دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

پروژه کارشناسی

ادغام بدون نظارت دوربین و لیدار برای پیشنهاد ناحیه سریع

نگارش

بردیا اردکانیان

استاد راهنما

مهدی جوانمردی

شهریور ۱۴۰۳


بسم الله الرحمن الرحيم

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ: شهریور ۱۴۰۳

اینجانب **بردیا اردکانیان** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

بردیا اردکانیان

امضا

تقدیم به

آنان که الفبای انسانیت و چگونه زیستن را به من آموختند...

سپاس گزاری

بدین وسیله از زحمات و تلاش بی‌دریغ استاد محترم جناب دکتر مهدی جوانمردی و خانواده عزیزم صمیمانه سپاسگزاری می نمایم و همچنین از سایر همکاران و دوستانی که هر کدام به نحوی در تهیه این مجموعه با این جانب همکاری داشته اند تشکر نموده و موفقیت همه آنها را از خداوند متعال خواهانم.

برویا ارکانیان
شهریور ۱۴۰۳

چکیده

در سال‌های اخیر، ادغام داده‌های چند سنسوری مانند دوربین و لیدار در سامانه‌های تشخیص شیء توجه بسیاری را به خود جلب کرده است. این ادغام می‌تواند به بهبود دقت و کارایی در شناسایی اشیاء کمک کند. با این حال، جمع‌آوری و برچسب‌گذاری دستی داده‌های مورد نیاز برای آموزش مدل‌های تشخیص شیء، فرآیندی زمان‌بر و هزینه‌بر است. در این پژوهش، هدف اصلی ارزیابی رویکردهای خوشه‌بندی بدون نظارت بر روی پیشنهاد ناحیه بر روی سامانه تشخیص اشیاست. به منظور انجام این ارزیابی، داده‌های دوبعدی تصاویر دوربین و داده‌های سه‌بعدی لیدار ادغام شده و با استفاده از الگوریتم‌های خوشه‌بندی بدون نظارت، مناطق حاوی اشیاء شناسایی می‌شوند. نتایج به دست آمده نشان می‌دهند که این روش می‌تواند بدون نیاز به داده‌های برچسب‌دار، مناطق حاوی اشیاء را با دقت قابل قبولی تشخیص دهد، که می‌تواند به کاهش هزینه‌ها و زمان مورد نیاز برای توسعه سامانه‌های تشخیص شیء کمک کند.

واژه‌های کلیدی:

پیشنهاد ناحیه، یادگیری بدون نظارت، ادغام داده‌های دوربین و لیدار، تشخیص شیء، خوشه‌بندی بدون نظارت

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فصل اول

مقدمه

۱-۱ شرح مسئله

در دهه‌های اخیر، خودروهای خودران[1]، به ویژه در حوزه تشخیص شیء[2]، پیشرفت‌های چشمگیری داشته است. تشخیص شیء یکی از مؤلفه‌های کلیدی در خودروهای خودران است، زیرا امکان درک محیط و تصمیم‌گیری‌های ایمن را فراهم می‌کند. الگوریتم‌های تشخیص شیء به‌طور کلی به دو دسته تقسیم می‌شوند: روش‌های تک‌مرحله‌ای و روش‌های دومرحله‌ای.

در روش‌های تک‌مرحله‌ای مانند یولو[3] [۱]، تشخیص و طبقه‌بندی اشیاء در یک مرحله و به‌صورت همزمان انجام می‌شود. این روش‌ها به دلیل سرعت بالای پردازش، برای کاربردهای بلادرنگ مناسب هستند، اما ممکن است دقت کمتری نسبت به روش‌های دومرحله‌ای داشته باشند.

در مقابل، روش‌های دومرحله‌ای مانند فَستر آر-سی ان ان[4] [۲] ابتدا با استفاده از تکنیک‌های پیشنهاد ناحیه[5]، مناطقی از تصویر را که احتمالاً حاوی اشیاء هستند شناسایی می‌کنند. سپس در مرحله دوم، این نواحی توسط یک طبقه‌بند مورد تحلیل و تشخیص قرار می‌گیرند. این رویکرد با تمرکز بر مناطق مرتبط، دقت بالاتری در تشخیص شیء ارائه می‌دهد، هرچند ممکن است زمان پردازش بیشتری نسبت به روش‌های تک‌مرحله‌ای نیاز داشته باشد.

پیشنهاد ناحیه نقش مهمی در خودروهای خودران ایفا می‌کند، زیرا با شناسایی دقیق مکان اشیائی مانند خودروها، عابران پیاده و موانع، به خودروهای خودران امکان می‌دهد تا تصمیمات ایمنی در مورد مسیریابی و جلوگیری از برخورد اتخاذ کنند.

این فرآیند تاکنون عمدتاً بر اساس اطلاعات دریافتی از سنسورهایی مانند دوربین‌ها انجام شده است. با این حال، ادغام داده‌های حاصل از سنسورهای مختلف، مانند دوربین و لیدار، می‌تواند به ایجاد مجموعه داده‌های غنی‌تری منجر شود که درک بهتری از محیط ارائه می‌دهند. هدف این پروژه، بررسی و ارزیابی عملکرد ادغام داده‌های دوربین و لیدار با استفاده از تکنیک‌های یادگیری بدون نظارت[6] است تا مشخص شود چگونه می‌توان با این رویکرد، پیشنهادات ناحیه‌ای دقیق‌تر و کارآمدتر ارائه داد.

در مراحل اولیه توسعه خودروهای خودران، تمرکز اصلی بر استخراج مستقیم ویژگی‌ها از تصاویر و سپس تشخیص اشیاء بود. این روش‌ها، که عمدتاً بر پایه تکنیک‌های سنتی پردازش تصویر[7] بنا شده بودند، در شرایطی که تصاویر دارای نویز کم و کنتراست بالا بودند، نتایج قابل قبولی ارائه می‌کردند. اما در

[1] Driverless Cars
[2] Object Detection
[3] YOLO (You Only Look Once)
[4] Faster R-CNN (Region-Based Convolutional Neural Network)
[5] Region Proposal
[6] Unsupervised Learning
[7] Image Processing

محیط‌های پیچیده‌تر و چالش‌برانگیز، این روش‌ها کارایی لازم را نداشتند و دقت پایینی در تشخیص شیء ارائه می‌دادند. محدودیت‌های این رویکردها در درک عمیق محتوای تصویر و تمایز بین اشیاء مختلف در سناریوهای متنوع بود.

با ظهور مدل‌هایی مانند آر-سی ان ان[۸] [۳]، تحول بزرگی در تکنیک‌های تشخیص شیء رخ داد. این مدل‌ها با پیاده‌سازی الگوریتمهای پیشنهاد ناحیه، توانستند مناطقی از تصاویر را که احتمالاً حاوی اشیاء هستند، شناسایی کرده و برای تحلیل بیشتر و استخراج ویژگی‌ها اولویت‌بندی کنند. این رویکرد نه تنها دقت تشخیص شیء را به‌طور قابل توجهی بهبود بخشید، بلکه به کاهش زمان پردازش نیز کمک کرد. پیشنهاد ناحیه به عنوان یک مفهوم کلیدی در بسیاری از مطالعات بعدی و توسعه مدل‌های پیشرفته‌تر تشخیص شیء مانند فَستر آر-سی ان ان و یولو استفاده شد که هر یک به نوبه خود نوآوری‌هایی را در این حوزه ارائه دادند.

امروزه، با دسترسی به سنسورهای پیشرفته‌تری مانند لیدار، که اطلاعات مکانی سه‌بعدی دقیقی را فراهم می‌کند، فرصت‌های جدیدی برای ادغام داده‌های چندگانه به منظور افزایش دقت و کارایی سامانه‌های تشخیص شیء فراهم شده است. این پروژه قصد دارد با ترکیب داده‌های دوربین و لیدار و استفاده از الگوریتم‌های یادگیری بدون نظارت، عملکرد روش‌های پیشنهادی در ارائه پیشنهادات ناحیه را مورد بررسی و ارزیابی قرار دهد تا مشخص شود چگونه این رویکرد می‌تواند به تشخیص‌های دقیق‌تر و سریع‌تری در محیط‌های مختلف و پویا منجر شود.

۱-۲ اهمیت و ضرورت پروژه

با وجود پیشرفت‌های قابل توجه در فناوری‌های خودروهای خودران و تشخیص شیء، همچنان چالش‌های عمده‌ای پابرجاست. یکی از مهم‌ترین این چالش‌ها، جمع‌آوری داده‌های مورد نیاز برای تشخیص شیء است که نیازمند فرآیند دستی و زمان‌بری برای پیدا کردن و علامت‌گذاری اشیاء مهم در تصاویر است. این فرآیند که شامل شناسایی دقیق موقعیت اشیاء و برچسب‌زنی آن‌ها به منظور آموزش مدل‌های تشخیص شیء است، نه تنها هزینه‌بر است، بلکه به منابع انسانی قابل توجهی نیز نیاز دارد. این امر موجب می‌شود که بهره‌وری و سرعت پیشرفت در این زمینه تحت تأثیر قرار گیرد.

در این میان، یادگیری بدون نظارت می‌تواند راهکاری ارزشمند برای کاهش نیاز به داده‌های برچسب‌دار فراهم آورد. این شیوه یادگیری به مدل‌های ماشینی امکان می‌دهد تا از داده‌های بدون برچسب برای کشف الگوها و ویژگی‌های پنهان در داده‌ها استفاده کنند. با پیاده‌سازی این روش‌ها، امکان پردازش و

[8] R-CNN (Region-Based Convolutional Neural Network)

تحلیل مجموعه‌های داده‌ای که در آن‌ها اشیاء به صورت دستی علامت‌گذاری نشده‌اند، فراهم می‌شود، که می‌تواند به کاهش هزینه‌ها و افزایش سرعت تحقیقات کمک کند. علاوه بر این، ادغام داده‌های دوبعدی و سه‌بعدی حاصل از دوربین‌ها و سنسورهای لیدار، دیدگاه جامع‌تر و دقیق‌تری از محیط را فراهم می‌آورد. این ترکیب اطلاعات، با افزایش دقت شناسایی و کاهش خطاها در تشخیص شیء، به فهم بهتر و دقیق‌تر از موقعیت‌های مختلف و پویای محیطی منجر می‌شود. استفاده از داده‌های مکمل از هر دو سنسور نه تنها کیفیت تشخیص شیء را بهبود می‌بخشد، بلکه در شرایط محیطی مختلف مانند نور کم یا مواقعی که دید دوربین محدود است، اثربخشی سامانه را افزایش می‌دهد.

این پروژه با هدف بررسی و ارزیابی اثربخشی ادغام بدون نظارت داده‌های دوبعدی و سه‌بعدی در پیشنهاد ناحیه مدل‌های تشخیص شیء، تلاش می‌کند تا به درک بهتری از عملکرد این رویکردها در محیط‌های واقعی دست یابد. این پروژه صرفاً به پیاده‌سازی و ارزیابی یک مدل تجربی می‌پردازد تا تأثیرات ادغام داده‌ها را در بهبود قابلیت‌های تشخیص و کاهش خطاها بررسی کند. این بررسی به ما امکان می‌دهد تا درکی عمیق‌تر از چالش‌ها و محدودیت‌های موجود در استفاده از سنسورهای مختلف در تشخیص شیء به دست آوریم، و این دانش می‌تواند در آینده به بهبود فناوری‌ها و رویکردهای موجود کمک کند.

۱-۳ اهداف پروژه

هدف اصلی این پروژه، ارزیابی عملکرد ادغام داده‌های دوربین و لیدار با استفاده از یادگیری بدون نظارت در فرآیند پیشنهاد ناحیه است. این پروژه به دنبال آن است که میزان تأثیرگذاری این تکنیک‌ها بر سرعت و دقت پردازش را مورد سنجش قرار دهد. به طور خاص، بررسی خواهد شد که چگونه ادغام داده‌ها می‌تواند در زمان لازم برای پردازش و دقت در شناسایی مناطق حاوی اشیاء مؤثر واقع شود. همچنین، پروژه معطوف به این است که ارزیابی کند چه درصدی از مناطق پیشنهاد شده واقعاً حاوی اشیاء مهم هستند و چه تعدادی از این پیشنهادات به اشتباه انجام شده‌اند، که می‌تواند شاخصی از کیفیت و کارایی مدل پیشنهادی باشد.

علاوه بر این، این پروژه به بررسی می‌پردازد که کدام یک از مدل‌های یادگیری بدون نظارت می‌توانند در فرآیند تشخیص و پیشنهاد مناطق مؤثرتر عمل کنند. این بخش از پروژه به ارزیابی عملکرد مدل‌های مختلف یادگیری بدون نظارت می‌پردازد تا مشخص شود کدام یک قادر به ارائه دقت و کارایی بالاتر در شناسایی مناطق حاوی اشیاء هستند.

۱-۴ ساختار گزارش

این گزارش در پنج فصل تنظیم شده است. فصل اول به معرفی کلی پروژه، بیان مسئله، اهمیت و ضرورت تحقیق، و اهداف پروژه می‌پردازد. فصل دوم به بررسی ادبیات و کارهای پیشین اختصاص دارد، که در آن مفاهیم کلیدی، تعاریف و پیش‌زمینه‌های نظری مرتبط با موضوع تحقیق به‌طور مفصل تشریح می‌شود. همچنین، مدل‌ها و تکنولوژی‌های مرتبط با پروژه معرفی و بررسی می‌شوند. فصل سوم روش‌ها و فرآیندهای به‌کارگرفته‌شده در پروژه را توضیح می‌دهد، که شامل طراحی مدل‌های پیشنهادی و تکنیک‌های استفاده‌شده برای پیاده‌سازی مسئله است. فصل چهارم به ارزیابی و تحلیل نتایج آزمایش‌ها می‌پردازد، که در آن کیفیت و کارایی مدل‌های طراحی‌شده مورد بررسی و تجزیه و تحلیل قرار می‌گیرد. در نهایت، فصل پنجم به جمع‌بندی کلی نکات مهم گزارش و ارائه پیشنهادهایی برای کارهای آتی می‌پردازد، که می‌تواند شامل اقدامات لازم برای بهبود عملکرد و کارایی مدل‌ها در آینده باشد.

فصل دوم

مروری بر ادبیات و پیشینه تحقیق

۲-۱ مقدمه

در سال‌های اخیر، پیشرفت‌های قابل توجهی در حوزه هوش مصنوعی و یادگیری عمیق صورت گرفته است که تأثیر بسزایی بر تکنولوژی‌های مختلف از جمله خودروهای خودران داشته است. یکی از چالش‌های اصلی در توسعه خودروهای خودران، درک دقیق محیط پیرامون و تشخیص اشیاء موجود در آن است. برای رسیدن به این هدف، ترکیب داده‌های حاصل از سنسورهای مختلف مانند دوربین و لیدار اهمیت ویژه‌ای دارد. در این فصل، به بررسی مفاهیم اساسی و کارهای پیشین مرتبط با خودروهای خودران، تشخیص شیء، پیشنهاد ناحیه و ادغام داده‌های دوربین و لیدار می‌پردازیم.

۲-۲ خودروهای خودران

خودروهای خودران، وسایل نقلیه‌ای هستند که قادر به حرکت و انجام وظایف رانندگی بدون نیاز به دخالت انسانی می‌باشند [۴]. این خودروها از دهه‌های گذشته موضوع پژوهش و توسعه بوده‌اند. اولین تلاش‌ها برای ایجاد خودروهای خودران به دهه ۱۹۸۰ میلادی بازمی‌گردد، زمانی که دانشگاه‌ها و مراکز تحقیقاتی شروع به آزمایش خودروهایی با قابلیت‌های خودکار کردند [۵]. یکی از پروژه‌های پیشگام در این زمینه، پروژه آلوین[1] بود که توسط دانشگاه کارنگی ملون توسعه یافت. با پیشرفت فناوری‌های سنسورها، پردازشگرها و الگوریتم‌های هوش مصنوعی، خودروهای خودران از مرحله آزمایشگاهی به واقعیت نزدیک‌تر شدند.

برای عملکرد صحیح، خودروهای خودران به ترکیبی از سامانه‌ها و اجزا نیاز دارند. این اجزا شامل سنسورها برای جمع‌آوری اطلاعات محیطی، سامانه‌های پردازش برای تحلیل داده‌ها، الگوریتم‌های تصمیم‌گیری برای برنامه‌ریزی مسیر و اقدامات، و سامانه‌های کنترلی برای اجرای دستورات هستند. یکی از مهم‌ترین وظایف در خودروهای خودران، تشخیص اشیاء و موانع در محیط پیرامون است که به خودرو امکان می‌دهد تا به‌طور ایمن و کارآمد حرکت کند.

[1] ALVINN (Autonomous Land Vehicle In a Neural Network)

۲-۳ سنسورهای مورد استفاده در خودروهای خودران

برای درک محیط پیرامون، خودروهای خودران از سنسورهای متعددی استفاده می‌کنند که هر کدام اطلاعات متفاوت و مکملی را فراهم می‌کنند این سنسورها شامل رادار، لیدار، دوربین، سنسورهای اولتراسونیک و سامانه‌های موقعیت‌یابی مانند جی پی اس[۲] هستند. هر یک از این سنسورها مزایا و محدودیت‌های خاص خود را دارند و با ترکیب داده‌های آن‌ها، خودرو می‌تواند تصویر جامعی از محیط خود داشته باشد.

در میان این سنسورها، دوربین و لیدار بیشترین کاربرد را در تشخیص شیء دارند. دوربین‌ها اطلاعات غنی بصری مانند رنگ، بافت و شکل را فراهم می‌کنند، در حالی که لیدار اطلاعات سه‌بعدی دقیق از فاصله و عمق اشیاء را ارائه می‌دهد.

لیدار[۳] یک تکنولوژی سنجش از راه دور است که از پالس‌های نور لیزری برای اندازه‌گیری فواصل تا اشیاء استفاده می‌کند [۶]. این سامانه با ارسال پالس‌های لیزری و دریافت بازتاب آن‌ها، فاصله تا اشیاء را با دقت بالا محاسبه می‌کند. با چرخش یا حرکت سنسور، لیدار قادر است یک نمای سه‌بعدی دقیق از محیط اطراف ایجاد کند که به آن ابر نقطه[۴] گفته می‌شود. مزایای استفاده از لیدار شامل دقت بالا در اندازه‌گیری فاصله، عملکرد خوب در شرایط نوری مختلف، حتی در تاریکی و شب، و قابلیت تشخیص اشیاء در محیط‌های پیچیده است. برخلاف بسیاری از سنسورهای دیگر، لیدار می‌تواند در شرایط نور کم یا بدون نور نیز داده‌های با کیفیت جمع‌آوری کند. با این حال، معایبی مانند هزینه بالا، که آن را به یک سنسور گران‌قیمت تبدیل می‌کند، حساسیت به شرایط آب و هوایی مانند باران و مه، و حجم بالای داده‌های تولید شده نیز وجود دارد.

دوربین‌ها سنسورهایی هستند که تصاویر دوبعدی از محیط را ضبط می‌کنند. این تصاویر حاوی اطلاعات غنی از جمله رنگ، بافت، شکل و سایر ویژگی‌های بصری هستند که برای تشخیص اشیاء و درک صحنه بسیار مفید می‌باشند [۷]. مزایای استفاده از دوربین‌ها شامل هزینه پایین، وضوح بالا و قابلیت تشخیص ویژگی‌های بصری است. با این حال، عملکرد دوربین‌ها به شدت به شرایط نوری وابسته است و در شرایط نور کم یا تغییرات شدید نور ممکن است کارایی کاهش یابد.

[2] GPS (Global Positioning System)

[3] LIDAR (Light Detection and Ranging)

[4] Point Cloud

ترکیب داده‌های حاصل از سنسورهای مختلف، که به آن ادغام داده‌ها[5] گفته می‌شود، نقش مهمی در بهبود دقت و قابلیت اطمینان سامانه‌های تشخیص شیء در خودروهای خودران دارد. ادغام داده‌های لیدار و دوربین می‌تواند مزایای هر دو سنسور را به ارمغان بیاورد و نقاط ضعف آن‌ها را جبران کند. با ادغام اطلاعات سه‌بعدی دقیق از لیدار و ویژگی‌های بصری غنی از دوربین، سامانه می‌تواند درک بهتری از محیط داشته باشد و دقت تشخیص اشیاء را افزایش دهد.

۲-۴ تشخیص شیء در خودروهای خودران

تشخیص شیء یکی از وظایف اصلی در بینایی ماشین است که به شناسایی و تعیین موقعیت اشیاء مختلف در یک تصویر یا ویدئو می‌پردازد [۸]. این فرآیند شامل دو بخش اصلی است: شناسایی کلاس شیء و تعیین محل دقیق آن در تصویر.

تاریخچه تشخیص شیء به اوایل دهه ۲۰۰۰ میلادی بازمی‌گردد، زمانی که الگوریتم‌های مبتنی بر ویژگی‌های دستی مانند روش ویولا-جونز[6] [۹] برای تشخیص اشیاء ساده به‌کار گرفته شدند. با پیشرفت یادگیری ماشین و گسترش شبکه‌های عصبی کانولوشنال (سی ان ان‌ها)[7]، مدل‌های شناسایی کلاس شیء مبتنی بر این شبکه‌ها مانند ایمیجنت[8] [۱۰] به‌وجود آمدند که دقت و کارایی تشخیص شیء را به‌طور قابل توجهی افزایش دادند.

روش‌های مدرن تشخیص شیء به‌طور کلی به دو دسته اصلی تقسیم می‌شوند: روش‌های تک‌مرحله‌ای و روش‌های دومرحله‌ای [۸]. در روش‌های دومرحله‌ای مانند فَستر آر-سی ان ان [۲]، ابتدا مناطقی از تصویر که احتمالاً حاوی اشیاء هستند شناسایی می‌شوند (پیشنهاد ناحیه)، و سپس در مرحله دوم، این مناطق توسط یک طبقه‌بند تحلیل و تشخیص داده می‌شوند. این رویکرد به دلیل دقت بالا در تشخیص اشیاء، به‌ویژه در کاربردهایی که دقت اهمیت بالایی دارد، مورد توجه قرار گرفته است.

در مقابل، روش‌های تک‌مرحله‌ای مانند یولو [۱] و اس اس دی[9] [۱۱] تشخیص و طبقه‌بندی اشیاء را به‌طور همزمان و در یک مرحله انجام می‌دهند. این روش‌ها به دلیل سرعت بالای پردازش، برای کاربردهای بلادرنگ مناسب هستند، هرچند ممکن است دقت کمتری نسبت به روش‌های دومرحله‌ای

[5] Data Fusion
[6] Viola-Jones
[7] CNNs (Convolutional Neural Networks)
[8] ImageNet
[9] SSD (Single Shot MultiBox Detector)

داشته باشند.

۲-۵ پیشنهاد ناحیه در خودروهای خودران

پیشنهاد ناحیه یک تکنیک مهم در تشخیص شیء است که به شناسایی بخش‌هایی از تصویر می‌پردازد که احتمالاً حاوی اشیاء هستند [۱۲]. این فرآیند با کاهش فضای جستجو و تمرکز بر مناطق مهم، کارایی و دقت تشخیص شیء را افزایش می‌دهد.

در مراحل اولیه، روش‌های پیشنهاد ناحیه بر پایه هیوریستیک‌ها و ویژگی‌های ساده تصویری مانند رنگ، بافت و لبه‌ها بودند [۱۳]. یکی از روش‌های معروف، روش پنجره کشویی[10] بود که در آن یک پنجره با اندازه ثابت یا متغیر بر روی تصویر حرکت داده می‌شد [۹]. این روش به دلیل تعداد زیاد پنجره‌ها، محاسبات سنگینی داشت و کارایی مناسبی نداشت.

با هدف بهبود کارایی، الگوریتم جستجوی انتخابی[11] توسط اویلیینگس و همکاران در سال ۲۰۱۳ معرفی شد [۱۲]. این الگوریتم با ترکیب مزایای تقسیم‌بندی تصویر و روش‌های هیوریستیک، تعداد مناطق پیشنهادی را کاهش داد و دقت را افزایش داد. در این روش، تصویر به سوپرپیکسل‌ها تقسیم می‌شود و سپس بر اساس معیارهای مشابهت، این سوپرپیکسل‌ها با هم ادغام می‌شوند تا مناطق بزرگ‌تری را تشکیل دهند.

با پیشرفت یادگیری عمیق، مدل‌های مبتنی بر شبکه‌های عصبی کانولوشنال معرفی شدند که بهبود قابل توجهی در دقت و کارایی تشخیص شیء داشتند. مدل آر-سی ان ان [۳] توسط گیرشیک و همکاران در سال ۲۰۱۴ معرفی شد. در این مدل، ابتدا با استفاده از الگوریتم جستجوی انتخابی، حدود ۲۰۰۰ ناحیه پیشنهادی استخراج می‌شود. سپس هر ناحیه به‌طور جداگانه توسط یک سی ان ان پردازش می‌شود تا ویژگی‌ها استخراج شوند. در نهایت، یک طبقه‌بند خطی مانند اس وی ام[12] برای طبقه‌بندی اشیاء به کار می‌رود. اگرچه آر-سی ان ان دقت بالایی داشت، اما به دلیل نیاز به پردازش هر ناحیه به‌طور جداگانه، بسیار زمان‌بر بود.

برای حل مشکل زمان پردازش، مدل فَست آر-سی ان ان[13] توسط گیرشیک در سال ۲۰۱۵ معرفی

[10] Sliding Window
[11] Selective Search
[12] SVM (Support vector machine)
[13] Fast R-CNN

شد [۱۴]. در این مدل، تصویر ورودی تنها یک بار از طریق یک سی ان ان عبور داده می‌شود و یک نقشه ویژگی تولید می‌شود. سپس نواحی پیشنهادی بر روی این نقشه ویژگی اعمال می‌شوند و با استفاده از عملیات تجمیع ناحیه مورد علاقه[۱۴]، ویژگی‌های مربوط به هر ناحیه استخراج می‌شوند. این روش زمان پردازش را به‌طور قابل توجهی کاهش داد.

مدل فَستر آر-سی ان ان توسط رن و همکاران در سال ۲۰۱۵ معرفی شد [۲]. این مدل با ادغام یک شبکه پیشنهاد ناحیه (ار پی ان)[۱۵] در ساختار سی ان ان، توانست فرآیند پیشنهاد ناحیه را به‌صورت انتها به انتها و با کارایی بالا انجام دهد. شبکه پیشنهاد ناحیه به‌طور مستقیم از نقشه‌های ویژگی سی ان ان استفاده می‌کند تا نواحی پیشنهادی را تولید کند، که منجر به افزایش سرعت و دقت تشخیص شیء شد.

در مقابل، مدل‌های تک‌مرحله‌ای مانند یولو [۱] و اس اس دی [۱۱] از پیشنهاد ناحیه استفاده نمی‌کنند و به‌جای آن، تصویر را به شبکه‌ای از سلول‌ها تقسیم می‌کنند و در هر سلول به پیش‌بینی اشیاء می‌پردازند. این روش‌ها به دلیل سرعت بالاتر، برای کاربردهای بلادرنگ مناسب هستند.

در سال‌های اخیر، مدل‌های مبتنی بر مبدل‌ها [۱۵] مانند دیتر[۱۶] [۱۶] معرفی شده‌اند که از مکانیزم توجه برای تشخیص شیء استفاده می‌کنند و نیاز به پیشنهاد ناحیه را حذف می‌کنند. با این حال، تحقیقات بعدی نشان داده‌اند که افزودن مکانیزم پیشنهاد ناحیه می‌تواند به بهبود بار محاسباتی و دقت در این مدل‌ها منجر شود [۱۷].

۲-۶ شبکه‌های عصبی و کاربرد آن‌ها در تشخیص شیء

با پیشرفت یادگیری ماشین و به‌ویژه یادگیری عمیق، شبکه‌های عصبی به ابزارهای قدرتمندی برای پردازش و تحلیل داده‌های پیچیده تبدیل شده‌اند. در حوزه تشخیص شیء، شبکه‌های عصبی و به‌خصوص شبکه‌های عصبی کانولوشنال نقش اساسی ایفا می‌کنند. در این بخش، به بررسی شبکه‌های عصبی، شبکه‌های عصبی کانولوشنال و مدل‌های مهمی که برای استخراج ویژگی و تشخیص شیء در خودروهای خودران استفاده می‌شوند، می‌پردازیم.

شبکه‌های عصبی مصنوعی[۱۷] مدل‌های محاسباتی الهام‌گرفته از ساختار و عملکرد شبکه‌های عصبی

[14] ROI Pooling
[15] RPN (Region Proposal Network)
[16] DETR
[17] ANN (Artificial Neural Networks)

بیولوژیکی هستند. این شبکه‌ها از لایه‌های متعددی از نورون‌ها تشکیل شده‌اند که با وزن‌های قابل تنظیم به یکدیگر متصل هستند. با آموزش مناسب، شبکه‌های عصبی می‌توانند الگوها و روابط پیچیده در داده‌ها را یاد بگیرند و برای وظایفی مانند طبقه‌بندی، پیش‌بینی و تشخیص الگوها مورد استفاده قرار گیرند.

۲-۶-۱ شبکه‌های عصبی کانولوشنال

شبکه‌های عصبی کانولوشنال نوع خاصی از شبکه‌های عصبی هستند که برای پردازش داده‌های دارای ساختار شبکه‌ای مانند تصاویر طراحی شده‌اند [۱۸]. سی ان ان‌ها از لایه‌های کانولوشن استفاده می‌کنند که فیلترهایی را بر روی ورودی اعمال می‌کنند تا ویژگی‌های محلی مانند لبه‌ها، بافت‌ها و الگوهای پیچیده‌تر را استخراج کنند.

ویژگی مهم سی ان ان‌ها در قابلیت یادگیری ویژگی‌های سلسله‌مراتبی از داده‌ها است؛ لایه‌های ابتدایی ویژگی‌های ساده را می‌آموزند، در حالی که لایه‌های عمیق‌تر ویژگی‌های سطح بالاتری را استخراج می‌کنند. این خصوصیت باعث شده است که سی ان ان‌ها ابزار بسیار مؤثری برای وظایفی مانند تشخیص شیء، بخش‌بندی تصویر و تشخیص چهره باشند [۱۰].

در مدل‌های تشخیص شیء مورد استفاده در خودروهای خودران، استخراج ویژگی‌های مؤثر از تصاویر نقش حیاتی دارد. این مدل‌ها برای دستیابی به دقت و کارایی بالا، نیازمند شبکه‌هایی هستند که بتوانند ویژگی‌های غنی و معناداری از تصاویر پیچیده و پویا استخراج کنند. در ادامه، به معرفی و توضیح معماری برخی از مهم‌ترین شبکه‌های عصبی کانولوشنال که به‌طور گسترده در تشخیص شیء و به‌ویژه در خودروهای خودران استفاده می‌شوند، می‌پردازیم.

۲-۶-۲ مدل رزنت

رزنت[۱۸] توسط هه و همکاران در سال ۲۰۱۶ معرفی شد [۱۹]. این مدل با معرفی اتصالات باقی‌مانده[۱۹] توانست مشکل ناپدید شدن گرادیان[۲۰] در شبکه‌های عمیق را برطرف کند و شبکه‌های بسیار عمیق با بیش از ۱۵۰ لایه را ممکن سازد.

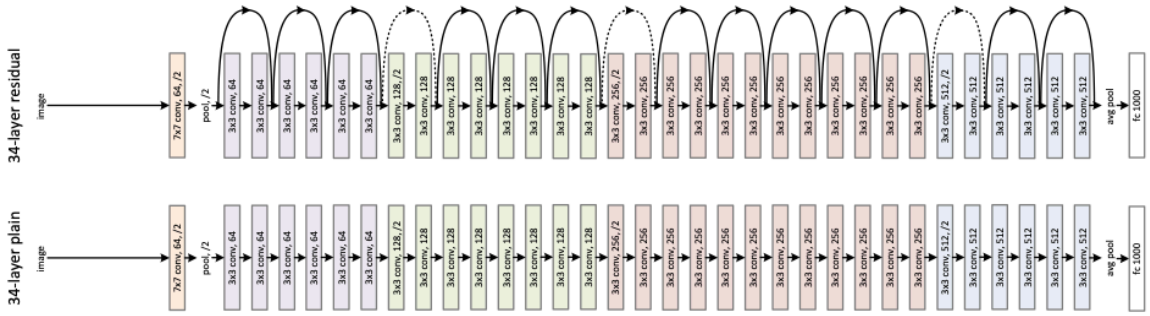
معماری رزنت شامل بلوک‌های باقی‌مانده است که در آن‌ها ورودی لایه به خروجی اضافه می‌شود.

[18] ResNet (Residual Neural Network)

[19] Residual Connections

[20] Vanishing Gradient Problem

این ساختار به شبکه امکان می‌دهد تا به‌طور مؤثری ویژگی‌های پیچیده را یاد بگیرد و دقت بالایی در وظایف بینایی ماشین ارائه دهد. معماری کلی مدل رزنت در شکل ۲-۱ نشان داده شده است.



شکل ۲-۱: معماری کلی مدل رزنت که شامل بلوک‌های باقی‌مانده برای حل مشکل ناپدید شدن گرادیان است.

در بلوک‌های باقی‌مانده، خروجی لایه‌های کانولوشن با ورودی اولیه جمع می‌شود. این اتصالات میانبر[۲۱] به شبکه اجازه می‌دهند تا گرادیان‌ها را به لایه‌های ابتدایی‌تر منتقل کند و یادگیری را در شبکه‌های عمیق تسهیل کند. این ویژگی باعث می‌شود که رزنت بتواند شبکه‌هایی با صدها لایه را آموزش دهد بدون اینکه دچار مشکل ناپدید شدن یا انفجار گرادیان شود.

در بسیاری از مدل‌های تشخیص شیء مانند فَستر آر-سی ان ان، از رزنت به عنوان ستون فقرات[۲۲] برای استخراج ویژگی استفاده می‌شود. این مدل با ارائه نقشه‌های ویژگی غنی، به بهبود دقت تشخیص شیء کمک می‌کند. در کاربردهای خودروهای خودران، استفاده از رزنت به عنوان استخراج‌کننده ویژگی، امکان تشخیص دقیق‌تر اشیاء در تصاویر پیچیده و پویا را فراهم می‌کند.

۲-۶-۳ مدل وی جی جی

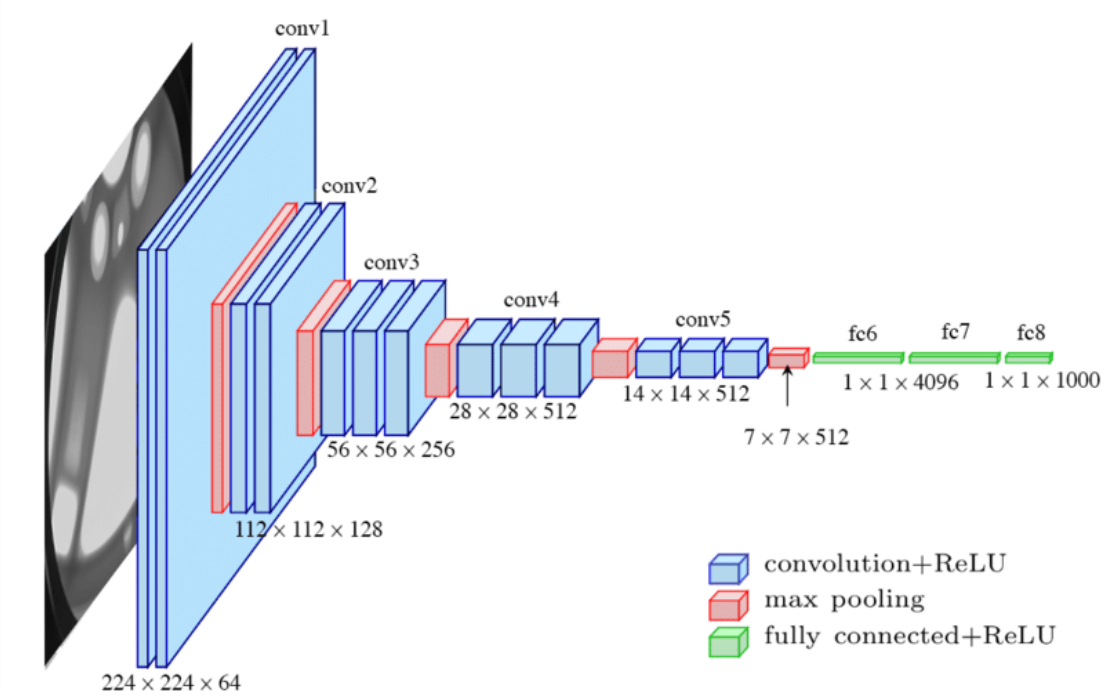
وی جی جی[۲۳] یکی از معماری‌های مهم شبکه‌های عصبی کانولوشنال است که توسط سیمونیان و زیسرمن در سال ۲۰۱۴ معرفی شد [۲۰]. این مدل به دلیل سادگی ساختار و عملکرد بالایش در مسابقات شناسایی تصویر ایمیج‌نت شهرت یافت. ایده اصلی پشت وی جی جی استفاده از فیلترهای کانولوشن بسیار کوچک با اندازه ۳ × ۳ در سراسر شبکه و افزایش عمق شبکه با افزودن لایه‌های کانولوشن بیشتر است.

معماری وی جی جی شامل لایه‌های کانولوشن متوالی است که هر یک توسط تابع فعال‌سازی واحد

[21] Skip Connections
[22] Backbone
[23] VGG

یکسو شده‌ی خطی[۲۴] دنبال می‌شوند. پس از چندین لایه کانولوشن، یک لایه ادغام ماکسیمم[۲۵] برای کاهش ابعاد مکانی به کار می‌رود. شبکه با لایه‌های تماماً متصل[۲۶] برای طبقه‌بندی نهایی به پایان می‌رسد. دو نسخه معروف از این مدل، وی جی جی-۱۶ و وی جی جی-۱۹ هستند که به ترتیب شامل ۱۶ و ۱۹ لایه‌های وزن‌دار (کانولوشن و تماماً متصل) می‌باشند.

استفاده از فیلترهای کانولوشن کوچک به شبکه اجازه می‌دهد تا میدان دید مؤثری داشته باشد و در عین حال تعداد پارامترها را نسبت به فیلترهای بزرگ‌تر کاهش دهد. این طراحی امکان یادگیری ویژگی‌های پیچیده را فراهم می‌کند و در عین حال بهره‌وری محاسباتی را حفظ می‌نماید. معماری کلی مدل وی جی جی در شکل ۲-۲ نشان داده شده است.



شکل ۲-۲: معماری مدل وی جی جی که از فیلترهای کانولوشن و لایه‌های ادغام ماکسیمم استفاده می‌کند.

ویژگی بارز وی جی جی سادگی و یکنواختی معماری آن است. استفاده از فیلترهای کانولوشن با اندازه ثابت و ساختار بلوکی باعث شده است که این مدل به راحتی قابل پیاده‌سازی و توسعه باشد. علیرغم عمق شبکه، به دلیل استفاده از توابع فعال‌سازی ReLU، مشکل ناپدید شدن گرادیان به‌خوبی کنترل

[24] ReLU
[25] Max Pooling
[26] Fully Connected

می‌شود و آموزش شبکه امکان‌پذیر می‌گردد.

در زمینه تشخیص شیء، وی جی جی به‌طور گسترده به عنوان شبکه پایه برای استخراج ویژگی مورد استفاده قرار گرفته است. مدل‌هایی مانند آر-سی ان ان [۳]، فَست آر-سی ان ان [۱۴] و نسخه‌های اولیه فَستر آر-سی ان ان [۲] از وی جی جی-۱۶ به عنوان ستون فقرات بهره برده‌اند. توانایی وی جی جی در استخراج ویژگی‌های غنی و سلسله‌مراتبی از تصاویر، آن را برای وظایفی که نیاز به اطلاعات مکانی دقیق دارند، مناسب می‌سازد.

در کاربردهای مرتبط با خودروهای خودران، وی جی جی در سیستم‌های تشخیص شیء برای شناسایی و مکان‌یابی دقیق اشیائی مانند عابران پیاده، خودروها و علائم راهنمایی استفاده شده است. با این حال، به دلیل هزینه محاسباتی و نیاز به حافظه بالاتر نسبت به مدل‌های جدیدتر مانند رزنت، استفاده از وی جی جی در کاربردهای عملی کاهش یافته است. مدل‌های مدرن‌تر با معرفی اتصالات باقی‌مانده و بهینه‌سازی‌های معماری، عملکرد بهتری از نظر دقت و کارایی ارائه می‌دهند.

به طور خلاصه، مدل وی جی جی یک معماری پایه‌ای و تأثیرگذار در توسعه شبکه‌های عصبی کانولوشنال عمیق است. اصول طراحی آن بر بسیاری از مدل‌های بعدی تأثیر گذاشته و استفاده از آن در استخراج ویژگی به پیشرفت‌های قابل توجهی در وظایف تشخیص و شناسایی اشیاء منجر شده است.

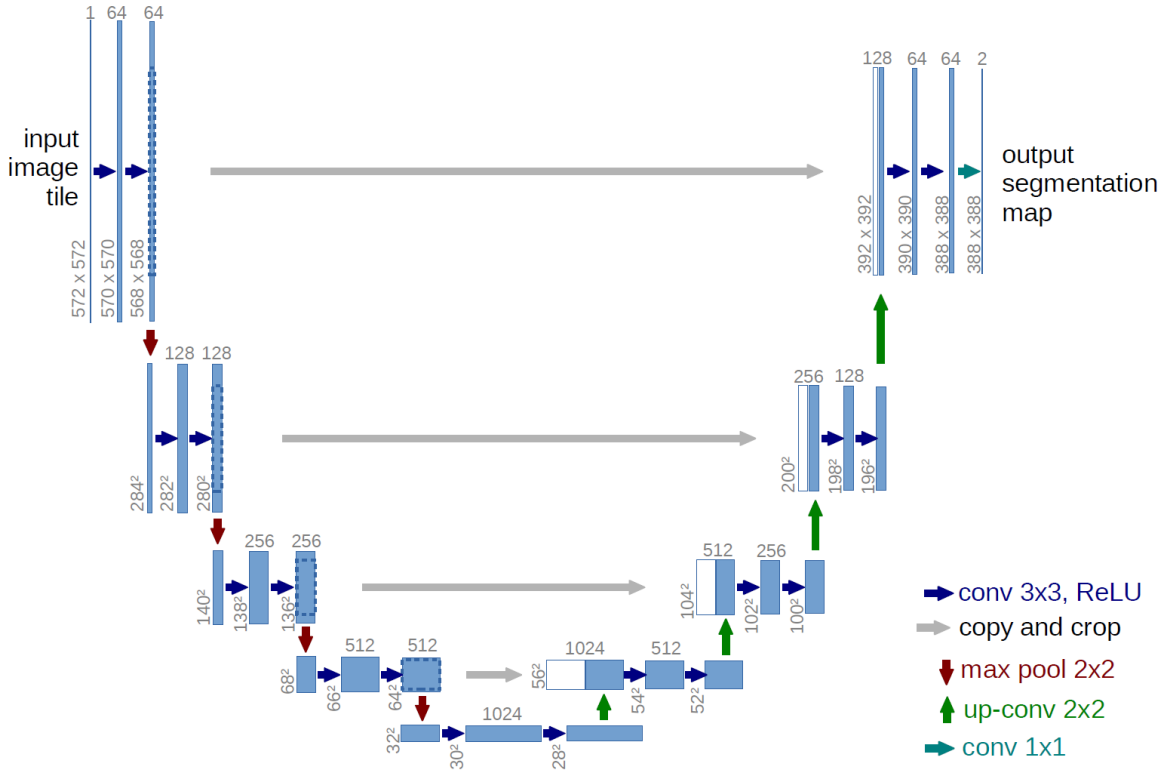
۲-۶-۴ مدل یو-نت

یو-نت[۲۷] یک معماری شبکه عصبی کانولوشنال است که برای بخش‌بندی تصویر طراحی شده است [۲۱]. این مدل در ابتدا برای کاربردهای پزشکی معرفی شد، اما به دلیل عملکرد بالایش در بخش‌بندی، در حوزه‌های دیگر نیز مورد استفاده قرار گرفت.

معماری یو-نت شامل دو مسیر متقارن است: یک مسیر انقباضی[۲۸] و یک مسیر گسترشی[۲۹]. معماری کلی یو-نت در شکل ۲-۳ نشان داده شده است.

در مسیر انقباضی، شبکه با استفاده از لایه‌های کانولوشن و نمونه‌برداری پایین[۳۰]، ویژگی‌های مکانی و متنی را استخراج می‌کند. در هر مرحله، ابعاد مکانی کاهش می‌یابد و تعداد فیلترها افزایش می‌یابد تا ویژگی‌های سطح بالا استخراج شوند.

[27] U-Net
[28] Contracting Path
[29] Expanding Path
[30] Downsampling



شکل ۲-۳: معماری مدل یو-نت که شامل مسیر انقباضی و مسیر گسترشی برای بخش‌بندی دقیق تصویر است.

در مسیر گسترشی، شبکه با استفاده از لایه‌های کانولوشن و نمونه‌برداری بالا[۳۱]، وضوح مکانی را بازیابی می‌کند. در هر مرحله، ویژگی‌های استخراج‌شده در مسیر انقباضی با ویژگی‌های متناظر در مسیر گسترشی از طریق اتصالات میانبر[۳۲] ترکیب می‌شوند. این اتصالات به شبکه اجازه می‌دهند تا جزئیات مکانی دقیق را حفظ کند و به دقت بالایی در بخش‌بندی تصویر دست یابد.

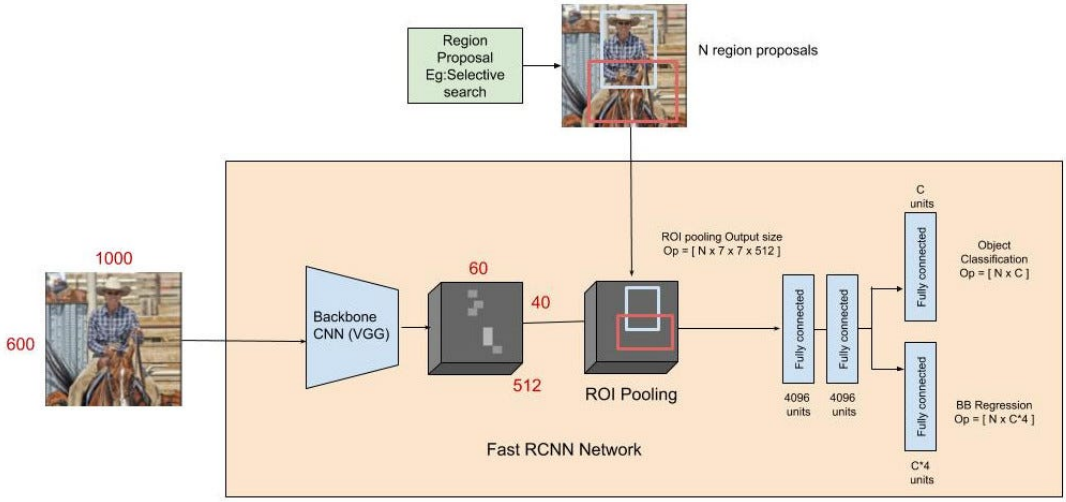
در پروژه‌های مرتبط با خودروهای خودران، یو-نت برای وظایفی مانند بخش‌بندی جاده، خطوط عبور و عابران پیاده به کار می‌رود. استفاده از این مدل به خودروهای خودران امکان می‌دهد تا محیط پیرامون خود را با دقت بالاتری درک کنند و تصمیمات ایمن‌تری اتخاذ کنند.

۲-۶-۵ مدل فَستر آر-سی ان ان

فَستر آر-سی ان ان یکی از مدل‌های دومرحله‌ای معروف در تشخیص شیء است که به‌طور گسترده در کاربردهای خودروهای خودران استفاده می‌شود. این مدل با معرفی شبکه پیشنهاد ناحیه، فرآیند پیشنهاد

[31] Upsampling
[32] Skip Connections

ناحیه را به‌صورت انتها به انتها و کارآمد انجام می‌دهد. معماری کلی فَستر آر-سی ان ان در شکل ۲-۴ نشان داده شده است.



شکل ۲-۴: معماری مدل فَستر آر-سی ان ان که شامل شبکه پیشنهاد ناحیه و شبکه تشخیص است.

معماری فَستر آر-سی ان ان شامل سه بخش اصلی است: ابتدا، تصویر ورودی از طریق یک شبکه عصبی کانولوشنال عمیق مانند رزنت یا وی جی جی عبور می‌کند تا نقشه‌های ویژگی استخراج شوند که اطلاعات مکانی و متنی غنی را برای مراحل بعدی فراهم می‌کنند. سپس، شبکه پیشنهاد ناحیه بر روی نقشه‌های ویژگی عمل می‌کند تا نواحی مستطیلی که احتمالاً حاوی اشیاء هستند را پیشنهاد دهد. این شبکه با اسلاید کردن یک پنجره بر روی نقشه‌های ویژگی و استفاده از آنکرها[۳۳] با ابعاد و نسبت‌های مختلف، به پیش‌بینی احتمال وجود شیء و اصلاح مختصات ناحیه می‌پردازد. در نهایت، نواحی پیشنهادی پس از عملیات تجمیع ناحیه مورد علاقه، که اندازه نواحی را به یک اندازه ثابت تبدیل می‌کند، به شبکه تشخیص وارد می‌شوند که وظیفه طبقه‌بندی اشیاء و پیش‌بینی دقیق مختصات جعبه‌های محدودکننده را بر عهده دارد.

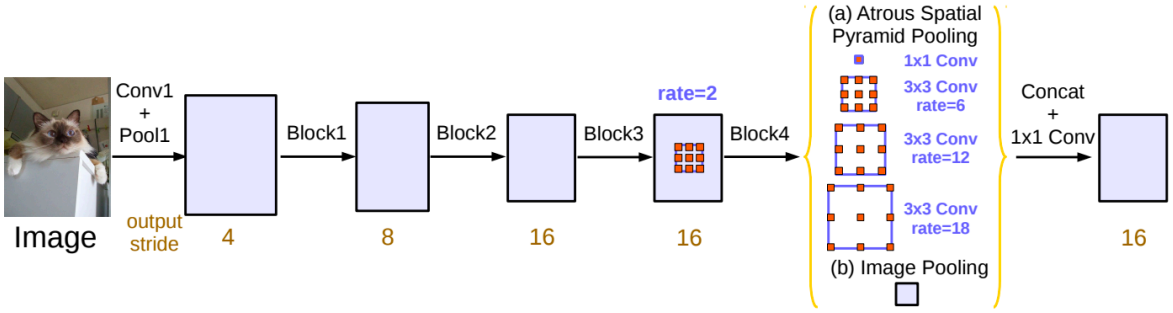
فَستر آر-سی ان ان به دلیل دقت بالا در تشخیص اشیاء و کارایی مناسب، به‌طور گسترده در سیستم‌های بینایی ماشین خودروهای خودران استفاده می‌شود. این مدل می‌تواند به‌طور مؤثر اشیائی مانند خودروهای دیگر، عابران پیاده، دوچرخه‌ها و علائم راهنمایی را تشخیص دهد و مکان‌یابی کند. در پروژه ما، از ویژگی‌های استخراج‌شده توسط فَستر آر-سی ان ان نیز استفاده می‌کنیم.

[33] Anchors

۲-۶-۶ مدل دیپ‌لب

دیپ‌لب[۳۴] یکی از پیشرفته‌ترین معماری‌ها در زمینه بخش‌بندی معنایی تصویر است که توسط چن و همکاران توسعه یافته است [۲۲]. این مدل با به‌کارگیری تکنیک‌های نوآورانه مانند کانولوشن‌های حفره‌دار[۳۵] و ای اس پی پی[۳۶]، دقت بالایی در تشخیص و بخش‌بندی اشیاء در تصاویر ارائه می‌دهد.

معماری کلی مدل دیپ‌لب‌وی۳ در شکل ۲-۵ نشان داده شده است.



شکل ۲-۵: معماری کلی مدل دیپ‌لب‌وی۳ که از کانولوشن‌های حفره‌دار و ای اس پی پی برای استخراج ویژگی‌های چندمقیاسی استفاده می‌کند.

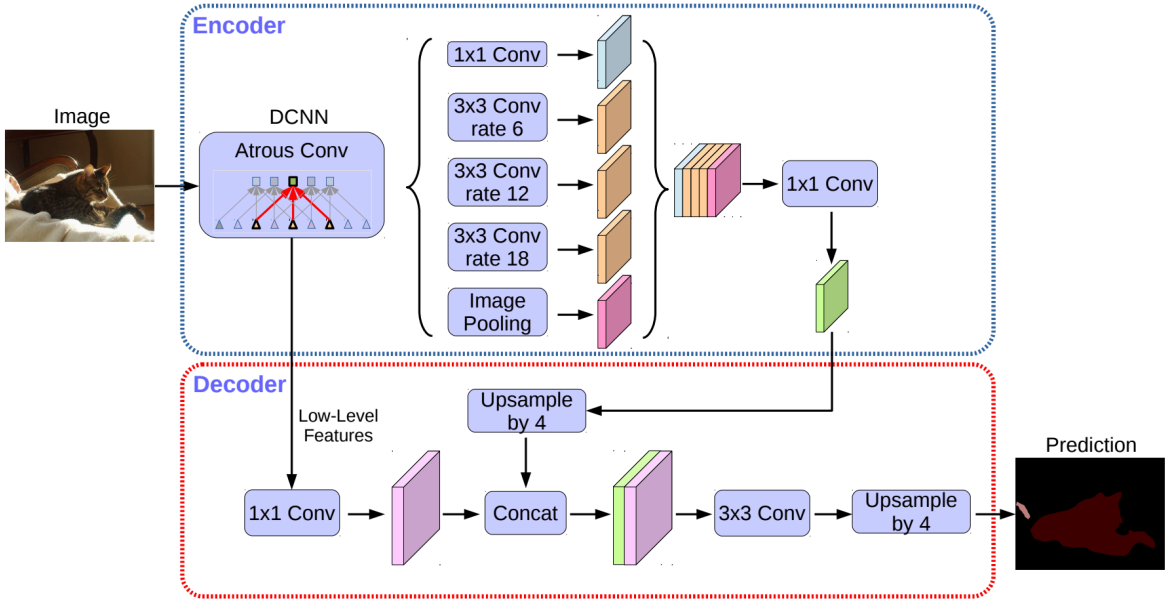
در دیپ‌لب‌وی۳، از شبکه‌های عصبی کانولوشنال عمیق مانند رزنت به عنوان بخش استخراج ویژگی استفاده می‌شود. این بخش وظیفه استخراج ویژگی‌های غنی و سلسله‌مراتبی از تصویر ورودی را بر عهده دارد. با استفاده از کانولوشن‌های حفره‌دار، میدان دید مؤثر فیلترها افزایش می‌یابد بدون اینکه رزولوشن مکانی کاهش یابد [۲۲].

یکی از اجزای کلیدی در معماری دیپ‌لب‌وی۳، استفاده از تکنیک ای اس پی پی است. ای اس پی پی با به‌کارگیری کانولوشن‌های حفره‌دار با نرخ‌های گسترش مختلف، امکان استخراج ویژگی‌ها در مقیاس‌های متفاوت را فراهم می‌کند. این تکنیک با ترکیب اطلاعات از مقیاس‌های مختلف، به مدل کمک می‌کند تا اشیاء با اندازه‌های متفاوت را به‌طور دقیق تشخیص دهد.

در نسخه دیپ‌لب‌وی۳+، یک رمزگشا به معماری اضافه شده است که با ترکیب ویژگی‌های سطح بالا و پایین، جزئیات مکانی را بهبود می‌بخشد [۲۳]. معماری دیپ‌لب‌وی۳+ در شکل ۲-۶ نشان داده شده است.

در بخش رمزگشا، ویژگی‌های با وضوح بالا از مراحل ابتدایی شبکه با ویژگی‌های سطح بالا ترکیب

[34] DeepLab
[35] Atrous Convolution
[36] Atrous Spatial Pyramid Pooling



شکل ۲-۶: معماری مدل دیپ‌لب‌وی۳+ که با افزودن بخش رمزگشا، جزئیات مکانی بخش‌بندی را بهبود می‌بخشد.

می‌شوند. این ترکیب از طریق اتصالات میانبر انجام می‌شود و به مدل امکان می‌دهد تا جزئیات مکانی دقیق‌تری را در خروجی بخش‌بندی حفظ کند.

در خودروهای خودران، تشخیص دقیق و سریع اشیاء مانند خودروهای دیگر، عابران پیاده، علائم راهنمایی و موانع ضروری است. برای دستیابی به این هدف، مدل‌های تشخیص شیء نیازمند استخراج ویژگی‌های قدرتمند از تصاویر هستند.

معماری‌هایی مانند رزنت، یو-نت و دیپ‌لب به دلیل توانایی‌شان در استخراج ویژگی‌های غنی و معنادار، به‌طور گسترده در سیستم‌های بینایی ماشین خودروهای خودران استفاده می‌شوند. این مدل‌ها با فراهم کردن نقشه‌های ویژگی با کیفیت بالا، به مدل‌های تشخیص شیء امکان می‌دهند تا با دقت بالایی اشیاء را شناسایی و مکان‌یابی کنند.

۲-۷ کارهای پیشین مرتبط با ادغام داده‌های دوربین و لیدار

ادغام داده‌های دوربین و لیدار به منظور بهبود تشخیص شیء و پیشنهاد ناحیه، در سال‌های اخیر توجه بسیاری را به خود جلب کرده است. ترکیب این دو منبع داده به دلیل مکمل بودن اطلاعات آن‌ها می‌تواند به بهبود دقت و قابلیت اطمینان سامانه‌های تشخیص شیء کمک کند. در این بخش، به بررسی تعدادی از تحقیقات انجام‌شده در این زمینه می‌پردازیم.

در پژوهشی توسط تیان و همکاران [۲۴]، روشی برای تشخیص اشیاء به صورت بدون نظارت و با استفاده از سرنخ‌های لیدار ارائه شده است. این روش از ابر نقطه‌های سه‌بعدی برای ایجاد بخش‌بندی‌هایی استفاده می‌کند که ممکن است به اشیاء تعلق داشته باشند. سپس با استفاده از یک فرآیند، به این بخش‌ها برچسب‌هایی می‌زند. این روش شامل آموزش یک شبکه برچسب‌گذاری با ویژگی‌های ادغام‌شده تصویر دوبعدی و ابر نقاط سه‌بعدی است که به تشخیص اشیاء کمک می‌کند. نتایج این پژوهش نشان می‌دهد که این رویکرد دقت قابل قبولی نسبت به روش‌های نظارت‌شده دارد. این روش با کار ما در ترکیب داده‌های دوربین و لیدار برای خوشه‌بندی اشیاء شباهت دارد و به چالش‌های مربوط به ابهام مرز اشیاء و توزیع دسته‌ای می‌پردازد.

در مطالعه‌ای دیگر، ژانگ و همکاران [۲۵] یک تکنیک برای شناسایی اشیاء با استفاده از داده‌های لیدار را معرفی کرده‌اند که نیازی به برچسب‌زنی دستی ندارد. این روش با بهره‌گیری از خوشه‌بندی نقاط، توانایی شناسایی اشیاء را در فاصله‌های دور و نزدیک با دقت بالایی دارد. آزمایش‌های انجام‌شده نشان می‌دهد که این روش عملکرد بهتری نسبت به تکنیک‌های مبتنی بر نظارت دارد. این رویکرد با روش ما در استفاده از خوشه‌بندی نقاط هم‌راستا است و به چالش‌های مربوط به شناسایی اشیاء در فواصل متفاوت می‌پردازد.

بای و همکاران [۲۶] روشی برای تشخیص سه‌بعدی اشیاء با استفاده از ترکیب داده‌های لیدار و دوربین ارائه داده‌اند. این روش با استفاده از مبدلها، دقت تشخیص را در شرایط مختلف محیطی مانند نور کم و عدم تطابق حسگرها بهبود می‌بخشد و برای شرایط واقعی طراحی شده است. این پژوهش به دلیل استفاده از ترکیب داده‌های لیدار و دوربین با کار ما شباهت دارد و نشان می‌دهد که استفاده از مدل‌های مبتنی بر مبدل می‌تواند به بهبود عملکرد در شرایط متغیر کمک کند.

لی و همکاران [۲۷] مدلی به نام دیپ‌فیوژن[۳۷] را معرفی کرده‌اند که بر ترکیب عمیق داده‌های لیدار و دوربین برای تشخیص اشیاء تمرکز دارد. این روش به طور قابل توجهی عملکرد را بر روی داده‌های خارج از توزیع بهبود می‌بخشد، که نشان‌دهنده اثربخشی آن در موقعیت‌های واقعی است که داده‌ها ممکن است به شدت متفاوت باشند. این مدل با استفاده از شبکه‌های عصبی عمیق و ادغام اطلاعات چندگانه، به بهبود دقت تشخیص در شرایط چالش‌برانگیز کمک می‌کند.

در پژوهشی دیگر، لیو و همکاران [۲۸] الگوریتمی برای تشخیص سه‌بعدی اشیاء با استفاده از ادغام

[37] DeepFusion

داده‌های دوربین و لیدار ارائه کرده‌اند. این پژوهش یک شبکه عصبی عمیق به نام فیو دی ان ان[۳۸] را معرفی می‌کند که از یک زیرشبکه تلفیق مبتنی بر توجه برای ادغام ویژگی‌های استخراج‌شده از تصاویر دوربین دوبعدی و ابر نقاط لیدار سه‌بعدی استفاده می‌کند. این الگوریتم دقت بالایی را در مجموعه داده کیتی[۳۹] نشان داده است و نشان می‌دهد که استفاده از مکانیزم‌های توجه می‌تواند به بهبود ادغام داده‌ها و دقت تشخیص کمک کند.

این پژوهش‌ها نشان می‌دهند که ادغام داده‌های لیدار و دوربین می‌تواند بهبودهای قابل توجهی در تشخیص شیء و پیشنهاد ناحیه ارائه دهد. با استفاده از تکنیک‌های مختلف مانند یادگیری بدون نظارت، خوشه‌بندی نقاط، مبدلها و مکانیزم‌های توجه، مدل‌های ارائه‌شده توانسته‌اند به دقت و کارایی بالاتری در شرایط مختلف دست یابند.

۲-۸ الگوریتم‌های خوشه‌بندی برای ادغام بدون نظارت

ادغام داده‌های لیدار و دوربین در تشخیص شیء و پیشنهاد ناحیه، نیازمند به‌کارگیری تکنیک‌های مؤثر برای تحلیل و پردازش داده‌های چندبعدی است. یکی از روش‌های اصلی در یادگیری بدون نظارت[۴۰] برای این منظور، استفاده از الگوریتم‌های خوشه‌بندی است. با توجه به اینکه ما قصد داریم از رویکردهای بدون نظارت برای استخراج ساختارهای پنهان در داده‌ها استفاده کنیم، بررسی و انتخاب الگوریتم‌های خوشه‌بندی مناسب اهمیت ویژه‌ای دارد.

الگوریتم **کا میانگین**[۴۱] یک روش ساده و کارآمد برای خوشه‌بندی داده‌ها است که با هدف تقسیم داده‌ها به K خوشه انجام می‌شود. در این الگوریتم، ابتدا K مرکز اولیه به صورت تصادفی انتخاب می‌شود. سپس هر داده به نزدیک‌ترین مرکز اختصاص می‌یابد و مراکز خوشه‌ها با محاسبه میانگین نقاط متعلق به هر خوشه به‌روزرسانی می‌شوند. این فرآیند تکرار می‌شود تا زمانی که مراکز خوشه‌ها تغییر چندانی نکنند یا به یک معیار همگرایی برسند. کا میانگین به دلیل سادگی و سرعت اجرا، یکی از پرکاربردترین الگوریتم‌های خوشه‌بندی است، اما نیازمند تعیین تعداد خوشه‌ها به صورت پیشینی است و در مواجهه با خوشه‌هایی با شکل‌های پیچیده یا چگالی‌های متفاوت عملکرد مناسبی ندارد.

[38] FuDNN
[39] KITTI
[40] Unsupervised Learning
[41] K-Means

الگوریتم **دی‌بی‌اسکن**[۴۲] یک روش خوشه‌بندی مبتنی بر چگالی است که می‌تواند خوشه‌هایی با شکل‌های دلخواه و اندازه‌های مختلف را شناسایی کند. در این الگوریتم، نقاط با چگالی بالا به عنوان هسته‌های خوشه در نظر گرفته می‌شوند و نقاطی که در همسایگی این هسته‌ها قرار دارند به خوشه مربوطه اختصاص می‌یابند. پارامترهای اصلی دی‌بی‌اسکن شامل فاصله حداکثر ϵ و حداقل تعداد نقاط در همسایگی ($MinPts$) هستند. این الگوریتم قادر است نویز و نقاط پرت را تشخیص داده و آن‌ها را به هیچ خوشه‌ای نسبت ندهد. دی‌بی‌اسکن نیازی به تعیین تعداد خوشه‌ها ندارد، اما حساسیت آن به تنظیم پارامترها ممکن است بر عملکرد آن تأثیر بگذارد.

خوشه‌بندی طیفی[۴۳] روشی است که از خواص طیفی ماتریس شباهت یا همسایگی داده‌ها برای خوشه‌بندی استفاده می‌کند. در این الگوریتم، ابتدا یک ماتریس شباهت بر اساس فاصله یا شباهت بین داده‌ها ساخته می‌شود. سپس با محاسبه مقادیر ویژه و بردارهای ویژه ماتریس لاپلاسین مرتبط، داده‌ها به فضای ویژگی جدیدی نگاشت می‌شوند. در این فضای جدید، داده‌ها با استفاده از الگوریتم‌هایی مانند کا میانگین خوشه‌بندی می‌شوند. خوشه‌بندی طیفی می‌تواند ساختارهای پیچیده و غیرخطی در داده‌ها را شناسایی کند و برای داده‌هایی با شکل‌های دلخواه مناسب است. با این حال، به دلیل نیاز به محاسبه مقادیر ویژه، از نظر محاسباتی پرهزینه است و برای مجموعه داده‌های بزرگ ممکن است کارایی مناسبی نداشته باشد.

با توجه به اینکه در پروژه ما از یادگیری بدون نظارت برای خوشه‌بندی داده‌های ادغام‌شده لیدار و دوربین استفاده می‌کنیم، بررسی این الگوریتم‌ها و انتخاب مناسب‌ترین آن‌ها برای دستیابی به نتایج مطلوب اهمیت دارد. هر یک از این الگوریتم‌ها ویژگی‌ها، مزایا و محدودیت‌های خاص خود را دارند که در زمینه تشخیص شیء و پیشنهاد ناحیه باید مورد توجه قرار گیرند. به‌عنوان مثال، کا میانگین با سادگی و سرعت بالایش، برای داده‌های با خوشه‌های کروی و همگن مناسب است، در حالی که دی‌بی‌اسکن می‌تواند با خوشه‌های با شکل‌های نامنظم و حضور نویز به خوبی کار کند. خوشه‌بندی طیفی نیز با قابلیت شناسایی ساختارهای غیرخطی، در مسائل پیچیده‌تر مفید است، هرچند هزینه محاسباتی بالایی دارد.

[42] DBSCAN

[43] Spectral Clustering

۲-۹ جمع‌بندی

در این فصل، به بررسی مفاهیم اساسی و کارهای پیشین مرتبط با خودروهای خودران، سنسورهای مورد استفاده، تشخیص شیء، پیشنهاد ناحیه و شبکه‌های عصبی پرداختیم. خودروهای خودران برای درک محیط پیرامون خود از سنسورهای مختلفی مانند دوربین و لیدار استفاده می‌کنند. ترکیب داده‌های حاصل از این سنسورها، امکان درک بهتر و دقیق‌تری از محیط را فراهم می‌سازد.

تشخیص شیء به عنوان یکی از وظایف اصلی در سیستم‌های خودروهای خودران، نقش حیاتی در ایمنی و کارایی این خودروها دارد. روش‌های مختلفی برای تشخیص شیء توسعه یافته‌اند که به دو دسته کلی **تک‌مرحله‌ای** و **دومرحله‌ای** تقسیم می‌شوند. روش‌های تک‌مرحله‌ای، به دلیل سرعت بالایشان در کاربردهای بلادرنگ مورد استفاده قرار می‌گیرند، اما دقت آن‌ها نسبت به روش‌های دومرحله‌ای کمتر است. در مقابل، روش‌های دومرحله‌ای، با استفاده از پیشنهاد ناحیه، دقت بالاتری در تشخیص اشیاء ارائه می‌دهند، اما زمان پردازش بیشتری نیاز دارند و بخش قابل توجهی از این زمان صرف تولید پیشنهادات ناحیه می‌شود.

علاوه بر این، برای آموزش این مدل‌ها، نیاز به داده‌های برچسب‌دار و بزرگ است که تهیه و برچسب‌گذاری آن‌ها هزینه‌بر و زمان‌بر است. به منظور کاهش این هزینه‌ها، استفاده از روش‌های **یادگیری بدون نظارت** مانند خوشه‌بندی مورد توجه قرار گرفته است، که نیاز به داده‌های برچسب‌دار ندارند و می‌توانند به طور خودکار الگوها و ساختارهای موجود در داده‌ها را کشف کنند.

همچنین، داده‌های دوربین به تنهایی با محدودیت‌هایی مانند حساسیت به شرایط نوری و آب‌وهوایی مواجه هستند. بنابراین، حرکت به سمت **ادغام داده‌های چند سنسور** مانند دوربین و لیدار، به منظور بهبود دقت و قابلیت اطمینان سیستم‌های تشخیص شیء، اهمیت یافته است. با پیشرفت‌های اخیر در قدرت پردازشی، هزینه ادغام داده‌های چند سنسور نسبت به مزایای حاصل از آن کمتر شده است.

در بخش کارهای پیشین، پژوهش‌های مرتبط با ادغام داده‌های دوربین و لیدار را بررسی کردیم. این پژوهش‌ها نشان می‌دهند که ادغام داده‌های حاصل از این دو سنسور می‌تواند به بهبود دقت و قابلیت اطمینان سیستم‌های تشخیص شیء کمک کند. استفاده از تکنیک‌های یادگیری بدون نظارت، خوشه‌بندی نقاط و مدل‌های مبتنی بر مبدل، از جمله راهکارهایی است که در این زمینه به کار گرفته شده است.

با توجه به بررسی‌های انجام‌شده، مشخص شد که هر یک از روش‌های موجود دارای مزایا و معایبی

هستند. روش‌های تک‌مرحله‌ای سریع ولی کم‌دقت، روش‌های دومرحله‌ای دقیق ولی کند، و روش‌های مبتنی بر داده‌های تک‌سنسور دارای محدودیت‌هایی هستند. همچنین، هزینه بالای تهیه داده‌های برچسب‌دار یک چالش اساسی است.

در روش‌ما، با در نظر گرفتن این نکات، سعی داریم تا با ترکیب داده‌های دوربین و لیدار و استفاده از الگوریتم‌های یادگیری بدون نظارت مانند خوشه‌بندی، به بهبود فرآیند پیشنهاد ناحیه و در نهایت تشخیص شیء در خودروهای خودران بپردازیم. در فصل بعدی، روش پیشنهادی خود را که به منظور رفع برخی از این محدودیت‌ها طراحی شده است، ارائه خواهیم کرد.

فصل سوم

روش انجام پروژه

۳-۱ مقدمه

در این فصل به توضیح مراحل اولیه پردازش داده‌ها در پروژه می‌پردازیم. ابتدا به انتخاب مجموعه داده کیتی[۲۹] و دلایل انتخاب آن اشاره می‌کنیم. سپس فرآیند بارگذاری و آماده‌سازی داده‌های خام، شامل تصاویر و ابر نقاط لیدار، را شرح می‌دهیم. در ادامه، نحوه انتقال ابر نقاط به فضای دوربین، فیلتر کردن نقاط در میدان دید دوربین، و نگاشت آن‌ها بر روی تصاویر دوربین را بررسی می‌کنیم. این مراحل ابتدایی برای آماده‌سازی داده‌ها جهت استفاده در مراحل بعدی پروژه ضروری هستند.

۳-۲ مجموعه داده

اولین گام در اجرای این پروژه، انتخاب مجموعه داده مناسب برای پیاده‌سازی و ارزیابی روش پیشنهادی بود. مجموعه داده کیتی یکی از معروف‌ترین و جامع‌ترین مجموعه داده‌ها در حوزه خودروهای خودران و بینایی ماشین است که شامل اطلاعات گسترده‌ای از سنسورهای مختلف مانند دوربین‌های استریو، لیدار است. این مجموعه داده داده‌های متنوعی برای وظایف مختلف ارائه می‌دهد، از جمله تشخیص شیء دوبعدی و سه‌بعدی با تصاویر برچسب‌گذاری‌شده حاوی جعبه‌های محدودکننده برای اشیاء مختلف مانند خودروها، عابران پیاده و دوچرخه‌سواران؛ بخش‌بندی معنایی با تصاویر دارای برچسب‌های پیکسلی که نشان‌دهنده کلاس اشیاء در هر پیکسل هستند؛ تخمین عمق با تصاویر دارای نقشه‌های عمق که فاصله تا هر پیکسل را نشان می‌دهد؛ و مسیریابی و تخمین حرکت با داده‌های مرتبط با موقعیت و حرکت خودرو در طول زمان.

در این پروژه، از داده‌های خام کیتی استفاده شده است. دلیل این انتخاب این است که روش ما بر پایه یادگیری بدون نظارت است و نیازی به داده‌های برچسب‌دار یا داده مرجع[1] ندارد. داده‌های خام کیتی شامل تصاویر و ابر نقاط لیدار بدون برچسب‌گذاری دستی است که برای اهداف ما مناسب است.

۳-۳ بارگذاری و نمایش داده‌ها

داده‌های خام کیتی شامل چهار تصویر برای هر فریم است: تصویر ۰۰ (تصویر سمت چپ سیاه و سفید)، تصویر ۰۱ (تصویر سمت راست سیاه و سفید)، تصویر ۰۲ (تصویر سمت چپ رنگی) و تصویر ۰۳ (تصویر

[1] Ground Truth

سمت راست رنگی). تصاویر سیاه و سفید از دوربین‌های مونوکروم برای کاربردهایی مانند استریو ویژن استفاده می‌شوند، در حالی که تصاویر رنگی از دوربین‌های قرمز، سبز، آبی[۲] برای استخراج ویژگی‌های بصری غنی‌تر به‌کار می‌روند. تصویر سمت چپ سیاه و سفید نمایی از داده‌های عمق یا فاصله اشیاء از حسگر را نشان می‌دهد که اطلاعاتی درباره‌ی فاصله اجسام تا لیدار را به صورت مقیاس رنگی یا عددی نمایش می‌دهد.[۳] تصویر سمت راست سیاه و سفید شدت برگشتی از پالس‌های لیدار را نشان می‌دهد که میزان انرژی برگشتی از سطح اجسام را نمایش می‌دهد و می‌تواند به درک ویژگی‌های سطح کمک کند.[۴]

نمونه‌ای از این تصاویر در شکل ۳-۱ نشان داده شده است. در این شکل، قسمت آ) تصویر سمت چپ سیاه و سفید، قسمت ب) تصویر سمت راست سیاه و سفید، قسمت ج) تصویر سمت چپ رنگی و قسمت د) تصویر سمت راست رنگی را نشان می‌دهد.

شکل ۳-۱: نمونه‌ای از تصاویر دوربین‌های مختلف در مجموعه داده کیتی

علاوه بر تصاویر، ابر نقاط سه‌بعدی حاصل از سنسور لیدار نیز بارگذاری می‌شوند. این ابر نقاط شامل مختصات سه‌بعدی نقاط محیط اطراف خودرو هستند که توسط لیدار جمع‌آوری شده‌اند.

برای نمایش ابر نقاط و درک بهتر پراکندگی آن‌ها، از نمای بالا یا نمای پرنده[۵] استفاده می‌کنیم. شکل ۳-۲ نمای پرنده از ابر نقاط لیدار مربوط به همان فریم تصویر قبلی را نشان می‌دهد. در این شکل،

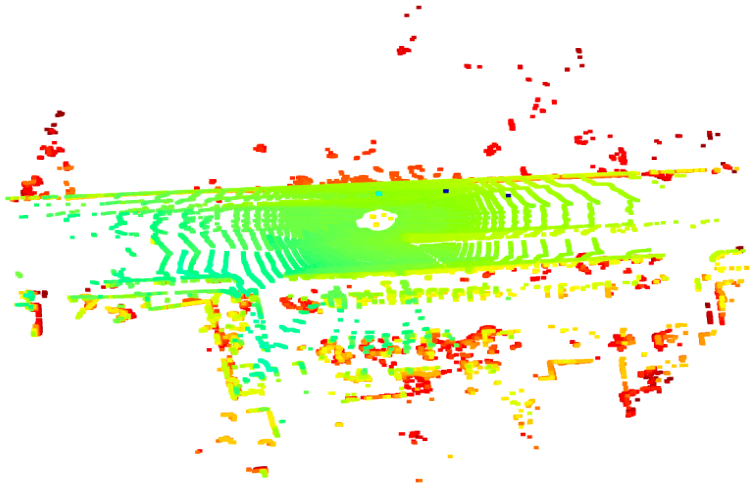
[2] RGB (Red, Green, Blue)

[۳] نمایی از داده‌های عمق یا فاصله اشیاء از حسگر را نشان می‌دهد که اطلاعاتی درباره‌ی فاصله اجسام تا لیدار را به صورت مقیاس رنگی یا عددی نمایش می‌دهد.

[۴] تصویر شدت برگشتی از پالس‌های لیدار است که میزان انرژی برگشتی از سطح اجسام را نشان می‌دهد و می‌تواند به درک ویژگی‌های سطح کمک کند (مثل تفاوت مواد یا سطح صاف و زبر).

[۵] نمای پرنده

نقاط در اطراف ماشین پراکنده شده‌اند و فضای خالی در مرکز تصویر، محل قرارگیری دوربین و ماشین را نشان می‌دهد. این تصویر ابر نقاط ۳۶۰ درجه اطراف ماشین را ثبت کرده و وضعیت توزیع نقاط در محیط پیرامون را نمایش می‌دهد.



شکل ۳-۲: نمای پرنده از ابر نقاط لیدار

۳-۴ پردازش و تطبیق ابر نقاط با دوربین

برای استفاده از اطلاعات ابر نقاط در کنار تصاویر دوربین، نیاز به انتقال و تطبیق ابر نقاط به فضای دوربین و نگاشت آن‌ها بر روی تصویر داریم. این فرآیند شامل انتقال ابر نقاط به فضای دوربین، فیلتر کردن نقاط در میدان دید دوربین، و نگاشت ابر نقاط بر روی تصویر دوربین است.

ابر نقاط اولیه در سامانه مختصات لیدار تعریف شده‌اند. برای انتقال این نقاط به سامانه مختصات دوربین، از ماتریس‌های کالیبراسیون بین لیدار و دوربین استفاده می‌کنیم. این ماتریس‌ها شامل ماتریس چرخش (R) و بردار انتقال (T) هستند که رابطه فضایی بین سنسور لیدار و دوربین را مشخص می‌کنند. تابعی پیاده‌سازی شده است که با استفاده از این ماتریس‌ها، مختصات ابر نقاط را به سامانه مختصات دوربین تبدیل می‌کند. این تبدیل به صورت زیر انجام می‌شود:

$$Points_{cam} = R \times Points_{velo} \quad (۱-۳)$$

که در آن $Points_{velo}$ مختصات ابر نقاط در سامانه لیدار، R ماتریس چرخش با ابعاد ۳ × ۳، T بردار انتقال با ابعاد ۱ × ۳، و $Points_{cam}$ مختصات ابر نقاط در سامانه دوربین هستند.

پس از انتقال ابر نقاط به فضای دوربین، لازم است تنها نقاطی را انتخاب کنیم که در میدان دید دوربین قرار دارند. این کار با استفاده از پارامترهای کالیبراسیون داخلی دوربین و ابعاد تصویر انجام می‌شود. ابتدا، مختصات نقاط در فضای دوربین را به مختصات همگن تبدیل می‌کنیم و سپس با استفاده از ماتریس پروجکشن (P) دوربین، مختصات تصویری نقاط را محاسبه می‌کنیم:

$$Points_{img} = P \times \begin{bmatrix} Points_{cam} \\ ۱ \end{bmatrix}$$

در این معادله، (u, v) مختصات تصویری نقاط بر روی صفحه تصویر هستند، و $Points_{cam}$ مختصات ابر نقاط در فضای دوربین است.

مختصات تصویری (u, v) را با تقسیم بر مولفه عمق (P_l) به دست می‌آوریم، که P_l نشان‌دهنده مختصات عمق نقاط ابر نقاط است:

$$u = \frac{\left(P \times \begin{bmatrix} Points_{cam} \\ ۱ \end{bmatrix} \right)_x}{P_l}$$

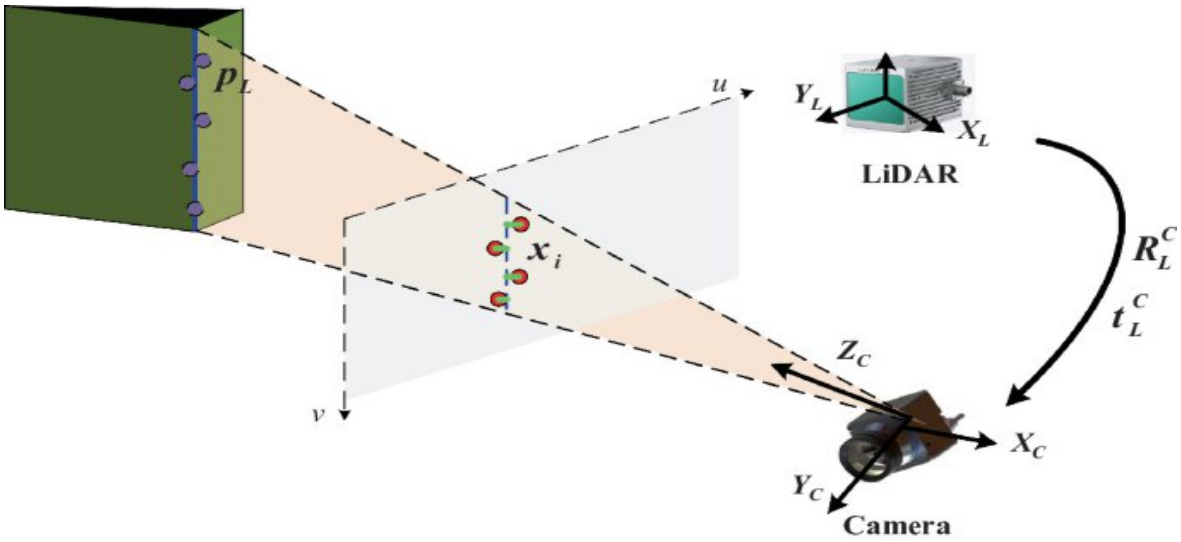
$$v = \frac{\left(P \times \begin{bmatrix} Points_{cam} \\ ۱ \end{bmatrix} \right)_y}{P_l}$$

در اینجا، مختصات فضایی X, Y, Z را با u, v, P_l جایگزین کرده‌ایم، به طوری که u و v مختصات دوبعدی تصویر هستند که از نگاشت X و Y به صفحه تصویر به دست می‌آیند، و P_l نشان‌دهنده عمق یا فاصله نقطه از دوربین در راستای محور نوری (معادل Z) است. این نشانه گذاری در حوزه بینایی ماشین رایج است و فرآیند پیش افکنی نقاط سه‌بعدی به فضای دوبعدی تصویر همراه با حفظ اطلاعات عمق را مدل‌سازی می‌کند.

در نهایت، نقاطی را که در محدوده تصویر قرار دارند، یعنی عرض تصویر $۰ \leq u <$ و $۰ \leq v <$ ارتفاع تصویر، و همچنین عمق مثبت دارند ($۰ < P_l$)، انتخاب می‌کنیم. با استفاده از مختصات تصویری

نقاط فیلترشده، می‌توانیم ابر نقاط را بر روی تصویر دوربین نگاشت کنیم. این کار با رسم نقاط بر روی تصویر انجام می‌شود، به طوری که هر نقطه در مختصات (u, v) بر روی تصویر نمایش داده می‌شود.

در شکل ۳-۳، یک دیاگرام نشان داده شده است که در آن یک سمت دوربین قرار دارد، سمت دیگر صفحه u, v و در انتها ابر نقاط با نام P_l مشخص شده‌اند. این شکل کمک می‌کند تا فرآیند نگاشت ابر نقاط از فضای سه‌بعدی به تصویر دوبعدی را بهتر درک کنیم.



شکل ۳-۳: دیاگرام تطبیق ابر نقاط با تصویر؛ در این شکل، یک سمت دوربین، سمت دیگر صفحه u, v و در انتها ابر نقاط با نام P_l مشخص شده‌اند

همچنین، در شکل ۴-۳، نتیجه نگاشت ابر نقاط بر روی تصویر دوربین نشان داده شده است. این تصویر نشان‌دهنده توزیع نقاط ابر نقاط بر روی تصویر اصلی است؛ نقاط حاصل از لیدار بر روی تصویر دوربین پروجکت شده‌اند و نحوه قرارگیری این نقاط در فضای دوبعدی تصویر را نشان می‌دهد.

۳-۵ استخراج ویژگی با استفاده از مدل دیپ‌لب‌وی۳

پس از ترکیب اولیه داده‌های لیدار و دوربین، نیاز به استخراج ویژگی‌های تصویری داریم که بتوانند در خوشه‌بندی و پیشنهاد ناحیه کمک کنند. با توجه به شباهت مسئله ما به بخش‌بندی تصویر، از مدل دیپ‌لب‌وی۳ [۲۲] برای استخراج ویژگی‌های تصویر استفاده می‌کنیم. این مدل که در فصل قبل به‌تفصیل توضیح داده شد، یکی از پیشرفته‌ترین معماری‌ها در حوزه بخش‌بندی معنایی تصویر است و عملکرد بسیار خوبی در استخراج ویژگی‌های غنی از تصاویر دارد.

در پروژه ما، از مدل دیپ‌لب‌وی۳ به عنوان استخراج‌کننده ویژگی‌های تصویری استفاده می‌کنیم.

شکل ۳-۴: نمایش ابر نقاط بر روی تصویر دوربین

همان‌طور که در معماری این مدل توضیح داده شد، خروجی بخش رمزگذار شامل نقشه‌های ویژگی با اطلاعات غنی در مورد محتوای تصویر است. از آنجا که هدف ما انجام بخش‌بندی نهایی نیست، بلکه استخراج ویژگی‌های مناسب برای خوشه‌بندی است، خروجی لایه ماقبل آخر مدل را به‌کار می‌بریم.

این خروجی شامل نقشه ویژگی با ۲۵۶ کانال و ابعاد مکانی کاهش‌یافته است. برای هم‌تراز کردن این نقشه ویژگی با ابعاد تصویر اصلی، از عملیات نمونه افزایی[۶] استفاده می‌کنیم. پس از نمونه افزایی، نقشه ویژگی استخراج‌شده با ابعاد تصویر اصلی هماهنگ می‌شود و می‌توانیم آن را با داده‌های لیدار ادغام کنیم.

این روش به ما امکان می‌دهد تا با استفاده از ویژگی‌های غنی استخراج‌شده از تصویر، خوشه‌بندی مؤثرتری را در ترکیب با داده‌های لیدار انجام دهیم.

۳-۶ ادغام داده‌های لیدار و ویژگی‌های تصویری

پس از استخراج ویژگی‌های تصویر و نمونه‌افزایی آن به ابعاد تصویر اصلی، می‌توانیم داده‌های لیدار و ویژگی‌های تصویری را با یکدیگر ادغام کنیم. هر نقطه در نقشه ویژگی استخراج‌شده دارای مختصات دوبعدی u, v و یک بردار ویژگی با ۲۵۶ کانال است. از طرفی، نقاط لیدار نگاشت‌شده بر روی تصویر نیز دارای مختصات تصویری u, v و مختصات فضایی u, v, P_l هستند.

[6] Upsampling

برای ادغام این دو مجموعه داده، نقاطی را که مختصات تصویری یکسانی دارند، با یکدیگر ترکیب می‌کنیم. به این ترتیب، هر نقطه ادغام‌شده شامل مختصات فضایی u, v, P_l و بردار ویژگی ۲۵۶ کاناله از تصویر است. خروجی نهایی ما یک مجموعه داده است که در آن هر نقطه دارای ۳ کانال مختصات فضایی و ۲۵۶ کانال ویژگی تصویری است.

۳-۷ خوشه‌بندی داده‌های ادغام‌شده

با در دست داشتن داده‌های ادغام‌شده، هدف ما انجام خوشه‌بندی بر روی این نقاط است تا ناحیه‌های حاوی اشیاء موجود در صحنه را شناسایی کنیم. برای این منظور، از الگوریتم دی‌بی‌اسکن استفاده می‌کنیم.

با اعمال الگوریتم دی‌بی‌اسکن بر روی داده‌های ادغام‌شده، نتایج اولیه نشان داد که خوشه‌بندی به خوبی انجام نشده است. اکثر نقاط به عنوان نویز شناسایی شده‌اند و خوشه‌های تشکیل‌شده پراکنده و نامرتبط هستند. نمونه‌ای از نتیجه خوشه‌بندی اولیه در شکل ۳-۵ نشان داده شده است.

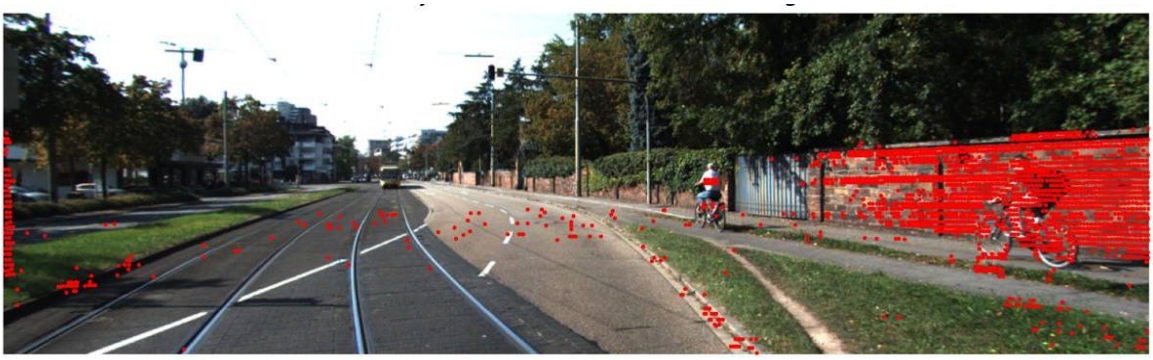
شکل ۳-۵: خروجی خوشه‌بندی ابر نقاط به همراه تصویر با استفاده از روش خوشه‌بندی دی‌بی‌اسکن و ویژگی‌های تصویر استخراج‌شده با مدل دیپ‌لب‌وی۳

با بررسی نتایج خوشه‌بندی، متوجه شدیم که بزرگ‌ترین مشکل مربوط به نقاط زمین است. این نقاط با داشتن مقادیر P_l کم و ویژگی‌های تصویری مشابه، باعث ایجاد اختلال در خوشه‌بندی می‌شوند. از آنجا که مدل یادگیری بدون نظارت ما تمایل دارد بر روی نقاط زمین تمرکز کند، خوشه‌های مربوط به اشیاء مهم مانند خودروها و عابران پیاده به خوبی شناسایی نمی‌شوند.

برای بهبود دقت خوشه‌بندی، تصمیم گرفتیم نقاط مربوط به زمین را حذف کنیم. این کار به ما امکان می‌دهد تا تمرکز مدل را بر روی اشیاء مهم‌تر قرار دهیم. نقاط زمین معمولاً دارای ویژگی‌های تصویری

مشابهی هستند، زیرا سطح آسفالت در تصاویر دوربین تقریباً یکنواخت است. همچنین، تعداد زیادی از نقاط لیدار مربوط به زمین هستند که می‌تواند باعث ایجاد خوشه‌های بزرگ و نامرتبط شود. علاوه بر این، نقاط زمین دارای مقادیر P_l کم و تقریباً یکسانی هستند که باعث می‌شود الگوریتم دی‌بی‌اسکن آن‌ها را به عنوان یک خوشه بزرگ در نظر بگیرد.

برای حذف نقاط زمین، چند روش را امتحان کردیم. ابتدا از روشی مبتنی بر تحلیل نرمال‌های سطح [۷] استفاده کردیم. در این روش، نرمال‌های هر نقطه با استفاده از همسایگی آن محاسبه می‌شود. نقاطی که نرمال آن‌ها با جهت عمودی مطابقت دارد (یعنی زاویه نرمال با محور P_l کم است)، به عنوان نقاط زمین در نظر گرفته می‌شوند و حذف می‌گردند. با این حال، این روش منجر به حذف برخی نقاط دور از سنسور نیز شد که منطقی نبود، زیرا نقاطی که بر روی سطوح افقی دیگر قرار دارند نیز ممکن است به اشتباه حذف شوند. همان‌طور که در تصویر ۳-۶ مشاهده می‌شود، خروجی مطلوبی نگرفتیم.



شکل ۳-۶: ابر نقاط منطبق شده بر روی تصویر بعد از حذف نقاط زمین با روش نرمال‌های سطح

روش دیگر، حذف نقاط با مقادیر P_l کمتر از یک مقدار آستانه بود. این روش نیز مشکلاتی داشت، زیرا ممکن است برخی از نقاط زمین که ارتفاع کمی دارند باقی بمانند و نقاط اشیاء دور که ارتفاع کم دارند حذف شوند. همان‌طور که در تصویر ۳-۷ مشاهده می‌شود، این روش نیز به نتایج دلخواه منجر نشد.

در نهایت، از الگوریتم اجتماع تصادفی نمونه‌ها[۸] برای تخمین صفحه زمین و حذف نقاط متعلق به آن استفاده کردیم [۳۰]. الگوریتم اجتماع تصادفی نمونه‌ها یک روش تکراری برای تخمین یک مدل از داده‌های دارای نویز است. در اینجا، هدف ما یافتن بهترین صفحه‌ای است که نمایانگر زمین باشد، که به

[7] Surface Normals

[8] Random Sample Consensus

شکل ۳-۷: ابر نقاط منطبق شده بر روی تصویر بعد از حذف نقاط زمین با روش فیلتر کردن بر اساس مقدار P_l

آن **برآورد صفحه**[۹] گفته می‌شود.

الگوریتم به این صورت کار می‌کند که ابتدا به صورت تصادفی حداقل تعداد نقاط مورد نیاز برای تعریف یک صفحه (سه نقطه) انتخاب می‌شود. سپس، یک مدل صفحه با استفاده از این نقاط انتخاب‌شده تخمین زده می‌شود. معادله یک صفحه در فضای سه‌بعدی به صورت $ax + by + cz + d = ۰$ است که a، b، c و d ضرایب صفحه هستند. سپس، برای تمامی نقاط ابر نقاط، فاصله عمودی آن‌ها تا صفحه تخمین‌زده‌شده محاسبه می‌شود. نقاطی که فاصله آن‌ها از صفحه کمتر از یک آستانه مشخص (ϵ) باشد، به عنوان درون‌پوش‌ها[۱۰] در نظر گرفته می‌شوند. تعداد درون‌پوش‌ها شمارش می‌شود. این فرآیند تا تعداد مشخصی تکرار یا تا زمانی که تعداد درون‌پوش‌ها به حداکثر برسد ادامه می‌یابد. در نهایت، صفحه‌ای که بیشترین تعداد درون‌پوش‌ها را دارد به عنوان تخمین نهایی انتخاب می‌شود. پس از تخمین صفحه زمین، نقاطی که در فاصله عمودی کمی از این صفحه قرار دارند به عنوان نقاط زمین در نظر گرفته شده و حذف می‌شوند.

این فرآیند به ما امکان می‌دهد تا نقاط مربوط به زمین را که برای تشخیص اشیاء مورد نظر ما اهمیت کمتری دارند، از ابر نقاط حذف کنیم و بدین ترتیب عملکرد الگوریتم‌های خوشه‌بندی را بهبود بخشیم.

با استفاده از روش اجتماع تصادفی نمونه‌ها، توانستیم نقاط زمین را به طور مؤثری حذف کنیم بدون اینکه نقاط مربوط به اشیاء دیگر تحت تأثیر قرار بگیرند. شکل ۳-۸ نمایانگر ابر نقاط پس از حذف نقاط زمین است.

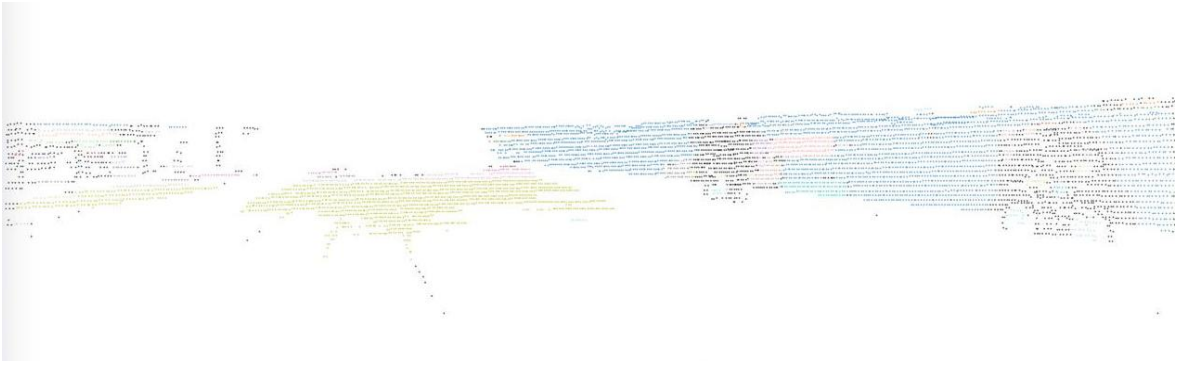
[9] Plane Fitting

[10] Inliers

شکل ۳-۸: ابر نقاط منطبق شده بر روی تصویر بعد از حذف نقاط زمین با روش اجتماع تصادفی نمونه‌ها

۳-۸ خوشه‌بندی پس از حذف نقاط زمین

شکل ۳-۹ خروجی حاصل از خوشه‌بندی پس از حذف نقاط زمین را نشان می‌دهد. همان‌طور که در تصویر مشاهده می‌شود، نقاط مربوط به دوچرخه‌سوار به اشتباه به عنوان نویز دسته‌بندی شده‌اند. با اینکه هدف اصلی پروژه دسته‌بندی صحیح اشیاء و رهگذران در تصویر است، این یک قدم در جهت درست است، زیرا خوشه‌بندی نسبت به قبل بهبود زیادی یافته است. با این حال، هنوز می‌توان عملکرد را بهبود بخشید.



شکل ۳-۹: خروجی خوشه‌بندی ابر نقاط بعد از حذف نقاط زمین با استفاده از روش خوشه‌بندی دی‌بی‌اسکن و ویژگی‌های تصویری استخراج‌شده با مدل دیپ‌لب‌وی۳.

پس از حذف نقاط زمین و مشاهده بهبود در نتایج خوشه‌بندی، متوجه شدیم که تنظیم مناسب وزن‌های مختصات فضایی (u, v, P_l) و ویژگی‌های تصویری استخراج‌شده می‌تواند تأثیر قابل توجهی بر کیفیت خوشه‌بندی داشته باشد. از آنجا که داده‌های ما شامل دو نوع ویژگی متفاوت است (مختصات فضایی که نشان‌دهنده موقعیت سه‌بعدی نقاط هستند و ویژگی‌های تصویری با ابعاد بالا که اطلاعات غنی از محتوای بصری تصویر را در بر دارند) لازم است تا وزن‌های مناسبی به هر یک اختصاص دهیم تا

الگوریتم خوشه‌بندی بتواند به درستی از هر دو نوع اطلاعات بهره‌برداری کند.

اگر وزن‌های این دو نوع ویژگی به درستی تنظیم نشوند، ممکن است یکی از آن‌ها بر دیگری غالب شود و الگوریتم خوشه‌بندی نتواند ساختار واقعی داده‌ها را تشخیص دهد. به عنوان مثال، اگر وزن مختصات فضایی بسیار بیشتر از وزن ویژگی‌های تصویری باشد، خوشه‌بندی بیشتر بر اساس موقعیت مکانی نقاط انجام می‌شود و ممکن است اشیائی که از نظر بصری مشابه هستند ولی در فواصل نزدیک قرار دارند، در خوشه‌های جداگانه‌ای قرار گیرند. بالعکس، اگر وزن ویژگی‌های تصویری غالب باشد، ممکن است نقاطی که از نظر مکانی دور هستند ولی ویژگی‌های تصویری مشابهی دارند، در یک خوشه قرار گیرند که می‌تواند منجر به خوشه‌بندی نادرست شود.

برای یافتن ترکیب بهینه وزن‌ها، از روش جستجوی شبکه‌ای[11] استفاده کردیم. جستجوی شبکه‌ای یک روش جامع برای تنظیم فراپارامترها است که در آن، محدوده‌ای از مقادیر ممکن برای هر پارامتر تعریف می‌شود و تمام ترکیبات ممکن از این مقادیر بررسی می‌شوند [۳۱]. در اینجا، با تنظیم وزن‌های مختلف برای مختصات فضایی و ویژگی‌های تصویری، سعی کردیم به ترکیبی برسیم که بهترین نتایج را در خوشه‌بندی ارائه دهد. جزئیات مربوط به مقادیر وزن‌ها و پارامترهای استفاده‌شده در فصل چهارم ارائه خواهند شد.

برای ارزیابی کیفیت خوشه‌بندی، نیاز به یک معیار داشتیم که بتواند میزان تطابق خوشه‌ها با اشیاء واقعی در تصویر را اندازه‌گیری کند. از آنجا که داده‌های برچسب‌دار[12] در دسترس نبود، از یک مدل بخش‌بندی پیش‌آموزش‌دیده به نام سم[13] استفاده کردیم [۳۲]. سم یک مدل قدرتمند در بخش‌بندی تصویر است که قادر است بدون نیاز به ورودی‌های دستی، بخش‌بندی دقیق اشیاء را انجام دهد. با اعمال این مدل بر روی تصاویر، نقشه‌های بخش‌بندی تولید شد که در آن هر پیکسل به یک بخش خاص تعلق دارد. شکل ۳-۱۰ خروجی حاصل از بخش‌بندی تصویر توسط مدل سم را نشان می‌دهد.

با استفاده از نقشه‌های بخش‌بندی تولیدشده توسط سم، توانستیم کیفیت خوشه‌بندی را ارزیابی کنیم. به طور کلی، مشاهده کردیم که افزایش وزن ویژگی‌های تصویری نسبت به مختصات فضایی منجر به بهبود خوشه‌بندی اشیاء کوچک مانند خودروها و عابران پیاده می‌شود. این امر منطقی است زیرا ویژگی‌های تصویری اطلاعات دقیق‌تری در مورد محتوای بصری نقاط ارائه می‌دهند. از سوی دیگر، برای اشیاء بزرگ‌تر مانند ساختمان‌ها و دیوارها، تنظیم وزن‌های مختصات فضایی اهمیت بیشتری داشت.

[11] Grid Search

[12] Ground Truth

[13] SAM (Segment Anything Model)

شکل ۳-۱۰: بخش‌بندی انجام‌شده توسط مدل سم

جزئیات دقیق مربوط به تنظیم وزن‌ها، پارامترهای استفاده‌شده، و نتایج عددی ارزیابی در فصل چهارم ارائه خواهند شد.

۳-۹ استفاده از الگوریتم‌های خوشه‌بندی دیگر

برای بهبود عملکرد خوشه‌بندی و بررسی تأثیر الگوریتم‌های مختلف، تصمیم گرفتیم از الگوریتم‌های کا میانگین و خوشه‌بندی طیفی نیز استفاده کنیم. این الگوریتم‌ها در فصل قبل به تفصیل توضیح داده شدند و الگوریتم‌های آن‌ها تشریح گردید.

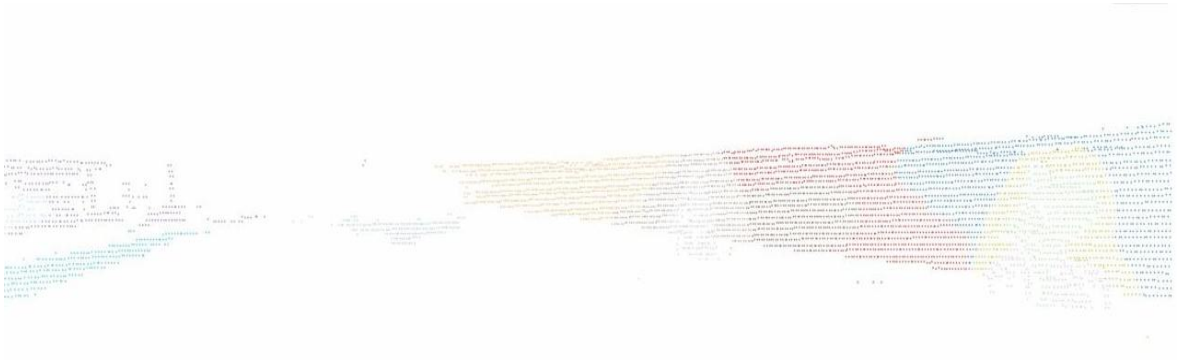
در ابتدا، الگوریتم کا میانگین را به کار گرفتیم. مقادیر مختلفی برای تعداد خوشه‌ها بین ۲۵ تا ۲۰۰ را امتحان کردیم. از وزن‌های بهینه به دست آمده در جستجوی شبکه‌ای برای مقیاس‌دهی ویژگی‌ها استفاده کردیم و الگوریتم کا میانگین را با استفاده از روش کا میانگین++ برای انتخاب مراکز اولیه اجرا نمودیم. نتایج خوشه‌بندی را با معیارهای قبلی ارزیابی کردیم.

شکل ۳-۱۱ خروجی حاصل از خوشه‌بندی با روش کا میانگین را نشان می‌دهد. مشاهده شد که با افزایش مقدار K، الگوریتم کا میانگین در تشخیص اشیاء کوچک مانند خودروها و دوچرخه‌ها عملکرد بهتری دارد. این امر به این دلیل است که تعداد خوشه‌های بیشتر، امکان تفکیک دقیق‌تر نقاط را فراهم می‌کند. با این حال، تعیین مقدار مناسب K یک چالش است. اگر K بسیار بزرگ باشد، اشیاء بزرگ‌تر مانند دیوارها و ساختمان‌ها به چندین خوشه تقسیم می‌شوند. اگر K بسیار کوچک باشد، اشیاء کوچک در خوشه‌های بزرگ‌تر گم می‌شوند.

سپس الگوریتم خوشه‌بندی طیفی را مورد استفاده قرار دادیم. این الگوریتم که در فصل قبل توضیح داده شد، قادر است ساختارهای پیچیده و غیرخطی را شناسایی کند. با استفاده از این روش، خروجی

شکل ۳-۱۱: خروجی خوشه‌بندی ابر نقاط به همراه تصویر با استفاده از روش خوشه‌بندی کا میانگین

حاصل نشان داد که اشیاء بسیار کوچک با دقت بهتری تشخیص داده می‌شوند. به عنوان مثال، حتی لباس و دوچرخه فرد دوچرخه‌سوار به صورت جداگانه خوشه‌بندی شده‌اند، در حالی که در مدل کا میانگین، کل دوچرخه و فرد دوچرخه‌سوار به عنوان یک خوشه دسته‌بندی شده بودند. خروجی حاصل از این روش در شکل ۳-۱۲ آمده است.



شکل ۳-۱۲: خروجی خوشه‌بندی ابر نقاط به همراه تصویر با استفاده از روش خوشه‌بندی خوشه‌بندی طیفی

با این حال، به دلیل پیچیدگی محاسباتی بالا، الگوریتم خوشه‌بندی طیفی برای داده‌های بزرگ مانند ابر نقاط مناسب نبود. بار محاسباتی این روش (حدود پنج برابر بیشتر از کا میانگین) استفاده از آن را در عمل غیرمقرون به صرفه می‌کند، زیرا افزایش زمان محاسباتی برای بهبود جزئی در نتایج، توجیه‌پذیر نیست.

الگوریتم دی‌بی‌اسکن که قبلاً استفاده شده بود، به دلیل عدم نیاز به تعیین تعداد خوشه‌ها و قابلیت شناسایی نویز، مزایایی دارد. این الگوریتم می‌تواند به راحتی اشیاء بسیار بزرگ یا بسیار کوچک را شناسایی کند، که در برخی موارد از مزیت‌های آن به شمار می‌آید. با این حال، پارامترهای حساس آن (ϵ و MinPts) باید با دقت تنظیم شوند. اگر تصویر دارای ویژگی‌های خاصی باشد، مانند روشنایی یا تاریکی

بیش از حد، ممکن است اکثر ابر نقاط به عنوان نویز شناسایی شوند. این امر منجر به کاهش عملکرد الگوریتم در شرایط نوری نامناسب می‌شود.

در مجموع، مقایسه الگوریتم‌های مختلف نشان داد که هر یک از آن‌ها مزایا و معایب خاص خود را دارند. کا میانگین با سادگی و سرعت اجرای بالا، برای داده‌های با خوشه‌های کروی و همگن مناسب است، اما تعیین تعداد خوشه‌ها یک چالش است. خوشه‌بندی طیفی با توانایی شناسایی ساختارهای پیچیده، در مسائل پیچیده‌تر مفید است، اما هزینه محاسباتی بالایی دارد. دی‌بی‌اسکن بدون نیاز به تعیین تعداد خوشه‌ها و با قابلیت شناسایی نویز، برای داده‌های با توزیع‌های مختلف مناسب است، اما حساسیت آن به تنظیم پارامترها می‌تواند بر عملکرد آن تأثیر بگذارد.

برای بهبود نتایج خوشه‌بندی، تصمیم گرفتیم خوشه‌های بسیار کوچک و بسیار بزرگ را حذف کنیم. خوشه‌هایی با تعداد نقاط کم را به عنوان نویز در نظر گرفتیم و حذف کردیم؛ این خوشه‌ها معمولاً ناشی از نویز سنسورها یا خطاهای خوشه‌بندی هستند. همچنین، خوشه‌های بسیار بزرگ احتمالاً مربوط به ساختارهایی مانند دیوارها یا زمین هستند که برای کاربرد ما اهمیت کمتری دارند. با حذف این خوشه‌ها، تمرکز بر روی اشیاء مهم مانند خودروها و عابران پیاده افزایش یافت.

۳-۱۰ تبدیل خوشه‌ها به پیشنهاد ناحیه

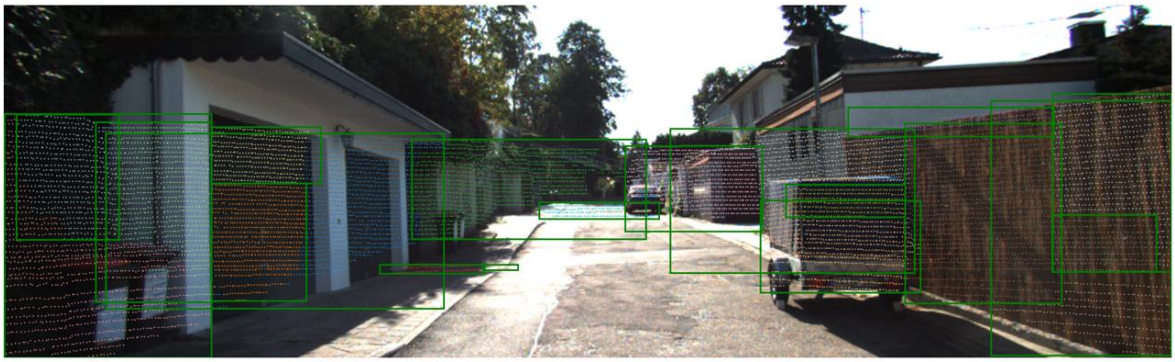
با داشتن خوشه‌های بهبودیافته، می‌توانیم آن‌ها را به پیشنهادات ناحیه تبدیل کنیم. برای هر خوشه، جعبه محدودکننده[۱۴] را با مراحل زیر تعیین کردیم: ابتدا مختصات تصویری u و v نقاط خوشه را استخراج کردیم. سپس کمینه و بیشینه u و v را به دست آوردیم. در نهایت، مستطیلی با گوشه‌های (u_{min}, v_{min}) و (u_{max}, v_{max}) رسم کردیم.

شکل ۳-۱۳ خروجی پیشنهاد ناحیه انجام‌شده با روش ذکرشده را نشان می‌دهد و شکل ۳-۱۴ مربوط به جعبه‌های محدودکننده داده‌های واقعی است.

۳-۱۱ ارزیابی پیشنهادات ناحیه با استفاده از داده‌های واقعی

برای ارزیابی دقیق عملکرد مدل و سنجش کیفیت پیشنهادات ناحیه، از برچسب‌های واقعی موجود در مجموعه داده کیتی استفاده کردیم. این برچسب‌ها شامل جعبه‌های محدودکننده دوبعدی اشیاء مختلف

[۱۴] Bounding Box



شکل ۳-۱۳: خروجی پیشنهاد ناحیه انجام‌شده با مدل K-Means

شکل ۳-۱۴: جعبه‌های محدودکننده داده‌های واقعی

مانند خودروها، عابران پیاده و دوچرخه‌سواران هستند. با استفاده از این داده‌های واقعی، معیارهای ارزیابی استانداردی مانند تلاقی بر اتحاد[۱۵]، مثبت واقعی[۱۶]، منفی مثبت[۱۷] و منفی کاذب[۱۸] را محاسبه کردیم. برای ما مهم است که بدانیم مدل ما چگونه در مقایسه با این معیارها عمل می‌کند و چه مقدار از جعبه‌های پیشنهادی ما با جعبه‌های واقعی هم‌پوشانی دارند.

استفاده از داده‌های واقعی برای ارزیابی مدل از اهمیت بالایی برخوردار است، زیرا امکان مقایسه عینی و قابل اعتماد عملکرد مدل را فراهم می‌کند. همچنین با تحلیل نتایج ارزیابی، می‌توانیم نقاط قوت و ضعف مدل را شناسایی کرده و برای بهبود آن اقدام کنیم. جزئیات مربوط به پارامترهای ارزیابی، نحوه محاسبه آن‌ها و فرمول‌های دقیق در فصل چهارم ارائه خواهد شد.

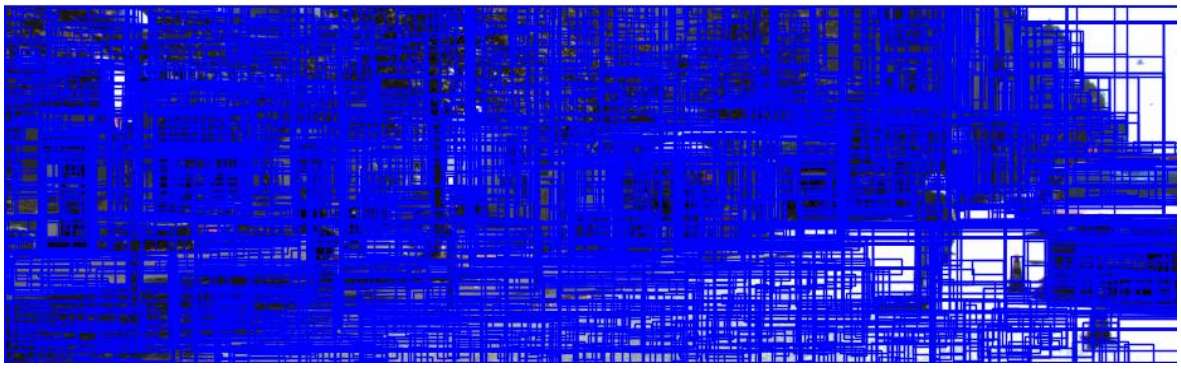
نتایج کمی و تحلیل دقیق معیارهای ارزیابی را در فصل چهارم ارائه خواهیم کرد. در آن فصل، به بررسی عملکرد مدل در شرایط مختلف، تأثیر تنظیمات متفاوت و مقایسه با مدل‌های دیگر خواهیم

[15] IoU (Intersection over Union)
[16] True Positive
[17] True Negative
[18] False Negative

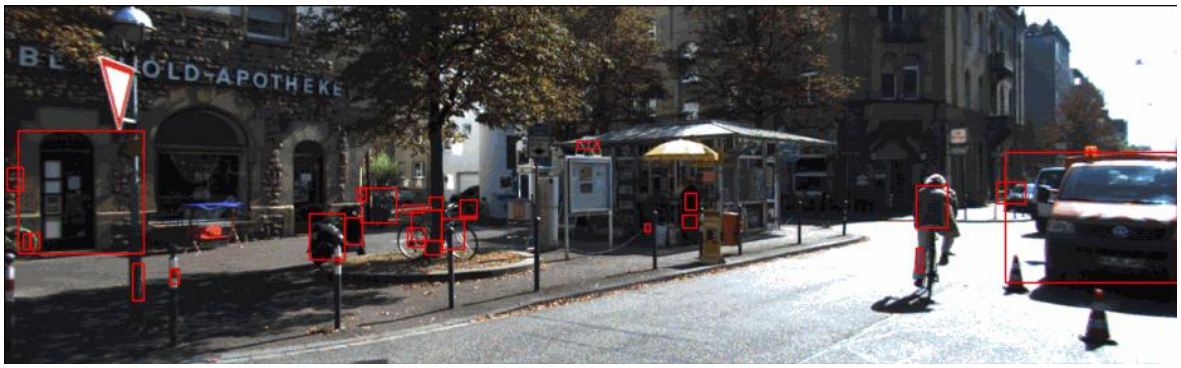
پرداخت.

برای ارزیابی عملکرد مدل خود و اطمینان از کارایی آن در مقایسه با مدل‌های پیشرفته، تصمیم گرفتیم تا مدل خود را با **مدل فَستر آر-سی ان ان** و روش **جستجوی انتخابی** مقایسه کنیم. این مقایسه به ما امکان می‌دهد تا نقاط قوت و ضعف روش پیشنهادی خود را در مقابل تکنیک‌های معتبر و شناخته‌شده بسنجیم.

در مقایسه با مدل فَستر آر-سی ان ان، که یکی از پیشرفته‌ترین و پرکاربردترین مدل‌ها در حوزه تشخیص شیء است [۲]، خروجی پیشنهاد ناحیه حاصل از مدل ما و مدل آن‌ها را مقایسه کردیم. شکل ۳-۱۵ خروجی پیشنهاد ناحیه حاصل از مدل فَستر آر-سی ان ان و شکل ۳-۱۶ خروجی پیشنهاد ناحیه حاصل از مدل ما را نشان می‌دهد.



شکل ۳-۱۵: خروجی پیشنهاد ناحیه حاصل از مدل فَستر آر-سی ان ان



شکل ۳-۱۶: خروجی پیشنهاد ناحیه حاصل از خوشه‌بندی

با ارزیابی معیارهای تلاقی بر اتحاد، تعداد مثبت حقیقی، منفی حقیقی و منفی کاذب، مشاهده کردیم که مدل ما با تعداد کمتری پیشنهاد ناحیه، توانسته است دقت بالاتری را ارائه دهد. این نشان می‌دهد که روش ما در تشخیص نواحی مرتبط با اشیاء عملکرد مؤثری دارد و می‌تواند با مدل‌های پیشرفته رقابت

کند. جزئیات مربوط به این معیارها و نحوه محاسبه آن‌ها در فصل چهارم ارائه خواهد شد.

۳-۱۲ جمع‌بندی

در این بخش، با تبدیل خوشه‌ها به پیشنهادات ناحیه و ارزیابی آن‌ها با استفاده از داده‌های واقعی، عملکرد مدل پیشنهادی خود را بررسی کردیم. همچنین، با مقایسه با مدل‌های پیشرفته‌ای مانند فَستر آر-سی ان ان و روش‌های پایه‌ای مانند جستجوی انتخابی، نشان دادیم که روش ما قادر است با تعداد کمتری پیشنهاد ناحیه، دقت بالایی را در تشخیص نواحی مرتبط با اشیاء ارائه دهد.

توجه به دلایل انتخاب روش‌ها و معیارها به ما کمک کرد تا درک بهتری از مزایا و محدودیت‌های مدل خود داشته باشیم و زمینه را برای بهبودهای آینده فراهم کنیم. این ارزیابی‌ها نشان می‌دهند که استفاده از داده‌های ادغام‌شده دوربین و لیدار، همراه با روش‌های خوشه‌بندی مناسب، می‌تواند راهکاری مؤثر و کارآمد برای تولید پیشنهادات ناحیه در سامانه‌های تشخیص شیء باشد. در فصل بعدی، مقایسه مدل خود را به طور کامل با دیگر مدل‌ها انجام می‌دهیم و کارایی پیشنهادات ناحیه خود را می‌سنجیم.

فصل چهارم

ارزیابی و نتایج

در این فصل، به ارزیابی مدل‌های پیشنهادی و تحلیل نتایج حاصل از پیاده‌سازی آن‌ها می‌پردازیم. ابتدا به معرفی مجموعه داده کیتی و ساختار آن می‌پردازیم. سپس، نتایج مربوط به تنظیم وزن‌ها در الگوریتم‌های خوشه‌بندی مختلف را با جزئیات ارائه می‌دهیم. در ادامه، پارامترهای استفاده‌شده در جستجوی شبکه‌ای و مدل سم را معرفی کرده و تأثیر آن‌ها را بر کیفیت خوشه‌بندی بررسی می‌کنیم. همچنین، عملکرد مدل‌های مختلف را با یکدیگر مقایسه کرده و تحلیل جامعی از نتایج به‌دست‌آمده ارائه می‌کنیم.

۴-۱ مجموعه داده کیتی و محیط پیاده سازی

مجموعه داده کیتی یکی از معتبرترین و گسترده‌ترین منابع برای تحقیقات در زمینه بینایی کامپیوتر و خودروهای خودران است. این مجموعه داده شامل تصاویر دوبعدی، داده‌های سه‌بعدی لیدار، و برچسب‌های دقیق برای اشیاء مختلف مانند خودروها، عابران پیاده و دوچرخه‌سواران است.

مجموعه داده کیتی شامل دو بخش اصلی است: داده‌های آموزش که شامل ۵,۰۰۰ تصویر همراه با برچسب‌های دقیق برای اشیاء مختلف است، و داده‌های آزمایش که شامل ۲,۴۸۱ تصویر بدون برچسب است و برای ارزیابی مدل‌ها استفاده می‌شود. در این پروژه، از ۵,۰۰۰ تصویر از داده‌های آموزش برای ارزیابی مدل‌های خود استفاده کردیم. دلیل این انتخاب این است که ما قصد داریم کارایی مدل خود را با مدل‌های فَستر آر-سی‌اِن‌اِن و جستجوی انتخابی مقایسه کنیم، و برای انجام این مقایسه به درستی، نیاز به برچسب‌های دقیق برای اشیاء داخل تصویر داریم. بنابراین، منطقی نیست که از داده‌های آزمایشی که بدون برچسب هستند، استفاده کنیم، زیرا در این صورت نمی‌توانیم به‌طور دقیق متوجه شویم که مدل ما یا مدل‌های دیگر تا چه حد خوب عمل می‌کنند. تمامی ارزیابی‌ها و تنظیم وزن‌ها بر روی این ۵,۰۰۰ تصویر از داده‌های آموزش انجام شد.

پیاده‌سازی مدل‌ها در محیط پایتون ۳ و با استفاده از کتابخانه‌های پای‌تورچ برای شبکه‌های عصبی عمیق و سایکیت‌لرن برای الگوریتم‌های خوشه‌بندی انجام شد. تمامی آزمایش‌ها بر روی سامانه‌ای با مشخصات پردازنده Apple M1 Pro، ۱۶ گیگابایت رم، بدون کارت گرافیک، و سیستم‌عامل macOS Sonoma انجام گرفت.

۴-۲ تنظیم وزن‌ها در الگوریتم‌های خوشه‌بندی

همان‌طور که در فصل قبل اشاره شد، برای استخراج ویژگی‌ها از مدل دیپ‌لب‌وی۳ استفاده کردیم. در این بخش، نتایج به‌دست‌آمده از تنظیم وزن‌ها در الگوریتم دی‌بی‌اسکن را قبل و بعد از حذف نقاط زمین ارائه می‌دهیم. از ویژگی‌های استخراج‌شده توسط این مدل برای ترکیب با مختصات فضایی و انجام خوشه‌بندی استفاده کردیم.

برای یافتن ترکیب بهینه وزن‌های مختصات فضایی و ویژگی‌های تصویری، از روش جستجوی شبکه‌ای استفاده کردیم. در این روش، محدوده‌ای از مقادیر ممکن برای هر وزن تعریف شده و تمامی ترکیبات ممکن بررسی می‌شوند. دو پارامتر اصلی ما شامل w_{uvp} که وزن مختصات فضایی (u, v, P_l) است و w_{feat} که وزن ویژگی‌های تصویری استخراج‌شده از مدل دیپ‌لب‌وی۳ می‌باشد، بودند.

محدوده مقادیر ممکن برای هر وزن را از ۱ تا ۵ انتخاب کردیم، بنابراین تعداد کل ترکیبات ممکن برابر با $۵ \times ۵ = ۲۵$ ترکیب است. برای هر ترکیب وزن، ویژگی‌های مختصات فضایی و ویژگی‌های تصویری را با وزن‌های مربوطه مقیاس‌دهی کردیم:

$$Features = w_{uvp} \times \begin{bmatrix} u \\ v \\ P_l \end{bmatrix} + w_{feat} \times \begin{bmatrix} f_۰ \\ f_۱ \\ f_۲ \\ \vdots \\ f_{۲۵۵} \end{bmatrix}$$

سپس الگوریتم خوشه‌بندی را بر روی داده‌های مقیاس‌دهی‌شده اعمال کرده و کیفیت خوشه‌بندی را ارزیابی کردیم.

۴-۲-۱ ارزیابی کیفیت خوشه‌بندی با استفاده از مدل سم

برای ارزیابی کیفیت خوشه‌بندی، نیاز به یک معیار داشتیم که بتواند میزان تطابق خوشه‌ها با اشیاء واقعی در تصویر را اندازه‌گیری کند. از آنجا که داده‌های برچسب‌دار در دسترس نبود، از یک مدل بخش‌بندی پیش‌آموزش‌دیده به نام سم استفاده کردیم [۳۲]. مدل سم یک مدل قدرتمند در بخش‌بندی تصویر است که قادر است بدون نیاز به ورودی‌های دستی، بخش‌بندی دقیق اشیاء را انجام دهد. با اعمال این

مدل بر روی تصاویر، نقشه‌های بخش‌بندی تولید شد که در آن هر پیکسل به یک بخش خاص تعلق دارد. شکل ۳-۱۰ خروجی حاصل از بخش‌بندی تصویر توسط مدل سم را نشان می‌دهد.

برای ارزیابی کیفیت خوشه‌بندی، برای هر نقطه در ابر نقاط، با استفاده از مختصات تصویری (u, v) آن، برچسب بخش‌بندی متناظر را از خروجی سم استخراج کردیم. سپس، برای هر خوشه، بررسی کردیم که تا چه حد نقاط آن در یک بخش واحد از نقشه بخش‌بندی قرار دارند. برای محاسبه میزان تطابق بین خوشه‌ها و بخش‌ها، از شاخص اطلاعات متقابل نرمال‌شده[1] استفاده کردیم [۳۳]. این معیار بین ۰ و ۱ متغیر است که مقادیر نزدیک به ۱ نشان‌دهنده تطابق بیشتر است.

شاخص NMI به صورت زیر تعریف می‌شود:

$$NMI(u, v) = \frac{۲ \times I(u; v)}{H(u) + H(v)}$$

که در آن $I(u; v)$ اطلاعات متقابل بین دو متغیر تصادفی u و v است، و $H(u)$ و $H(v)$ آنتروپی‌های مربوط به u و v هستند. اطلاعات متقابل $I(u; v)$ نشان‌دهنده میزان اطلاعات مشترک بین u و v است و به صورت زیر تعریف می‌شود:

$$I(u; v) = \sum_{u \in U} \sum_{v \in V} p(u, v) \log \left(\frac{p(u, v)}{p(u)p(v)} \right)$$

و آنتروپی $H(u)$ به صورت زیر تعریف می‌شود:

$$H(u) = -\sum_{u \in U} p(u) \log p(u)$$

در اینجا، u و v به ترتیب برچسب‌های خوشه‌بندی ما و برچسب‌های بخش‌بندی مدل سم هستند. با محاسبه شاخص اطلاعات متقابل، می‌توانیم میزان تطابق بین دو دسته‌بندی را اندازه‌گیری کنیم. مقادیر بالاتر شاخص اطلاعات متقابل نشان‌دهنده‌ی همبستگی بیشتر بین خوشه‌ها و بخش‌ها است.

با اجرای جستجوی شبکه‌ای، ترکیبی از وزن‌ها که بهترین عملکرد را داشت، شناسایی شد. به طور کلی، مشاهده کردیم که افزایش وزن ویژگی‌های تصویری نسبت به مختصات فضایی منجر به بهبود خوشه‌بندی اشیاء کوچک مانند خودروها و عابران پیاده می‌شود. این امر منطقی است زیرا ویژگی‌های

[1] NMI (Normalized Mutual Information)

تصویری اطلاعات دقیق‌تری در مورد محتوای بصری نقاط ارائه می‌دهند. از سوی دیگر، برای اشیاء بزرگ‌تر مانند ساختمان‌ها و دیوارها، تنظیم وزن‌های مختصات فضایی اهمیت بیشتری داشت.

۴-۲-۲ ارزیابی کیفیت الگوریتم دی‌بی‌اسکن قبل و بعد از حذف نقاط زمین

در ابتدا، بدون حذف نقاط زمین، از الگوریتم دی‌بی‌اسکن برای خوشه‌بندی داده‌ها استفاده کردیم. جدول ۴-۱ نتایج به‌دست‌آمده را برای ۵ ترکیب وزنی که بهترین نتایج را داشتند نشان می‌دهد. در اینجا، با استفاده از جستجوی شبکه‌ای، بیش از ۲۵۰ ترکیب وزنی مختلف برای وزن‌های ویژگی‌های تصویری و مختصات ابر نقاط امتحان شده‌اند، و ۵ ترکیب وزنی با بهترین عملکرد در جدول فهرست شده‌اند.

جدول ۴-۱: نتایج الگوریتم دی‌بی‌اسکن قبل از حذف نقاط زمین با استفاده از مدل دیپ‌لب‌وی۳

Combination	Coordinate Weight	Feature Weight	ε	MinPts	NMI
1	1	4	0.5	10	0.3694
2	1	4	0.4	5	0.3659
3	1	4	0.4	6	0.3671
4	1	4	0.5	11	0.3632
5	1	4	0.4	7	0.3535

در این مرحله، وزن ویژگی‌های تصویری (۴) نسبت به وزن مختصات فضایی (۱) بیشتر در نظر گرفته شده است، که نشان می‌دهد ویژگی‌های تصویری تأثیر بیشتری در خوشه‌بندی دارند. مقادیر اطلاعات متقابل نرمال‌شده در حدود ۰.۳۶ قرار دارند، که نشان‌دهنده‌ی کیفیت متوسط خوشه‌بندی است. وزن بیشتر برای ویژگی‌های تصویری باعث شده است که الگوریتم بیشتر به شباهت‌های بصری توجه کند. همچنین، مقادیر اطلاعات متقابل نرمال‌شده نسبتاً پایین هستند، که این مسئله می‌تواند به دلیل تأثیر منفی نقاط زمین و نویزهای مرتبط باشد.

با حذف نقاط زمین، انتظار می‌رود که عملکرد خوشه‌بندی بهبود یابد، زیرا نویزها و نقاط غیرمرتبط حذف شده‌اند. جدول ۴-۲ نتایج به‌دست‌آمده را نشان می‌دهد.

جدول ۴-۲: نتایج الگوریتم دی‌بی‌اسکن پس از حذف نقاط زمین با استفاده از مدل دیپ‌لب‌وی۳

Combination	Coordinate Weight	Feature Weight	ε	MinPts	NMI
1	3	5	0.5	10	0.4070
2	2	5	0.5	10	0.4068
3	1	5	0.5	10	0.4068
4	3	4	0.5	10	0.3652
5	2	4	0.5	10	0.3648

با مقایسه جداول ۴-۱ و ۴-۲، مشاهده می‌شود که حذف نقاط زمین باعث بهبود مقادیر اطلاعات متقابل نرمال‌شده در تمامی بخش‌های مجموعه داده شده است. همچنین، افزایش وزن مختصات فضایی پس از حذف نقاط زمین نشان می‌دهد که موقعیت مکانی نقاط اهمیت بیشتری پیدا کرده است. مقادیر اطلاعات متقابل نرمال‌شده به حدود ۰.۴ افزایش یافته‌اند، که این مسئله نشان‌دهنده بهبود کیفیت خوشه‌بندی است.

با داشتن خوشه‌های بهبودیافته، می‌توانیم آن‌ها را به پیشنهادات ناحیه تبدیل کنیم. برای هر خوشه، جعبه محدودکننده[۲] را با استخراج مختصات تصویری u و v نقاط خوشه و محاسبه کمینه و بیشینه u و v تعیین کردیم. سپس مستطیلی با گوشه‌های (u_{min}, v_{min}) و (u_{max}, v_{max}) رسم کردیم. شکل ۳-۱۳ خروجی پیشنهاد ناحیه انجام‌شده با روش ذکرشده را نشان می‌دهد و شکل ۳-۱۴ مربوط به جعبه‌های محدودکننده داده‌های واقعی است.

۴-۳ ارزیابی پیشنهادات ناحیه با استفاده از داده‌های واقعی

برای ارزیابی دقیق عملکرد مدل و سنجش کیفیت پیشنهادات ناحیه، از برچسب‌های واقعی موجود در مجموعه داده کیتی استفاده کردیم. این برچسب‌ها شامل جعبه‌های محدودکننده دوبعدی اشیاء مختلف مانند خودروها، عابران پیاده و دوچرخه‌سواران هستند. با استفاده از این داده‌های واقعی، معیارهای ارزیابی استانداردی را محاسبه کردیم که به ما امکان می‌دهد عملکرد مدل را به صورت دقیق و قابل اعتماد بسنجیم.

برای مقایسه پیشنهادات ناحیه مدل با برچسب‌های واقعی، معیارهای زیر را محاسبه کردیم:

تلاقی بر اتحاد[۳]: این معیار میزان هم‌پوشانی بین جعبه‌های محدودکننده پیشنهادشده و جعبه‌های واقعی را اندازه‌گیری می‌کند و به صورت زیر تعریف می‌شود:

$$IoU = \frac{\text{مساحت تلاقی دو جعبه}}{\text{مساحت اتحاد دو جعبه}}$$

مقادیر تلاقی بر اتحاد بین ۰ و ۱ قرار دارند؛ مقادیر نزدیک به ۱ نشان‌دهنده هم‌پوشانی بیشتر و دقت بالاتر در تشخیص ناحیه صحیح است.

[2] Bounding Box

[3] IoU (Intersection over Union)

تعیین مثبت‌های واقعی، مثبت‌های کاذب و منفی‌های کاذب: اگر مقدار تلاقی بر اتحاد بین یک پیشنهاد ناحیه و یک جعبه واقعی بیش از یک آستانه تعیین‌شده (مثلاً ۰.۵) باشد، آن پیشنهاد ناحیه به عنوان مثبت واقعی در نظر گرفته می‌شود. پیشنهادات ناحیه‌ای که مقدار تلاقی بر اتحاد آن‌ها با هیچ یک از جعبه‌های واقعی از آستانه تعیین‌شده بیشتر نباشد، به عنوان مثبت کاذب در نظر گرفته می‌شوند. جعبه‌های واقعی که توسط هیچ یک از پیشنهادات ناحیه پوشش داده نشده‌اند، به عنوان منفی کاذب در نظر گرفته می‌شوند.

محاسبه معیارهای دقت[۴] **و بازخوانی**[۵]: با استفاده از مقادیر مثبت واقعی[۶]، منفی واقعی[۷] و منفی کاذب[۸]، معیارهای **دقت** و **بازخوانی** را محاسبه کردیم:

$$\text{دقت} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{بازخوانی} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

این معیارها به ما کمک می‌کنند تا عملکرد مدل را در تشخیص صحیح اشیاء و میزان خطای آن بسنجیم.

استفاده از برچسب‌های واقعی مجموعه داده کیتی برای ارزیابی مدل اهمیت زیادی دارد، زیرا با داشتن داده‌های واقعی، می‌توانیم عملکرد مدل را به صورت دقیق و قابل اعتماد ارزیابی کنیم. همچنین با تحلیل موارد مثبت منفی و منفی کاذب، می‌توانیم نقاط ضعف مدل را شناسایی کرده و برای بهبود آن‌ها اقدام کنیم. استفاده از معیارهای استاندارد و داده‌های واقعی، امکان مقایسه منصفانه مدل ما با مدل‌های دیگر را فراهم می‌کند.

همان‌طور که در فصل‌های قبل گفته شد، برای بهبود دقت خوشه‌بندی، خوشه‌های بسیار کوچک و بسیار بزرگ را حذف کردیم. خوشه‌هایی با تعداد نقاط کم به عنوان نویز در نظر گرفته شدند و حذف شدند، زیرا معمولاً ناشی از نویز سنسورها یا خطاهای خوشه‌بندی هستند. همچنین، حذف خوشه‌های

[4] Precision
[5] Recall
[6] TP (True Positive)
[7] FP (False Positive)
[8] FN (False Negative)

بسیار بزرگ که احتمالاً مربوط به ساختارهایی مانند دیوارها یا زمین هستند، تمرکز بر روی اشیاء مهمی مانند خودروها و عابران پیاده را افزایش داد.

۴-۴ مقایسه مدل پیشنهادی با مدل‌های پایه

در این بخش، عملکرد مدل پیشنهادی را با مدل فَستر آر-سی ان ان مقایسه می‌کنیم. برای این منظور، از ویژگی‌های استخراج‌شده توسط فَستر آر-سی ان ان در مدل خود استفاده کردیم و الگوریتم کا میانگین را برای خوشه‌بندی به‌کار بردیم.

در مقایسه با مدل فَستر آر-سی ان ان، از ویژگی‌های استخراج‌شده توسط شبکه رزنت-۵۰ به عنوان ستون‌فقرات استفاده کردیم. این شبکه، نقشه‌های ویژگی با ۲۵۶ کانال تولید می‌کند که اطلاعات غنی از محتوای تصویر را در بر دارند. برای داشتن یک مبنای مقایسه مشترک و اطمینان از اینکه تفاوت‌ها ناشی از روش پیشنهاد ناحیه است نه ویژگی‌های استخراج‌شده، ویژگی‌های استخراج‌شده توسط رزنت-۵۰ را در مدل خود به کار بردیم. به این ترتیب، تفاوت عملکرد مدل ما و فَستر آر-سی ان ان بیشتر به روش‌های پیشنهاد ناحیه مربوط می‌شود تا به تفاوت در ویژگی‌های استخراج‌شده. نتایج اطلاعات متقابل نرمال‌شده به‌دست‌آمده از مدل پیشنهادی برای ۵ ترکیب وزنی با بهترین عملکرد، در جدول ۴-۳ آمده است.

جدول ۴-۳: نتایج مدل پیشنهادی با استفاده از ویژگی‌های فَستر آر-سی ان ان و الگوریتم کا میانگین

Combination	Coordinate Weight	Feature Weight	Number of Clusters	NMI
1	3	4	13	0.5439
2	3	3	13	0.5406
3	2	3	13	0.5387
4	3	3	9	0.5359
5	3	3	11	0.5358

مقادیر اطلاعات متقابل نرمال‌شده کمی کمتر از نتایج با مدل دیپ‌لب‌وی۳ است، اما همچنان در محدوده قابل قبول قرار دارد. استفاده از ویژگی‌های فَستر آر-سی ان ان نیز به دلیل هماهنگی با مدل مرجع، مقایسه منصفانه‌تری را فراهم می‌کند.

برای مقایسه دقیق‌تر مدل‌ها، معیارهای مختلفی از جمله **تلاقی بر اتحاد (IoU)**، تعداد **مثبت‌های واقعی (TP)**، **مثبت‌های کاذب (FP)** و **منفی‌های کاذب (FN)**، **دقت**، **بازخوانی** و **زمان اجرا** را در هر سه روش محاسبه کردیم. این محاسبات بر روی تمامی داده‌های آموزش که شامل ۵۰۰۰ تصویر است، انجام شد. سپس مقادیر به‌دست‌آمده را میانگین‌گیری کرده و از آن‌ها برای محاسبه دقت و بازخوانی

استفاده کردیم.

مدل پیشنهادی ما به‌طور متوسط ۱۳۵ پیشنهاد ناحیه در هر تصویر تولید می‌کند، در حالی که مدل فَستر آر-سی ان ان تعداد ۲۰۰۰ و روش جستجوی انتخابی تعداد ۳۵۲۳ پیشنهاد ناحیه ارائه می‌دهند. در جدول ۴-۴ نتایج میانگین معیارهای ارزیابی برای مدل پیشنهادی، فَستر آر-سی ان ان و جستجوی انتخابی ارائه شده است.

جدول ۴-۴: مقایسه معیارهای ارزیابی بین مدل پیشنهادی و فَستر آر-سی ان ان و جستجو انتخابی

Model	اتحاد بر تلاقی	TP	FP	FN	Precision	Recall	# Boxes
Ours (K-Means)	0.075	93	17	25	84%	78%	135
Faster R-CNN	0.052	849	757	394	52%	68%	2000
Selective Search	0.050	1470	1224	829	54%	63%	3523

نتایج به‌دست‌آمده نشان می‌دهد که مدل ما با تولید تعداد کمتری پیشنهاد ناحیه و دقت بالاتر، مثبت‌های کاذب کمتری دارد، که در کاربردهای بلادرنگ با منابع محاسباتی محدود بسیار مفید است. میانگین تلاقی بر اتحاد در مدل پیشنهادی ما بالاتر از مدل‌های فَستر آر-سی ان ان و جستجوی انتخابی است (۰.۰۷۵ در مقابل ۰.۰۵۲ و ۰.۰۵۰). این نشان می‌دهد که جعبه‌های محدودکننده‌ی پیشنهادی ما هم‌پوشانی بیشتری با جعبه‌های واقعی دارند و در تعیین دقیق موقعیت اشیاء عملکرد بهتری ارائه می‌دهند.

در مورد تعداد مثبت‌های واقعی، مثبت‌های کاذب و منفی‌های کاذب، مدل ما به دلیل تعداد کمتر پیشنهادات ناحیه، تعداد مثبت‌های واقعی کمتری دارد. با این حال، نسبت مثبت‌های واقعی به مجموع مثبت‌های واقعی و مثبت‌های کاذب در مدل ما بالاتر است، که منجر به دقت بالاتری می‌شود. دقت مدل ما ۸۴ درصد است، در حالی که دقت مدل‌های فَستر آر-سی ان ان و جستجوی انتخابی به ترتیب ۵۲ درصد و ۵۴ درصد است. این امر نشان می‌دهد که مدل ما مثبت‌های کاذب کمتری تولید می‌کند.

بازخوانی مدل ما ۷۸ درصد است، که بیشتر از بازخوانی مدل فَستر آر-سی ان ان با ۶۸ درصد و مدل جستجوی انتخابی با ۶۳ درصد است. این افزایش بازخوانی نشان می‌دهد که مدل ما توانایی بیشتری در شناسایی اشیاء دارد. این می‌تواند به دلیل تعداد پیشنهادات ناحیه کافی و بهینه در مدل ما باشد که منجر به پوشش بهتر اشیاء می‌شود. با این حال، با تنظیم مناسب پارامترهای مدل، می‌توان به توازن بهتری بین دقت و بازخوانی دست یافت.

برای ارائه یک مقایسه جامع‌تر، جدول تجمیعی زیر را تهیه کردیم که عملکرد کلی هر مدل را بر روی کل مجموعه‌داده ۵,۰۰۰ تصویری نشان می‌دهد. در این جدول، تعداد کل اشیاء موجود در مجموعه‌داده

کیتی برابر با ۱۴,۸۱۱ شیء است که به طور میانگین هر تصویر بین ۲ الی ۳ شیء را در خود دارد. همچنین، تعداد کل پیشنهادات ناحیه[9]، تعداد مثبت‌های واقعی، مثبت‌های کاذب و منفی‌های کاذب که توسط هر مدل شناسایی شده‌اند، ارائه شده است. این جدول به ما امکان می‌دهد تا عملکرد کلی هر مدل را در تشخیص صحیح اشیاء و خطاهای مرتکب‌شده مشاهده کنیم.

جدول ۴-۵: مقایسه تجمعی معیارهای ارزیابی بین مدل پیشنهادی و فَستر آر-سی‌اِن‌اِن و جستجوی انتخابی

Model	Total Proposals	TP	FP	FN
Ours (K-Means)	685,137	10,058	2,197	3,253
Faster R-CNN	10,000,000	10,071	9,227	4,740
Selective Search	17,612,358	9,332	7,932	5,479

از جدول ۴-۵ مشاهده می‌شود که مدل ما تعداد مثبت‌های واقعی مشابهی با مدل‌های دیگر دارد (۱۰,۰۵۸ در مقابل ۱۰,۰۷۱ و ۹,۳۲۲). همچنین، مدل ما تعداد مثبت‌های کاذب بسیار کمتری تولید کرده است (۲,۱۹۷ در مقابل ۹,۲۲۷ و ۷۹۳۲)، که نشان‌دهنده دقت بالاتر آن است. تعداد منفی‌های کاذب مدل ما نیز کمتر است (۳,۵۲۳ در مقابل ۴,۷۴۰ و ۵,۴۷۹)، که به معنای بازخوانی بالاتر و از دست دادن اشیاء کمتر است.

مدل ما با تولید تعداد کمتری پیشنهاد ناحیه (۶۸۵,۱۳۷ در مقابل ۱۰,۰۰۰,۰۰۰ و ۱۷,۶۱۲,۳۵۸)، کارایی بهتری را نشان می‌دهد، زیرا منابع محاسباتی کمتری مصرف می‌کند و از ایجاد پیشنهادات ناحیه غیرضروری و تصادفی جلوگیری می‌کند. این امر نشان‌دهنده این است که مدل ما به جای پیشنهاد دادن تعداد زیادی ناحیه به امید پوشش دادن اشیاء، با دقت بیشتری نواحی مرتبط را شناسایی می‌کند.

در مجموع، نتایج جدول ۴-۵ نشان می‌دهد که مدل پیشنهادی ما با داشتن تعداد مثبت‌های واقعی مشابه با مدل‌های پیشرفته، تعداد مثبت‌های کاذب و منفی‌های کاذب کمتری تولید می‌کند و از تولید پیشنهادات ناحیه غیرضروری جلوگیری می‌کند. این امر نشان‌دهنده کارایی بالاتر و دقت بیشتر مدل ما در مقایسه با مدل‌های دیگر است.

همچنین، همان‌طور که قبلاً نشان دادیم، مدل ما به صورت میانگین و تجمعی تعداد کمتری پیشنهادات ناحیه تولید می‌کند که این امر می‌تواند در مراحل بعدی طبقه‌بندی باعث کاهش زمان و بار محاسباتی شود. با این حال، طبق جدول ۴-۶ زمان اجرای مدل ما بیشتر است که این افزایش زمان

[9] Total Proposals

به دلیل مراحل پیش‌پردازش اضافی مانند ادغام داده‌های لیدار و حذف نقاط زمین است. با توجه به کاهش تعداد پیشنهادات ناحیه، انتظار می‌رود که زمان کل پردازش در کل سامانه بهبود یابد.

جدول ۴-۶: مقایسه تعداد پیشنهادات ناحیه و زمان اجرا بین مدل‌ها

Model	Execution Time (seconds)
Ours (K-Means)	1.54
Faster R-CNN	0.51
Selective Search	11.20

۴-۵ تحلیل کلی و نتیجه‌گیری

در این فصل، نتایج به‌دست‌آمده از پیاده‌سازی مدل‌های مختلف را ارائه و تحلیل کردیم. مهم‌ترین یافته‌ها نشان دادند که حذف نقاط زمین تأثیر مثبتی بر عملکرد الگوریتم‌های خوشه‌بندی داشت. همچنین، تنظیم مناسب وزن‌های مختصات فضایی و ویژگی‌های تصویری منجر به بهبود کیفیت خوشه‌بندی شد. الگوریتم کا میانگین، با وجود پیچیدگی محاسباتی مناسب، نتایج بهتری نسبت به دی‌بی‌اسکن ارائه داد. در مقایسه با فَستر آر-سی ان ان و جستجوی انتخابی، مدل ما با تعداد کمتری پیشنهادات ناحیه، دقت بالاتری را ارائه کرد. با این حال، زمان اجرای مدل ما بیشتر بود، اما کاهش تعداد پیشنهادات ناحیه می‌تواند در مراحل بعدی پردازش جبران شود.

نتایج به‌دست‌آمده نشان می‌دهد که روش پیشنهادی می‌تواند به عنوان یک راهکار مؤثر برای تولید پیشنهادات ناحیه در سامانه‌های تشخیص شیء مورد استفاده قرار گیرد. استفاده از داده‌های ادغام‌شده دوربین و لیدار، همراه با روش‌های خوشه‌بندی مناسب، می‌تواند راهکاری مؤثر و کارآمد برای تولید پیشنهادات ناحیه در سامانه‌های تشخیص شیء باشد.

فصل پنجم

نتیجه‌گیری و کارهای آینده

در این فصل، به جمع‌بندی نتایج به‌دست‌آمده از پژوهش انجام‌شده می‌پردازیم و به‌طور خلاصه دستاوردهای تحقیق را مرور می‌کنیم. سپس، به بیان پیشنهاداتی برای کارهای آینده می‌پردازیم که می‌تواند به بهبود و توسعه بیشتر این پژوهش کمک کند.

۵-۱ نتیجه‌گیری

در این پژوهش، به بررسی کارایی روش‌های خوشه‌بندی بر روی داده‌های ادغام‌شده‌ی دوربین و لیدار برای تولید پیشنهادات ناحیه در سامانه‌های تشخیص شیء پرداختیم. هدف اصلی ما ارزیابی میزان مؤثر بودن این روش در مقایسه با روش‌های پیشرفته‌ای مانند فَستر آر-سی ان ان و جستوجوی انتخابی بود.

ابتدا با استفاده از مدل دیپ‌لب‌وی۳ و سپس ویژگی‌های استخراج‌شده توسط فَستر آر-سی ان ان، داده‌های تصویری را پردازش کردیم و با ادغام آن‌ها با داده‌های لیدار، به خوشه‌بندی نقاط پرداختیم. سپس با استفاده از الگوریتم‌های مختلف خوشه‌بندی مانند کا میانگین و دی‌بی‌اسکن، پیشنهادات ناحیه را تولید کردیم.

نتایج به‌دست‌آمده نشان داد که روش پیشنهادی ما توانسته است با تولید تعداد کمتری پیشنهاد ناحیه، دقت بالاتری را در تعیین موقعیت اشیاء ارائه دهد. با این حال، زمان اجرای مدل ما نسبت به روش‌های پیشرفته کندتر بود؛ به‌طوری‌که زمان اجرای آن تقریباً سه برابر بیشتر از فَستر آر-سی ان ان بود.

با توجه به این نتایج، می‌توان گفت که هدف ما در این پژوهش ارزیابی کارایی این روش و سنجش میزان مؤثر بودن آن بود. نتایج به‌دست‌آمده تا حدی امیدوارکننده است، اما برای استفاده عملی از این روش، نیاز به بهبودهای بیشتری وجود دارد. به‌ویژه، باید بررسی کنیم که اگر مدل ما پیشنهادات ناحیه را به یک طبقه‌بند مانند فَستر آر-سی ان ان ارائه دهد، سرعت کلی سامانه چقدر بهبود می‌یابد و آیا کاهش تعداد پیشنهادات ناحیه می‌تواند زمان کل پردازش را کاهش دهد یا خیر.

۵-۲ پیشنهادات برای کارهای آینده

برای بهبود و توسعه بیشتر این پژوهش، پیشنهادات زیر ارائه می‌شود:

۵-۲-۱ استفاده از تکمیل عمق برای بهبود کیفیت پیشنهادات ناحیه

یکی از راه‌های بهبود کیفیت پیشنهادات ناحیه، استفاده از تکنیک‌های **تکمیل عمق**[۱] است. این تکنیک‌ها با استفاده از داده‌های تصویری، نقاط لیدار را تکمیل کرده و نقشه‌های عمق با رزولوشن بالاتر تولید می‌کنند. استفاده از این نقشه‌های عمق تکمیل‌شده می‌تواند به بهبود دقت خوشه‌بندی و در نتیجه، بهبود کیفیت پیشنهادات ناحیه منجر شود.

مدل‌هایی مانند تکمیل عمیق[۲] [۳۴] و تبدیل پراکنده به متراکم[۳] [۳۵] می‌توانند برای تکمیل عمق مورد استفاده قرار گیرند. افزودن این مرحله به مدل پیشنهادی، احتمالاً خروجی‌های دقیق‌تری را فراهم خواهد کرد.

۵-۲-۲ استفاده از الگوریتم‌های خوشه‌بندی پیشرفته

استفاده از الگوریتم‌های خوشه‌بندی سریع‌تر که نیاز به تعیین تعداد خوشه‌ها نداشته باشند، می‌تواند به بهبود سرعت مدل کمک کند. الگوریتم‌هایی مانند شیفت متوسط[۴] [۳۶] و انتشار همبستگی[۵] [۳۷] از این دسته هستند. این الگوریتم‌ها بدون نیاز به تعیین تعداد خوشه‌ها، داده‌ها را بر اساس چگالی نقاط خوشه‌بندی می‌کنند. با به‌کارگیری این الگوریتم‌ها، می‌توان سرعت پردازش را افزایش داد و نیاز به تنظیم تعداد خوشه‌ها را برطرف کرد.

۵-۲-۳ آموزش مدل مانند فَستر آر-سی ان ان با استفاده از روش پیشنهادی

یکی دیگر از راه‌های بهبود، آموزش یک مدل تشخیص شیء مانند فَستر آر-سی ان ان با استفاده از پیشنهادات ناحیه تولیدشده توسط مدل ما است. با این کار، می‌توان سرعت یادگیری، سرعت اجرای مدل و دقت آن را پس از آموزش بر روی مجموعه داده‌هایی مانند کوکو[۶] [۳۸] یا کیتی ارزیابی و با مدل‌های استاندارد مقایسه کرد. این ارزیابی می‌تواند نشان دهد که کاهش تعداد پیشنهادات ناحیه چگونه بر عملکرد کلی سامانه تأثیر می‌گذارد.

[1] Depth Completion
[2] Deep Completion
[3] Sparse-to-Dense
[4] Mean Shift
[5] Affinity Propagation
[6] COCO (Common Objects in Context)

۵-۳ جمع‌بندی

به‌طور کلی، نتایج این پژوهش نشان می‌دهد که استفاده از روش‌های خوشه‌بندی بر روی داده‌های ادغام‌شده دوربین و لیدار می‌تواند بهبودهایی در دقت تعیین موقعیت اشیاء ارائه دهد. با این حال، به دلیل کندتر بودن مدل پیشنهادی نسبت به روش‌های پیشرفته، نیاز به بهینه‌سازی و بهبودهای بیشتری وجود دارد. با اجرای پیشنهادات ارائه‌شده در بخش کارهای آینده، می‌توان امیدوار بود که مدل بهبود یافته و به کارایی عملی نزدیک‌تر شود.

منابع و مراجع

[1] Redmon, Joseph, Divvala, Santosh, Girshick, Ross, and Farhadi, Ali. You only look once: Unified, real-time object detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, 2016.

[2] Ren, Shaoqing, He, Kaiming, Girshick, Ross, and Sun, Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6):1137–1149, 2016.

[3] Girshick, Ross, Donahue, Jeff, Darrell, Trevor, and Malik, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, 2014.

[4] Meschtscherjakov, Alexander, Tscheligi, Manfred, Pfleging, Bastian, Borojeni, Shadan Sadeghian, Ju, Wendy, Palanque, Philippe, Riener, Andreas, Mutlu, Bilge, and Kun, Andrew L. Interacting with autonomous vehicles: Learning from other domains. in Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI EA '18', p. 1–8, New York, USA, 2018.

[5] Pomerleau, Dean. Alvinn: An autonomous land vehicle in a neural network. in Proceedings of Neural Information Processing Systems, pp. 305–313, 1989.

[6] Thrun, Sebastian, Montemerlo, Mike, Dahlkamp, Hendrik, Stavens, David, Aron, Andrei, Diebel, James, Fong, Philip, Gale, John, Halpenny, Morgan, Hoffmann, Gabriel, Lau, Kenny, Oakley, Celia, Palatucci, Mark, Pratt, Vaughan, Stang, Pascal, Strohband, Sven, Dupont, Cedric, Jendrossek, Lars-Erik, Koelen, Christian, Markey, Charles, Rummel, Carlo, Niekerk, Joe Van, Jensen, Eric, Alessandrini, Philippe, Bradski, Gary,

Davies, Bob, Ettinger, Scott, Kaehler, Adrian, Nefian, Ara, and Mahoney, Pamela. Stanley: The robot that won the darpa grand challenge. Journal of Field Robotics, 23(9):661–692, 2006.

[7] Levinson, Jesse, Askeland, Jake, Becker, Jan, Dolson, Jennifer, Held, David, Kammel, Soeren, Kolter, J. Zico, Langer, Dirk, Pink, Oliver, Pratt, Vaughan, Sokolsky, Michael, Stanek, Ganymed, Stavens, David, Teichman, Alex, Werling, Moritz, and Thrun, Sebastian. Towards fully autonomous driving: Systems and algorithms. in 2011 IEEE Intelligent Vehicles Symposium, pp. 163–168, 2011.

[8] Liu, Li, Ouyang, Wanli, Wang, Xiaogang, Fieguth, Paul, Chen, Jie, Liu, Xinwang, and Pietikäinen, Matti. Deep learning for generic object detection: A survey. International Journal of Computer Vision, 128(2):261–318, 2020.

[9] Viola, Paul and Jones, Michael. Rapid object detection using a boosted cascade of simple features. in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. I–511–I–518, Kauai, Hawaii, USA, 2001.

[10] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017.

[11] Liu, Wei, Anguelov, Dragomir, Erhan, Dumitru, Szegedy, Christian, Reed, Scott, Fu, Cheng-Yang, and Berg, Alexander C. Ssd: Single shot multibox detector. in Leibe, Bastian, Matas, Jiri, Sebe, Nicu, and Welling, Max, eds. , Computer Vision – ECCV 2016, pp. 21–37, Cham, Switzerland, 2016. Springer International Publishing.

[12] Uijlings, Jasper R. R., van de Sande, Koen E. A., Gevers, Theo, and Smeulders, Arnold W. M. Selective search for object recognition. International Journal of Computer Vision, 104(2):154–171, 2013.

[13] Zitnick, C. Lawrence and Dollár, Piotr. Edge boxes: Locating object proposals from edges. in Fleet, David, Pajdla, Tomas, Schiele, Bernt, and Tuytelaars, Tinne, eds. , Computer Vision – ECCV 2014, pp. 391–405, Cham, Switzerland, 2014. Springer International Publishing.

[14] Girshick, Ross. Fast r-cnn. in Proceedings of the IEEE International Conference on Computer Vision, 2015.

[15] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Łukasz Kaiser, and Polosukhin, Illia. Attention is all you need. in Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds. , Advances in Neural Information Processing Systems, vol. 30, 2017.

[16] Carion, Nicolas, Massa, Francisco, Synnaeve, Gabriel, Usunier, Nicolas, Kirillov, Alexander, and Zagoruyko, Sergey. End-to-end object detection with transformers. in Vedaldi, Andrea, Bischof, Horst, Brox, Thomas, and Frahm, Jan-Michael, eds. , Computer Vision – ECCV 2020, pp. 213–229, Cham, Switzerland, 2020. Springer International Publishing.

[17] Liu, Shilong, Li, Feng, Zhang, Hao, Yang, Xiao, Qi, Xianbiao, Su, Hang, Zhu, Jun, and Zhang, Lei. Dab-detr: Dynamic anchor boxes are better queries for detr. in International Conference on Learning Representations, 2022.

[18] LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[19] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[20] Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. Clinical Orthopaedics and Related Research, abs/1409.1556, 2014.

[21] Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. in Navab, Nassir, Hornegger, Joachim, Wells, William M., and Frangi, Alejandro F., eds. , Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241, Cham, Switzerland, 2015. Springer International Publishing.

[22] Chen, Liang-Chieh, Papandreou, George, Schroff, Florian, and Adam, Hartwig. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.

[23] Chen, Liang-Chieh, Zhu, Yukun, Papandreou, George, Schroff, Florian, and Adam, Hartwig. Encoder-decoder with atrous separable convolution for semantic image segmentation. in Ferrari, Vittorio, Hebert, Martial, Sminchisescu, Cristian, and Weiss, Yair, eds. , Computer Vision – ECCV 2018, pp. 833–851, Cham, Switzerland, 2018. Springer International Publishing.

[24] Tian, Haofei, Chen, Yuntao, Dai, Jifeng, Zhang, Zhaoxiang, and Zhu, Xizhou. Unsupervised object detection with lidar clues. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5958–5968, 2021.

[25] Zhang, Lunjun, Yang, Anqi Joyce, Xiong, Yuwen, Casas, Sergio, Yang, Bin, Ren, Mengye, and Urtasun, Raquel. Towards unsupervised object detection from lidar point clouds. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9317–9328, Vancouver, Canada, 2023. IEEE.

[26] Bai, Xuyang, Hu, Zeyu, Zhu, Xinge, Huang, Qingqiu, Chen, Yilun, Fu, Hangbo, and Tai, Chiew-Lan. Transfusion: Robust lidar-camera fusion for 3d object detection with

transformers. in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1080–1089, 2022.

[27] Li, Yingwei, Yu, Adams Wei, Meng, Tianjian, Caine, Ben, Ngiam, Jiquan, Peng, Daiyi, Shen, Junyang, Lu, Yifeng, Zhou, Denny, Le, Quoc V., Yuille, Alan, and Tan, Mingxing. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17161–17170, New Orleans, USA, 2022. IEEE.

[28] Liu, Leyuan, He, Jian, Ren, Keyan, Xiao, Zhonghua, and Hou, Yibin. A lidar–camera fusion 3d object detection algorithm. Information, 13(4), 2022.

[29] Geiger, Andreas, Lenz, Philip, Stiller, Christoph, and Urtasun, Raquel. Vision meets robotics: The kitti dataset. International Journal of Robotics Research, 2013.

[30] Fischler, Martin A. and Bolles, Robert C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981.

[31] Bergstra, James, Bardenet, Rémi, Bengio, Yoshua, and Kégl, Balázs. Algorithms for hyper-parameter optimization. in Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K.Q., eds. , Advances in Neural Information Processing Systems, vol. 24, 2011.

[32] Kirillov, Alexander, Mintun, Eric, Ravi, Nikhila, Mao, Hanzi, Rolland, Chloe, Gustafson, Laura, Xiao, Tete, Whitehead, Spencer, Berg, Alexander C., Lo, Wan-Yen, Dollár, Piotr, and Girshick, Ross B. Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3992–4003, 2023.

[33] Strehl, Alexander and Ghosh, Joydeep. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3:583–617, 2002.

[34] Ma, Fangchang, Cavalheiro, Guilherme V., and Karaman, Sertac. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. Proceedings of the International Conference on Robotics and Automation, pp. 3288–3295, 2019.

[35] Uhrig, Jonas, Schneider, Nick, Schneider, Lukas, Franke, Uwe, Brox, Thomas, and Geiger, Andreas. Sparsity invariant cnns. in Proceedings of the International Conference on 3D Vision, pp. 11–20, 2017.

[36] Comaniciu, Dorin and Meer, Peter. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):603–619, 2002.

[37] Frey, Brendan J. and Dueck, Delbert. Clustering by passing messages between data points. Science, 315(5814):972–976, 2007.

[38] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge J., Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C. Lawrence. Microsoft coco: Common objects in context. in Proceedings of the European Conference on Computer Vision, 2014.

واژه‌نامه‌ی انگلیسی به فارسی

A

Affinity Propagation . . . . انتشار همبستگی

Autonomous Land Vehicle In a . . . . آلوین
Neural Network

ANN (Artificial . . . . شبکه عصبی مصنوعی
Neural Network)

Artificial Intelligence . . . . هوش مصنوعی

ASPP (Atrous Spatial Pyramid ای‌اس‌پی‌پی
Pooling)

B

Backbone . . . . . ستون فقرات

Bounding Box . . . . . جعبه محدودکننده

C

CNNs . . . . . شبکه‌های عصبی کانولوشنال
(Convolutional Neural Networks)

COCO (Common Objects in . . . . . کوکو
Context)

Computer Vision . . . . . بینایی ماشین

D

Data Fusion . . . . . ادغام داده‌ها

Deep Learning . . . . . یادگیری عمیق

DeepLab . . . . . دیپ‌لب

DETR . . . . . دیتر

Driverless Cars . . . . . خودروهای خودران

F

False Negative . . . . . منفی کاذب

False Positive . . . . . مثبت کاذب

Fast R-CNN . . . . . فست آر-سی‌ان‌ان

Faster R-CNN . . . . . فَستر آر-سی‌ان‌ان

G

GPS (Global . . . . سامانه موقعیت‌یاب جهانی
Positioning System)

Grid Search . . . . . جستجوی شبکه‌ای

Ground Truth . . . . . داده مرجع

H

Homogeneous Coordinates مختصات همگن

ابر نقطه . . . . . Point Cloud

دقت . . . . . Precision

پای‌تورچ . . . . . PyTorch

R

الگوریتم اجتماع تصادفی نمونه‌ها . RANSAC (Random Sample Consensus)

آر-سی‌ان‌ان . . . . . RCNN (Region-Based Convolutional Neural Network)

بازخوانی . . . . . Recall

واحد یکسوکننده‌ی خطی . ReLU (Rectified Linear Unit)

اتصالات باقی‌مانده . . Residual Connections

رزنت . ResNet (Residual Neural Network)

قرمز، سبز، آبی . . . . . RGB

شبکه پیشنهاد ناحیه . . . . Region Proposal Network

S

سم . . . . . Segment Anything Model

سایکیت‌لرن . . . . . Scikit-learn

جستجوی انتخابی . . . . . Selective Search

پنجره کشویی . . . . . Sliding Window

خوشه‌بندی طیفی . . . . Spectral Clustering

فراپارامتر . . . . . Hyperparameter

I

پردازش تصویر . . . . . Image Processing

ایمیج‌نت . . . . . ImageNet

تلاقی بر اتحاد . . . . . IoU (Intersection over Union)

K

کا میانگین . . . . . K-Means

کیتی . . . . . KITTI

L

لیدار . . . . . LIDAR (Light Detection and Ranging)

M

ادغام ماکسیمم . . . . . Max Pooling

شیفت متوسط . . . . . Mean Shift

N

شبکه عصبی . . . . . Neural Network

اطلاعات متقابل نرمال‌شده . . . . . NMI (Normalized Mutual Information)

O

تشخیص شیء . . . . . Object Detection

P

اس‌اس‌دی Single Shot MultiBox Detector

نزول گرادیان تصادفی . Stochastic Gradient Descent

T

مثبت حقیقی . . . . . True Positive

مبدل . . . . . Transformer

U

یو-نت . . . . . U-Net

یادگیری بدون نظارت . . . . . Unsupervised Learning

نمونه‌برداری بالا . . . . . Upsampling

V

مشکل ناپدید شدن گرادیان . . . . Vanishing Gradient Problem

وی‌جی‌جی . . . . . VGG

Y

یولو . . . . . YOLO (You Only Look Once)

Abstract

In recent years, the integration of multisensory data such as camera and LiDAR in object detection systems has garnered significant attention. This integration can help improve accuracy and efficiency in object recognition. However, the collection and manual labeling of data required to train object detection models is a time-consuming and costly process. In this research, we evaluate the performance of a proposed method for region proposal using unsupervised learning and the fusion of camera and LiDAR data. The main objective is to determine how this approach can be effective in the region proposal process and how it affects the accuracy and speed of object detection. To this end, two-dimensional camera images and three-dimensional LiDAR data are fused, and unsupervised clustering algorithms are employed to identify regions containing objects. The results indicate that this method can detect object-containing regions with acceptable accuracy without the need for labeled data, which can help reduce the costs and time required for developing object detection systems.

Key Words:

Region Proposal, Unsupervised Learning, Camera and LiDAR Data Fusion, Object Detection, Unsupervised Clustering