دچار انسداد شدید هستند، اشتباهاتی در برچسب‌گذاری مشاهده می‌شود. همچنین در فریم‌های انتهایی دنباله، مواردی از رانش در شبه‌برچسب‌ها مشاهده شده که به خطا در مرزبندی میان جاده و پیاده‌رو منجر شده است. این نقاط ضعف نشان می‌دهد که هنوز فضای قابل‌توجهی برای بهبود از طریق ترکیب ردیابی نقاط با سایر قیود هندسی یا اطلاعات عمقی وجود دارد. مقایسه بصری خروجی مدل پایه و مدل SSL-Tracker در شکل‌های Error! Reference source not found. و Error! Reference source not found. قابل مشاهده است.



شکل ۱۰: مقایسه بصری ماسک‌های قطعه‌بندی بخش



شکل ۱۱: مقایسه بصری ماسک‌های قطعه‌بندی بخش

مطالعه حذف مؤلفه

تأثیر تراکم شبکه نمونه‌برداری بر عملکرد مدل در جدول ۱ نشان داده شده است.

جدول ۱: بررسی تأثیر تراکم نقاط نمونه‌برداری

شبکه نمونه‌برداری اولیه	MioU %	PA %
42 * 42	26.35	70.71
84 * 84	26.62	70.41
300 * 300	27.67	72.58

نتایج تغییر مقدار ضریب λ در جدول ۲ گزارش شده است.

جدول ۲: مقدار $mIoU$ به ازای مقادیر مختلف λ

مقدار λ	SSL-Tracker
0.5	26.04
1	27.67

نتیجه گیری

در این مقاله، یک چارچوب نیمه نظارتی برای قطعه بندی معنایی تصاویر جاده ای ارائه شد که بر بهره برداری از ردیابی مشترک نقاط در ویدیو با استفاده از CoTracker3 استوار است. تنها با استفاده از برچسب پیکسلی در فریم اول هر دنباله و تولید شبه برچسب برای فریم های بعدی، شبکه قطعه بندی LR-ASPP با ستون فقرات MobileNetV3 توانست عملکردی فراتر از مدل پایه کاملاً نظارت شده با همان حجم برچسب دستی کسب کند. نتایج تجربی روی مجموعه داده KITTI-STEP نشان داد که این رویکرد بهبود ملموسی در معیار mIoU به خصوص برای کلاس های پویا، به همراه دارد. تحلیل کیفی نیز نشان داد که مدل در حفظ سازگاری معنایی اشیاء در طول فریم ها، حتی در حضور حرکات پیچیده، عملکرد قابل قبولی فراهم می کند. برای کارهای آینده، می توان چند جهت را پیشنهاد کرد. نخست، ترکیب همین قیود حرکتی با قیود هندسی مبتنی بر عمق و حرکت دوربین^۱، به طوری که برای کلاس های ایستا از روش های مبتنی بر عمق و برای کلاس های پویا از ردیابی نقاط استفاده شود. دوم، طراحی مکانیزم های پالایش شبه برچسب ها، مانند اعمال سازگاری دوطرفه در زمان (هم به جلو و هم به عقب در دنباله) و استفاده از مدل های تخمین عدم قطعیت^۲ برای وزن دهی به شبه برچسب ها در تابع خطا. سوم، بهره گیری از ردیاب های جدیدتر یا ترکیب چند ردیاب برای بهبود کیفیت برچسب های کاذب.

References

- Badrinarayanan, V. a. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
- Bay, H. a. (2006). SURF: Speeded Up Robust Features. European Conference on Computer Vision (ECCV), 404--417.
- Chen, L.-C. a. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587.
- Chen, L.-C. a. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 834--848.
- Dosovitskiy, A. a. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2758--2766.
- Dosovitskiy, A. a. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).
- Duckworth, D. a. (2023). TAPIR: Tracking Any Point with per-frame Initialization and Refinement. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
- Duckworth, D. a. (2024). BootsTAPIR: Self-supervised Tracking Any Point. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Geiger, A. a. (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3354--3361.
- Horn, B. K. (1981). Determining optical flow. Artificial Intelligence, 185--203.
- Howard, A. a.-C. (2019). Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1314--1324.
- Kotar, G. B. (2023). CoTracker: It is Better to Track Together. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Lee, S. C.-Y. (2024). Local All-Pair Correspondence for Point Tracking. Proceedings of the European Conference on Computer Vision (ECCV).
- Long, J. a. (2015). Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431--3440.

¹ Depth and Ego-Motion
² Uncertainty Estimation

- Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 91--110.
- Lucas, B. D. (1981). An iterative image registration technique with an application to stereo vision. *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, 1981.
- Nasim, S. a. (2017). Semi Supervised Semantic Segmentation Using Generative Adversarial Network. *IEEE International Conference on Computer Vision (ICCV)*.
- Nikita, K. a. (2024). COTRACKER3: Simpler and Better Point Tracking by Pseudo-Labeling Real Videos. *International Conference on Computer Vision (ICCV)*.
- Robbins, H. a. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 400--407.
- Ronneberger, O. a. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234--241.
- Shi, J. a. (1994). Good features to track. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 593--600.
- Teed, Z. a. (2020). RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. *European Conference on Computer Vision (ECCV)*, 402--419.
- Vaswani, A. a. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 5998--6008.
- Wei-Chih, H. (2018). Adversarial Learning for Semi-Supervised Semantic Segmentation. *British Machine Vision Conference (BMVC)*.
- Yassine, O. a. (2020). Semi-Supervised Semantic Segmentation With Cross-Consistency Training. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuan, L. W. (2022). {U2PL}: A Unified Framework for Pseudo-Label-based Semi-Supervised Medical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2022*, 385--395.
- Yunchao, W. a. (2022). Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuyuan, L. a. (2022). Perturbed and Strict Mean Teachers for Semi-supervised Semantic Segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.