دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

پروژه تحقیقاتی درس شبکه عصبی و یادگیری عمیق

شبکه‌های عصبی کانولوشنی مبتنی بر منطقه

نگارش

زهرا اخلاقی

استاد درس

دکتر رضا صفابخش

دی ۱۴۰۲


بسم الله الرحمن الرحيم

چکیده

از حدود سال ۲۰۱۲ شبکه‌های عمیق در حوزه بینایی ماشین ظهور پیدا کردند، بینایی ماشین به ما کمک می‌کند پیچیدگی قوه بینایی انسان را درک کنیم و سیستم‌های رایانه‌ای را به‌گونه‌ای تربیت کنیم که بتوانند تصاویر و دیدیوهای دیجیتال را تفسیر و درک کنند. یکی از مسائل بینایی ماشین مساله کشف اشیاء است. حل این مساله به معنی پیدا کردن اشیاء مختلف در یک تصویر و تعیین یک مرز برای آن شکل است و تشخیص اشیاء با استفاده از یادگیری عمیق، یکی از چالش برانگیزترین مشکلات بینایی کامپیوتری است.

شبکه‌های عصبی کانولوشنی مبتنی بر منطقه یا نوعی از مدل‌های یادگیری عمیق هستند که به صورت خاص برای مسائل تشخیص اشیاء مشابه طراحی شده‌اند. این شبکه‌ها از مفاهیمی چون استخراج ویژگی‌ها، تجمیع مناطق و آموزش دقیق بر اساس اطلاعات مناطق مشخص برای تشخیص اشیاء در تصاویر استفاده می‌کنند. در این پژوهش به بررسی مدل شبکه عصبی کانولوشنال مبتنی بر منطقه و چند مورد از توسعه‌هایی که مبتنی بر آن صورت گرفته است می‌پردازیم.

واژه‌های کلیدی:

تشخیص اشیا، تکنیک های یادگیری عمیق، شبکه‌های عصبی کانولوشنی مبتنی بر منطقه

فهرست مطالب

عنوان صفحه

۱ پیشگفتار ۱

۱-۱ مقدمه ۲

۱-۲ سازماندهی گزارش ۴

۲ شبکه های عصبی کانولوشنی مبتنی بر منطقه ۵

۲-۱ مقدمه ۶

۲-۲ معرفی شبکه های عصبی کانولوشنی مبتنی بر منطقه ۶

۳ توسعه های شبکه های عصبی کانولوشنی مبتنی بر منطقه ۱۰

۳-۱ مقدمه ۱۱

۳-۲ مدل شبکه عصبی کانولوشنی مبتنی بر منطقه سریع ۱۱

۳-۳ مدل شبکه عصبی کانولوشن مبتنی بر منطقه سریع‌تر ۱۲

۳-۴ مدل تقویت‌شده شبکه عصبی کانولوشن مبتنی بر منطقه ۱۶

۳-۴-۱ استخراج ویژگی از تصاویر ۱۷

۳-۴-۲ RetinaRPN ۱۷

۳-۴-۳ خط لوله استنتاج احتمالی ۱۸

۳-۴-۴ وزن‌دهی مجدد ۱۸

۳-۵ توجه چند مقیاسی در شبکه عصبی کانولوشنی مبتنی بر منطقه ۱۹

۳-۵-۱ مؤلفه استخراج ویژگی ۲۲

۳-۵-۲ لایه نمونه‌گیری پایین ۲۳

۳-۵-۳ شبکه پیشنهادی منطقه ۲۴

۳-۵-۴ انتخاب منطقه موردنظر به صورت چند مقیاسی ۲۴

۳-۵-۵ لایه پیش بینی نهایی ۲۴

۳-۵-۶ شبکه ویژگی‌های هرمی اصلاح شده ۲۵

۳-۶ نقشه فعالسازی کلاس در شبکه عصبی کانولوشنی مبتنی بر منطقه ۲۶

۳-۶-۱ ماژول E-CAM ۲۷
۳-۶-۲ ماژول S-CAM ۲۸
۳-۷ تجمیع ویژگی ها بر اساس همسایگان برای تشخیص اشیاء سه بعدی ۳۰
۳-۷-۱ استخراج ویژگی ۳۱
۳-۷-۲ تجمیع ویژگی‌ها ۳۱

۴ جمع‌بندی و نتیجه‌گیری ۳۳
۴-۱ جمع‌بندی و نتیجه‌گیری ۳۴

منابع و مراجع ۳۶

فهرست اشکال

شکل صفحه

۱-۱ نمونه ای از خروجی مساله کشف اشیاء ۳

۱-۲ فرآیند کلی شبکه عصبی کانولوشنی مبتنی بر منطقه [۱] ۶

۲-۲ شبکه ImageNet برای استخراج ویژگی های نواحی پیشنهادی [۲] ۸

۲-۳ نمونه هایی از تصاویر کشیده شده برای تطبیق به ورودی شبکه ImageNet ۸

۳-۱ فرآیند کلی روش شبکه عصبی کانولوشنی مبتنی بر منطقه سریع [۳] ۱۱

۳-۲ مدل شبکه عصبی کانولوشن مبتنی بر منطقه سریعتر [۴] ۱۲

۳-۳ شبکه پیاده سازی شده با لایه کانولوشن [۵] ۱۳

۳-۴ اعمال شبکه کانولوشنی آموزش دیده روی تصاویر کوچکتر روی تصاویر بزرگتر ۱۴

۳-۵ نمونه ای از جعبه لنگرها ۱۵

۳-۶ فرمول محاسبه اناش ۱۵

۳-۷ تصاویر ثبت شده از زیر آب ۱۶

۳-۸ شبکه شعک منطقه تقویت‌شده [۶] ۱۷

۳-۹ ساختار RetinaRPN [۶] ۱۸

۳-۱۰ نمونه‌ای از وزن‌دهی مجدد برای تقویت پیشنهاد [۶] ۱۹

۳-۱۱ کاربرد درک صحنه در سنجش از دور ۲۰

۳-۱۲ ساختار کلی سیستم تشخیص شی برای درک تصاویر صحنه سنجش از دور [۷] . . . ۲۰

۳-۱۳ مدل تشخیص شی شعک منطقه چندمقیاسی [۷] ۲۱

۳-۱۴ ساختار کلی مدل SMENet برای استخراج ویژگی از تصاویر [۷] ۲۲

۳-۱۵ ساختار مرحله فردی و بخش بلوک SMENet [۷] ۲۳

۳-۱۶ ساختار لایه نمونه‌گیری پایین [۷] ۲۴

۳-۱۷ ساختار شویه اصلاح شده [۷] ۲۶

۳-۱۸ ماژول E-CAM [۸] ۲۷

۳-۱۹ ماژول S-CAM [۸] ۲۸

۳-۲۰ مدل تشخیص شی شعک منطقه نفک [۸] ۲۹

۳-۲۱مدل تشخیص شی NV2P-RCNN [۹] ۳۰
۳-۲۲استخراج یژگی در مدل NV2P-RCNN [۹] ۳۱

فهرست جداول

جدول صفحه

۴-۱ مقایسه مدل شعک منطقه سریع و سریع‌تر [۵] ۳۴
۴-۲ مقایسه مدل شعک منطقه تقویت‌شده با شعک منطقه سریع و سریع‌تر [۶] ۳۴

فهرست اختصارات

عنوان اختصاری	عنوان کامل
شعک	شبکه عصبی کانولوشنی
شعک منطقه	شبکه عصبی کانولوشنی مبتنی بر منطقه
مابرپ	ماشین بردار پشتیبان
اناش	انطباق بر روی اشتراک
دون	دور ترین نقطه نمونه
جمج	جمع‌بندی میانگین جهانی
شویه	شبکه ویژگی‌های هرمی
نفک	نقشه‌های فعال‌سازی کلاس

فصل اول

پیشگفتار

۱-۱ مقدمه

با بالغ شدن اینترنت در دهه ۱۹۹۰ و در دسترس قرار گرفتن مجموعه‌های زیادی از تصاویر به صورت آنلاین جهت تجزیه و تحلیل، مدل‌های بینایی ماشین [۱] رونق گرفت. همچنین، پیشرفت‌های سخت‌افزاری در کنار مجموعه داده‌های در حال رشد باعث شد ماشین‌ها بتوانند اجسام متنوعی را در عکس‌ها و فیلم‌ها شناسایی کنند.

در دهه‌های اخیر، تأثیرات این پیشرفت‌ها در زمینه بینایی ماشین، حیرت‌انگیز بوده بطوری‌که سیستم‌های امروزی در تشخیص و واکنش سریع به ورودی‌های بصری دقیق‌تر از انسان عمل می‌نمایند. لری رابرتز، نخستین فردی بود که با بررسی امکان استخراج اطلاعات هندسی سه بعدی از تصاویر دو بعدی بلوک‌ها (چندوجهی) حوزه بینایی کامپیوتر را مطرح نمود و پدر علم بینایی کامپیوتر لقب گرفت. پس از آن محققان بسیاری این کار را دنبال کرده و الگوریتم‌های مختلف بینایی کامپیوتر را در این زمینه مطرح کردند.

بینایی ماشین فرایند "دیدن" را در سیستم‌های هوشمند، امکان‌پذیر می‌نماید که به کمک آن می‌توان بسیاری از فعالیت‌ها که نیازمند شناخت بصری است را به‌طور خودکار انجام داد و به ماشین‌ها قدرت مشاهده محیط اطراف و تحلیل و پردازش اطلاعات پیرامون آن را می‌دهد. تأکید سیستم‌های بینایی ماشین، بیشتر بر روی قابلیت‌های تحلیل تصاویر، استخراج اطلاعات مفید از آن‌ها و درک و فهم اجسام موجود در آن‌هاست.

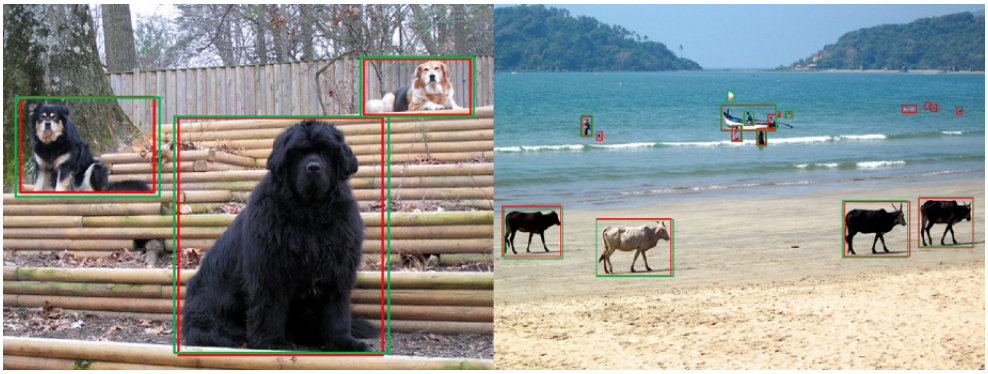
یکی از مسائل بینایی ماشین مساله کشف اشیاء [۲] است. تصویر زیر نمونه ای از خروجی این مساله است. ایده اولیه برای حل این مساله این است که فرض کنیم یک شبکه عصبی کانولوشنی [۳] داشته باشیم که برای دسته بندی تصاویر آموزش دیده است. حالا می توانیم یک پنجره کوچک را روی تصویر بلغزانیم هر بخش از تصویر را به شبکه مفروض بدهیم و از آنجا که اشیاء ممکن است به دلیل جلو و عقب بودن اندازه های مختلفی داشته باشند، همچنین می توانیم با پنجره های با اندازه های مختلف همین کار را تکرار کنیم.

---

[۱] Computer Vision

[۲] Object Detection

[۳] Convolutional Neural Network (CNN)



شکل ۱-۱: نمونه ای از خروجی مساله کشف اشیاء

ایراد اصلی که بر ایده خام بالا وارد است هزینه بسیار بالای محاسباتی است. شبکه‌های عصبی کانولوشنی مبتنی بر منطقه [۴] نوعی از مدل‌های یادگیری عمیق هستند که به صورت خاص برای مسائل تشخیص اشیاء و نظائر مشابه طراحی شده‌اند. این شبکه‌ها از مفاهیمی چون استخراج ویژگی‌ها، تجمیع مناطق و آموزش دقیق بر اساس اطلاعات مناطق مشخص برای تشخیص اشیاء در تصاویر استفاده می‌کنند.

تشخیص اشیاء با استفاده از شبکه عصبی کانولوشنی مبتنی بر منطقه یکی از وظایف مهم در حوزه بینایی ماشین و پردازش تصویر است. این وظیفه در بسیاری از زمینه‌های کاربردی مانند خودروهای هوشمند، پزشکی، امنیت، و افزایش هوش مصنوعی به کار می‌رود.

شبکه‌های عصبی کانولوشنی مبتنی بر منطقه به عنوان یک مدل پیشرو در زمینه تشخیص اشیاء معرفی شده است و در سال‌های اخیر به شدت مورد استفاده در برنامه‌ها و پروژه‌های بینایی ماشین، پردازش تصویر، و هوش مصنوعی قرار گرفته است. یکی از نقاط قوت این است که قادر به تشخیص و تمیز دادن اشیاء در تصاویر با دقت بالا است و می‌تواند با دقت بالا به شناسایی اجسام مختلف در دسته‌های مختلف بپردازد.

در این مدل، ابتدا تصویر به مناطق مختلف تقسیم می‌شود و سپس برای هر منطقه، ویژگی‌های خاصی استخراج می‌شود. سپس از این ویژگی‌ها برای تشخیص و دسته‌بندی اشیاء استفاده می‌شود. با استفاده از یک شبکه کانولوشنی برای استخراج ویژگی‌ها و الگوریتم‌های مناسب برای تشخیص مناطق مهم، به عنوان یکی از مدل‌های پیچیده و قدرتمند در زمینه تشخیص تصاویر شناخته می‌شود.

[۴] Region-based Convolutional Neural Network (R-CNN)

۱-۲ سازماندهی گزارش

در فصل دوم، شبکه عصبی کانولوشنی مبتنی بر منطقه به همراه جزئیات آن بررسی شده است. ابتدا به معرفی شبکه عصبی کانولوشنی مبتنی بر منطقه می‌پردازیم، سپس در مورد معماری آن توضیح می‌دهیم. هم‌چنین در رابطه با بازنمایی ورودی و خروجی در مدل و نحوه پیش‌آموزش آن صحبت می‌کنیم. به دنبال آن آزمایش‌ها و نتایجی که با این روش گرفته شده است ارائه می‌شوند.

در فصل سوم، در رابطه با چند نسخه مختلف شبکه عصبی کانولوشنی مبتنی بر منطقه که در واقع هر یک توسعه‌هایی را بر روی آن ارائه کرده‌اند صحبت می‌کنیم. در مورد هر یک از این توسعه‌ها تغییرات و اصلاحاتی که بر روی مدل ارائه کرده‌اند توضیح می‌دهیم و آزمایش‌ها و بهبودهایی که در نتایج حاصل شده است را دنبال می‌کنیم.

فصل چهارم نیز شامل جمع‌بندی و نتیجه‌گیری مواردی است که در این پژوهش مورد بررسی قرار گرفته است.

فصل دوم

شبکه های عصبی کانولوشنی مبتنی بر منطقه

۲-۱ مقدمه

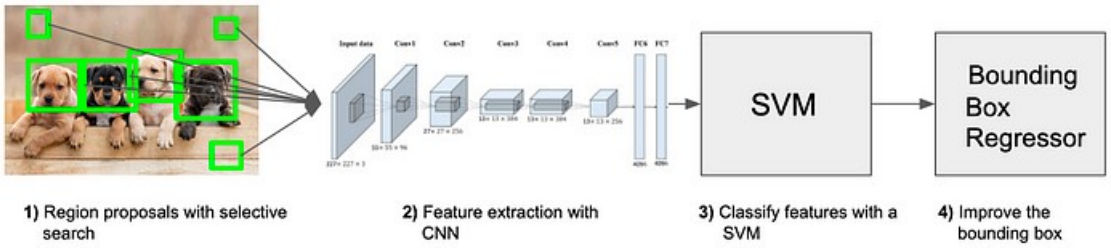
شبکه‌های عصبی کانولوشنی مبتنی بر منطقه یا نوعی از مدل‌های یادگیری عمیق هستند که به صورت خاص برای مسائل تشخیص اشیاء و نظائر مشابه طراحی شده‌اند. این شبکه‌ها از مفاهیمی چون استخراج ویژگی‌ها، تجمیع مناطق و آموزش دقیق بر اساس اطلاعات مناطق مشخص برای تشخیص اشیاء در تصاویر استفاده می‌کنند.

در ادامه این فصل، به جزئیات عملکرد و مزایا و معایب این شبکه خواهیم پرداخت تا خواننده با نحوه کارکرد و کاربردهای این مدل آشنا شود.

۲-۲ معرفی شبکه های عصبی کانولوشنی مبتنی بر منطقه

ایده اولیه برای حل مساله کشف اشیا در بینایی کامپیوتر این است که فرض کنیم یک شبکه عصبی کانولوشنی داشته باشیم که برای دسته بندی تصاویر آموزش دیده است. حالا می توانیم یک پنجره کوچک را روی تصویر بلغزانیم هر بخش از تصویر را به شبکه مفروض بدهیم و ببینیم آیا یک شی در آن پنجره قرار دارد یا خیر؟ و اگر پاسخ مثبت است آن شی چیست؟ و از آنجا که اشیاء ممکن است به دلیل جلو و عقب بودن اندازه های مختلفی داشته باشند می توانیم با پنجره های با اندازه های مختلف همین کار را تکرار کنیم. ایراد اصلی که بر این ایده وارد است هزینه بسیار بالای محاسباتی است.

مسئله‌ای که شبکه‌های عصبی کانولوشنی مبتنی بر منطقه سعی در حل آن دارد، مکان‌یابی اشیا در تصویر است. ایده اصلی این روش که در سال ۲۰۱۴ توسط راس گیرشیک و همکارانش مطرح شد، در این ایده گفته شده که به جای اینکه تمام قسمت های تصویر را برای پیدا کردن اشیاء بگردیم ابتدا نواحی از تصویر که احتمال وجود یک شی در آن ها وجود دارد را پیدا کنیم و سپس همان ایده ساده بالا رو فقط روی آن نواحی پیاده کنیم [۱].



شکل ۲-۱: فرآیند کلی شبکه عصبی کانولوشنی مبتنی بر منطقه [۱]

در این مدل، ابتدا تصویر به مناطق مختلف تقسیم می‌شود و سپس برای هر منطقه، ویژگی‌های خاصی استخراج می‌شود. سپس از این ویژگی‌ها برای تشخیص و دسته‌بندی اشیاء استفاده می‌شود. در این مدل با انتخاب یک شبکه کانولوشنی برای استخراج ویژگی‌ها و الگوریتم‌های مناسب برای تشخیص مناطق مهم، می‌تواند به عنوان یکی از مدل‌های پیچیده و قدرتمند در زمینه تشخیص تصاویر شناخته شود.
شبکه‌های عصبی کانولوشنی مبتنی بر منطقه از سه ماژول تشکیل شده:

- ماژول تولید نواحی پیشنهادی مستقل از اشیاء
- یک شبکه عصبی بزرگ کانولوشنی، برای استخراج بردار ویژگیهایی با طول ثابت از هر ناحیه
- تعدادی دسته بند ماشین بردار پشتیبان [۱] خطی برای هر کلاس از اشیاء

برای تعیین نواحی پیشنهادی روش‌های مختلفی ارائه شده است، یکی از این روش‌ها جستجوی انتخاب شده [۲] است. طراحی این الگوریتم به نحویست که شامل این ملاحظات می شود:

۱. اشیاء ممکن‌است در هر مقیاسی وجود داشته باشند، بنابراین باید الگوریتم طوری طراحی شود که ناحیه مربوط به هر شی با هر مقیاسی را تشخیص بدهد.

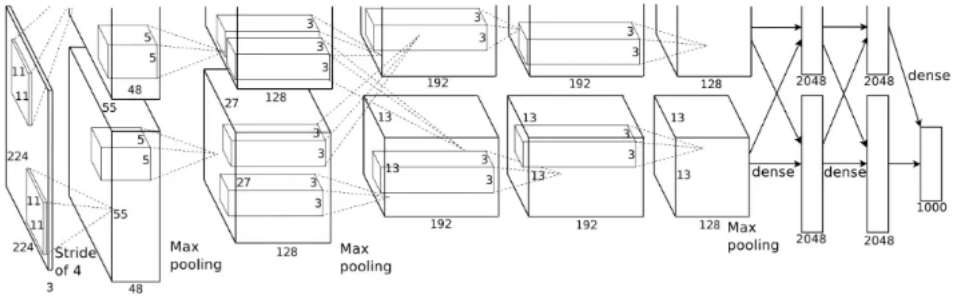
۲. تمایز نواحی میتواند به دلایل مختلفی مثل تفاوت رنگ یا زمینه یا فاصله مکانی یا موارد دیگر باشد. بنابراین به جای اینکه فقط یکی از اینها در نظر گرفته شود، به نحوی از تمام این ها استفاده می شود.

۳. با توجه به کاربردهای این روش در کشف و تشخیص اشیاء نباید هزینه محاسباتی زیادی داشته باشد و نسبتا باید سریع باشد.

جستجوی انتخاب کننده یک روش حریصانه است و به این صورت عمل می کند که ابتدا تعدادی ناحیه اولیه را مشخص می کند و سپس بر اساس میزان شباهتی که هر ناحیه با نواحی مجاورش دارد، هر ناحیه را با شبیه ترین ناحیه مجاور آن ادغام می کند و این کار را تا آنجا که کل تصویر در یک ناحیه ادغام شود ادامه می دهد [۲].
برای اندازه گیری میزان شباهت از ترکیب خطی شباهت‌های اندازه‌گیری شده استفاده می‌شود و سپس

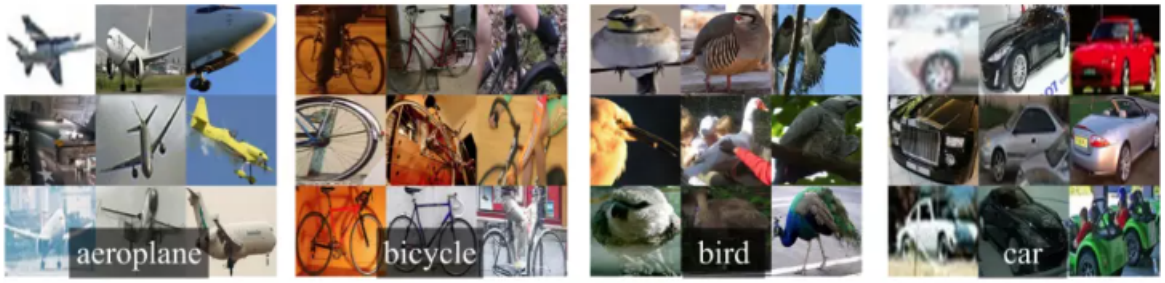
[۱] Support Vector Machines (SVM)
[۲] Selective Search

یک معیار شباهت کلی بین نواحی بدست می‌آید. این معیارها مربوط به ترکیب رنگ، زمینه، اندازه و میزان فیت شدن دو ناحیه به همدیگر است.



شکل ۲-۲: شبکه ImageNet برای استخراج ویژگی های نواحی پیشنهادی [۲]

بعد از تعیین نواحی پیشنهادی که حدود ۲۰۰۰ ناحیه است، هر ناحیه برای استخراج ویژگی‌ها به شبکه ImageNet داده می‌شود. ساختار این شبکه در شکل بالا نشان داده شده است، ولی لایه آخر شبکه کار کلاس بندی را انجام می‌دهد در اینجا حذف می شود و بردار ۴۰۹۶تایی از ویژگی ها بدست می آید.
با توجه به اینکه ابعاد تصویر ورودی به این شبکه باید مقدار ثابتی داشته باشد، در این مقاله از کشیدن تصویر و تغییر ابعاد برای اندازه کردن آن استفاده شده است که نمونه هایی از آن در شکل زیر قابل مشاهده است. ضمن اینکه می توان از اضافه کردن حاشیه نیز به جای کشیدن تصویر استفاده کرد.



شکل ۲-۳: نمونه هایی از تصاویر کشیده شده برای تطبیق به ورودی شبکه ImageNet

نهایتا نیز از این بردار ویژگی‌های با طول ثابت به عنوان ورودی به تعدادی ماشین بردار پشتیبان که برای دسته بندی هر نوع از اشیاء بهینه شده اند داده می شوند. البته به جای این کار می شود از یک لایه سافت مکس استفاده کرد که در مقاله نشان داده شده است که روش بردار پشتیبان عملکرد مناسب تری دارد [۱].

معایب شبکه‌های عصبی کانولوشنی مبتنی بر منطقه که در این مقاله ارائه شده عبارتند از:

۱. چند مرحله ای بودن فرآیند آموزش مدل.

۲. هزینه بالای آموزش مدل.

۳. کند بودن عملکرد مدل.

فصل سوم

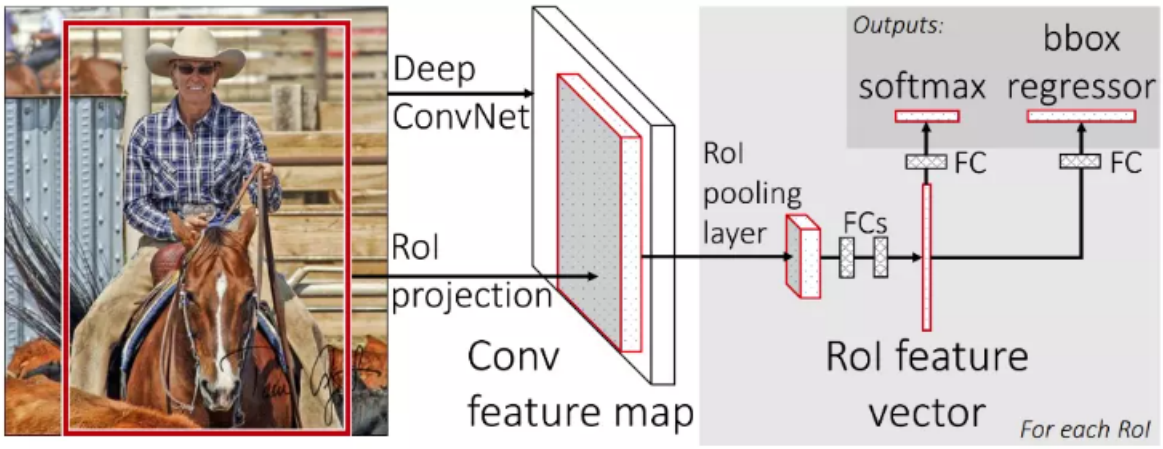
توسعه های شبکه های عصبی کانولوشنی مبتنی بر منطقه

۳-۱ مقدمه

در این فصل رابطه چند نسخه مختلف شبکه عصبی کانولوشنی مبتنی بر منطقه که در واقع هر یک توسعه‌هایی را بر روی آن ارائه کرده‌اند صحبت می‌کنیم. در مورد هر یک از این توسعه‌ها تغییرات و اصلاحاتی که بر روی مدل ارائه کرده‌اند توضیح می‌دهیم و آزمایش‌ها و بهبودهایی که در نتایج حاصل شده است را دنبال می‌کنیم.

۳-۲ مدل شبکه عصبی کانولوشنی مبتنی بر منطقه سریع

این روش نیز توسط راس گیرشیک یک سال بعد از روش شبکه های عصبی کانولوشنی مبتنی بر منطقه مطرح شد [۳]. ایده کلی این روش در شکل زیر آمده است:



شکل ۳-۱: فرآیند کلی روش شبکه عصبی کانولوشنی مبتنی بر منطقه سریع [۳]

ورودی این معماری تصویر کامل و مجموعه ای از نواحی پیشنهادی [۱] است. ابتدا کل تصویر توسط یک شبکه عصبی کانولوشنی از پیش آموزش دیده مثل همان ImageNet که در مدل قبلی استفاده شده داده می شود تا یک نقشه ویژگی از کل تصویر بدست بیاید. در واقع مدل ImagNet تا قبل از لایه های کاملا متصل [۲] آن مورد استفاده قرار می گیرد و آخرین لایه حداکثر ادغام [۳] با لایه ادغام نواحی پیشنهادی به ازاء هر ناحیه پیشنهادی جایگزین می شود تا یک بردار ویژگیهای با طول ثابت بدست آید.

[۱] Region of Interest (RoI)
[۲] fully connected
[۳] max pooling

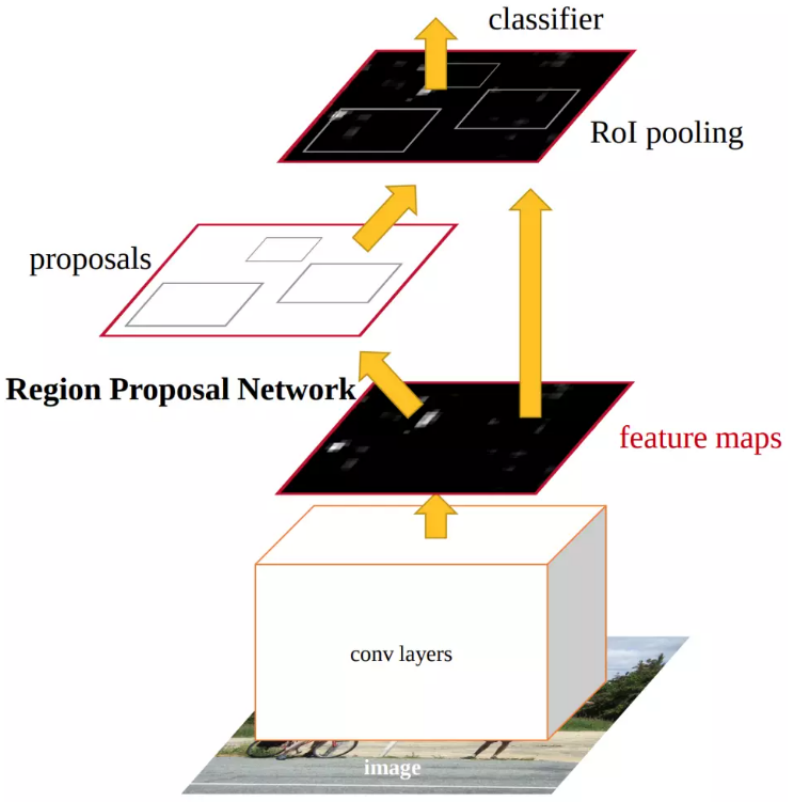
لایه ادغام نواحی پیشنهادی از ادغام حداکثر برای تبدیل ویژگی های داخل هر ناحیه به یک شکل دو بعدی با ابعاد ثابت استفاده می کند. این لایه هر ناحیه پیشنهادی را به تعدادی پنجره با ابعاد ثابت تقسیم می کند و در هر پنجره ماکسیمم مقدار ویژگیها را انتخاب می کند.

اگر ابعاد ناحیه پیشنهادی h*w باشد و ابعاد نهایی مورد نیاز H*W باشد ابعاد تقریبی پنجره ها h/H*w/W خواهد بود.

بعد از این لایه هم دو لایه کاملا متصل در کنار هم یکی برای دسته بندی اشیاء به همراه softmax و دیگری برای محل قرار گیری شیء به عنوان رگرسیور جعبه مرزی [۴] استفاده می شود.

۳-۳ مدل شبکه عصبی کانولوشن مبتنی بر منطقه سریع‌تر

این روش توسط شائوکینگ رن در سال ۲۰۱۶ مطرح شد [۴]. معماری کلی این مدل به صورت شکل زیر است.



شکل ۳-۲: مدل شبکه عصبی کانولوشن مبتنی بر منطقه سریعتر [۴]

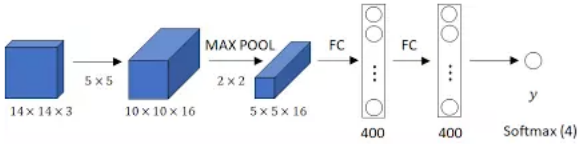
[۴] boundary box regressor

این مدل از دو ماژول تشکیل شده است. ماژول اول یک شبکه عصبی عمیق کانولوشنی که نواحی پیشنهادی را مشخص [۵] می کند و ماژول دوم همان شعک مبتنی بر منطقه سریع است که از نواحی پیشنهادی بدست آمده استفاده می کند.

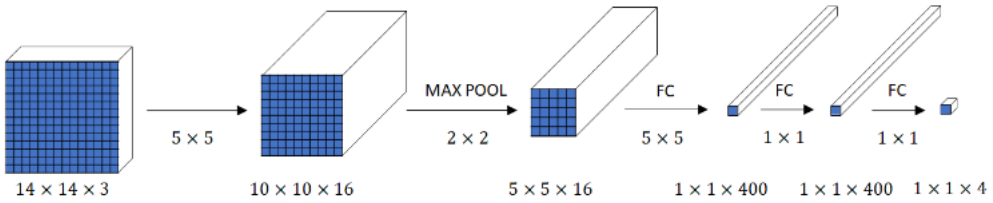
در اینجا دو ایده مورد استفاده قرار گرفته است. ایده اول برای حرکت دادن پنجره روی کل تصویر برای پیدا کردن نواحی پیشنهادی و ایده دوم برای پیدا کردن چند شی که دارای مرکز یکسانی در یک خانه هستند. که در ادامه به توضیح این دو می‌پردازیم:

- **پیاده سازی لایه کاملا متصل با استفاده از کانولوشن**

در صورتی که بخواهیم یک تنسور n*n*c را که از کانولوشن لایه های قبلی بدست آمده به یک لایه کاملا متصل بدهیم که دارای k پرسپترون است می توانیم از k فیلتر n*n استفاده کنیم. که در نتیجه به همان ابعاد میرسیم. در شکل زیر نمونه این کار قابل مشاهده است. البته در شکل اول آخرین لایه باید یک soft max با ۴ خروجی باشد.



شبکه پیاده سازی شده با fully connected



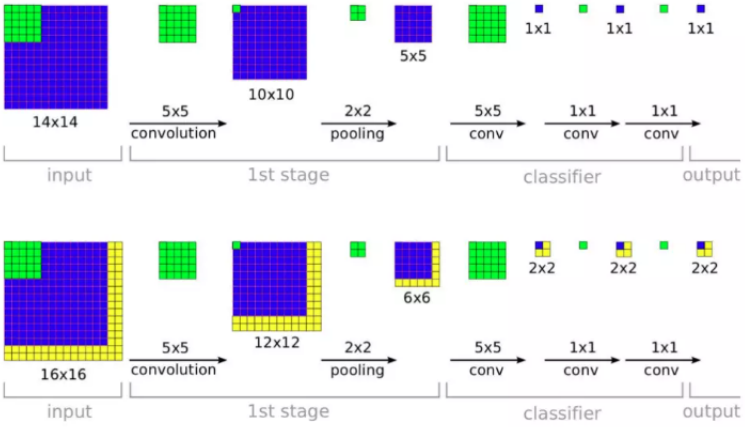
شکل ۳-۳: شبکه پیاده سازی شده با لایه کانولوشن [۵]

بنابراین به این طریق می توانیم کل شبکه را به صورت کانولوشنی پیاده سازی کنیم و در نتیجه بسیاری از محاسبات را به اشتراک گذاشته و عملکرد شبکه را برای اجرا رو GPU بهینه کنیم.

اما ایده مقاله به این صورت است که اگر ما یک شبکه کانولوشنی داشته باشیم که بطور مثال روی ابعاد ۱۴*۱۴ آموزش دیده است و خروجی ۱*۱ به ما میدهد را روی یک تصویر ۱۶*۱۶ اعمال کنیم خروجی آن به صورت ۲*۲ خواهد بود و این مثل اینست که ما یک پنجره ۱۴*۱۴ را روی

[۵] Region Proposal Network (RPN)

تصویر ۱۶*۱۶ لغزانده باشیم و به ازاء هر پنجره خروجی را بدست آورده باشیم. بنابراین با این شیوه به جای لغزاندن پنجره روی تصویر با یکبار گذراندن تصویر از شبکه می توانیم خروجی را به ازاء پنجره های مختلف تصویر بدست بیاوریم.



شکل ۳-۴: اعمال شبکه کانولوشنی آموزش دیده روی تصاویر کوچکتر روی تصاویر بزرگتر

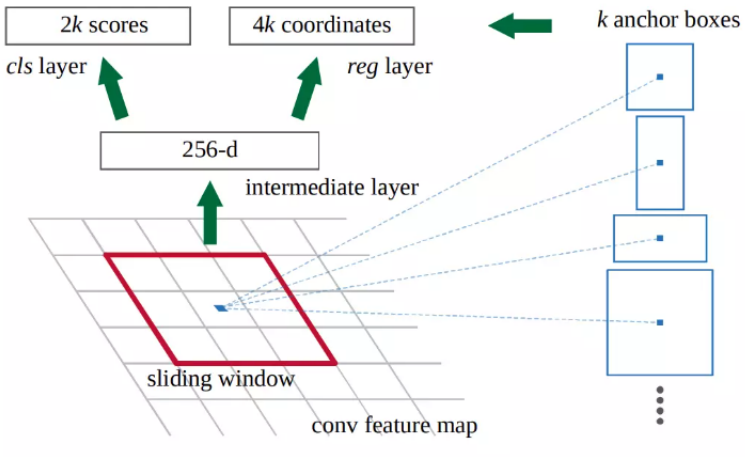
- **استفاده از جعبه لنگر** [۶]

برای ایده دوم ابتدا بهتر است، در مورد برچسب داده های آموزش یک مقدار توضیح دهیم. برای مثال ما یک تصویر داریم که آن را به تعداد خانه کوچکتر مثلا ۱۹*۱۹ خانه تقسیم کرده ایم. در خروجی مدل به ازاء هر خانه کوچک یک بردار شامل احتمال وجود شی در آن خانه و مختصات نقطه گوشه سمت چپ بالای تصویر و طول و عرض آن که مجموعا ۵ عدد می شود و به اندازه k+1 درایه دیگر برای k کلاس از شیاء و یکی هم برای زمینه که به صورت یک داغ نشان می دهد شی مورد نظر مربوط به کدام دسته است. پس در مجموع برای هر کدام از ۱۹*۱۹ خانه یک بردار 5+(k+1) بعدی خواهیم داشت. اگر یک شی در چند خانه قرار بگیرد فقط درایه اول در خانه ای ۱ خواهد بود که مرکز تصویر در آن قرار دارد.

مشکلی می تواند پیش بیاید اینست که مرکز دو شی در یکی از این خانه ها واقع شود. برای حل این مشکل ایده ارائه شده به این صورت است که می توانیم از تعداد جعبه لنگر استفاده کنیم که هر کدام از آنها مرکزشان در مرکز خانه قرار می گیرد و یک مقیاس و نسبت ابعاد متفاوتی دارند

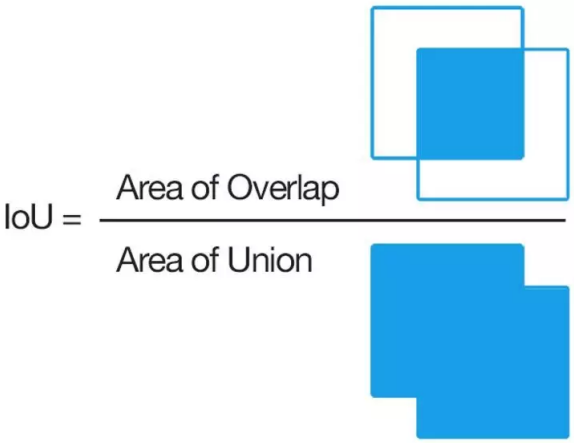
[۶] Anchor box

و به ازاء هرکدام از این جعبه لنگر ها در بردار بالا شبیه آن را به ادامه بردار اضافه کنیم. در مقاله شعک منطقه سریعتر از ۹ جعبه لنگر متفاوت استفاده شده است.



شکل ۳-۵: نمونه ای از جعبه لنگرها

برای محاسبه اناش که مخفف عبارت انطباق بر روی اشتراک [۷] است، از رابطه زیر استفاده می شود:



شکل ۳-۶: فرمول محاسبه اناش

برای آموزش شبکه منطق پیشنهادی به ازاء هر جعبه لنگر در هر خانه از تصاویر آموزشی عدد صفر یا یک اختصاص می دهیم. عدد ۱ را در حالت به یک جعبه لنگر اختصاص می دهیم:

[۷] Intersection over Union (IoU)

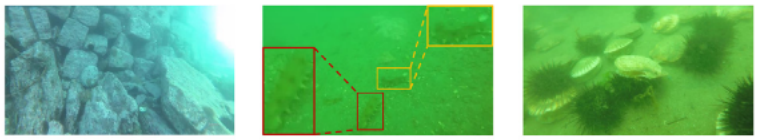
۱. در صورتی که جعبه لنگر مورد نظر بیشترین اناش را با محدوده واقعی آن شی داشته باشد.

۲. درصورتی که جعبه لنگر مورد نظر دارای اناش بالاتر از ۰.۷ با محدوده واقعی شی داشته باشد.

چون از این شبکه برای بخش شعک منطقه سریع هم می خواهیم استفاده کنیم دو روش برای اشتراک گذاری وجود دارد. روش اول اینست که اول برای در انتخاب نواحی پیشنهادی آموزش داده شود و سپس برای شعک منطقه سریع و این کار تکرار شود که در مقاله مذکرو همین روش مورد استفاده قرار گرفته است. روش دیگر هم اینست که همزمان با هم با خطای هر دو ماژول این ضرایب اصلاح شوند.

۳-۴ مدل تقویت‌شده شبکه عصبی کانولوشن مبتنی بر منطقه

محیط‌های پیچیده زیر آب چالش‌های جدیدی مانند شرایط نوری نامتعادل، کنتراست کم، انسداد و تقلید موجودات آبزی را برای تشخیص اشیا به ارمغان می‌آورند. در این شرایط، اشیاء ثبت شده توسط دوربین زیر آب مبهم می شوند و آشکارسازهای عمومی اغلب روی این اجسام مبهم از کار می افتند. در شکل زیر نمونه‌ای از تصاویز ثبت شده از زیر آب نشان‌داده شده است، همانطور که مشاهده می‌کنید تشخیص اشیا دشوار است.



شکل ۳-۷: تصاویر ثبت شده از زیر آب

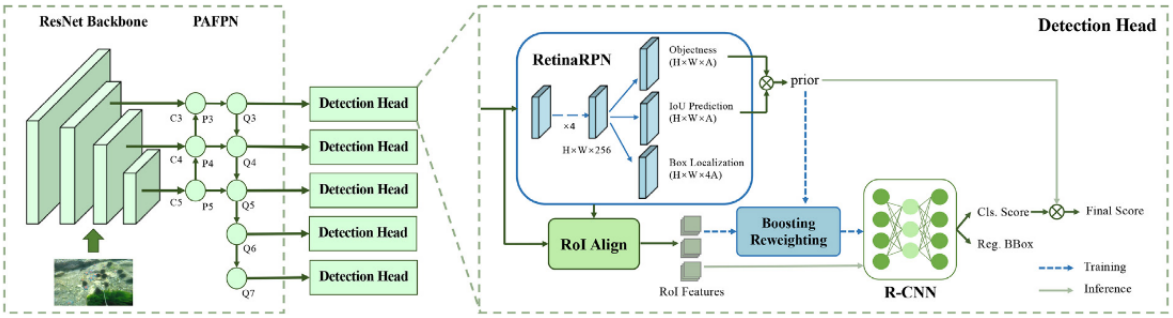
آشکارسازی دو مرحله‌ای شعک منطقه تقویت‌شده شامل سه جز کلیدی زیر می‌باشد [۶]:

۱. یک شبکه پیشنهادی منطقه جدید به نام RetinaRPN پیشنهاد می‌شود که پیشنهادهای با کیفیت بالا ارائه می‌کند و شیئ بودن و پیش‌بینی اناش را برای عدم قطعیت برای مدل‌سازی احتمال قبلی شی در نظر می‌گیرد.

۲. خط لوله استنتاج احتمالی [۸] برای ترکیب عدم قطعیت قبلی مرحله اول و امتیاز طبقه بندی مرحله دوم برای مدل سازی امتیاز تشخیص نهایی

[۸] Probabilistic Inference Pipeline

۳. استخراج نمونه سخت جدید (افزایش وزن مجدد)



شکل ۳-۸: شبکه شعک منطقه تقویت‌شده [۶]

ساختار شبکه شعک منطقه تقویت‌شده در شکل بالا نشان داده شده است. که به ترتیب شامل مراحل زیر می‌باشد:

۳-۴-۱ استخراج ویژگی از تصاویر

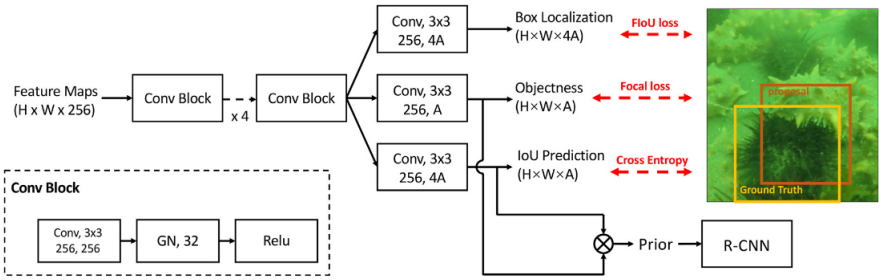
ویژگی‌های استخراج شده از تصاویر، می‌توانند سطح پایین یا سطح بالا باشند.استخراج ویژگی‌ها شامل دو بخش است: ۱. ویژگی‌های پایین به بالا ۲. ویژگی‌های بالا به پایین

۳-۴-۲ RetinaRPN

شبکه پیشنهادی منطقه مسئول ارائه پیشنهادهایی است که دارای اشیاء بالقوه هستند. تصاویر زیر آب تار، کم کنتراست و اعوجاج هستند که تشخیص اشیا از پس زمینه را دشوار می کند. در نتیجه، پیشنهادات با کیفیت بالا ممکن است توسط پیشنهادات با پسرفت ضعیف با شیئ بالاتر فیلتر شوند.

برای به دست آوردن پیشنهادات با کیفیت بالا با احتمالات قبلی دقیق، هدف ما ایجاد یک تشخیص اشیا قوی با الهام از طرح‌های آشکارساز یک مرحله‌ای فعلی است که شبکه پیشنهادی منطقه شبکیه نامیده می‌شود.

در شکل زیر ساختار RetinaRPN نشان داده شده است. RetinaRPN ویژگی ها را با چهار بلوک تبدیل استخراج می کند(هر بلوک تبدیل شامل ۳*۳ کانولوشن، GN با ۳۲ گروه و Relu است). محل جعبه توسط FIoU loss نظارت می شود. شیئ بودن با استفاده از focal loss و پیش بینی اناش توسط cross-entropy loss کنترل می شود [۶].



شکل ۳-۹: ساختار RetinaRPN [۶]

۳-۴-۳ خط لوله استنتاج احتمالی

برای آشکارساز دو مرحله ای، در مرحله اول، خروجی، kجعبه پیشنهادی را ارائه می‌کند. RPN یک احتمال را پیش‌بینی ارائه می‌کند، که در آن زمانی‌که احتمال برابر با ۱ باشد، نشان‌دهنده یک شی است و زمانی‌که احتمال برابر با فر باشد، پس‌زمینه را نشان می‌دهد. این نتیجه توسط یک طبقه‌بندی‌کننده باینری که با یک هدف log-lihood آموزش داده شده است، تحقق می‌یابد. شعک منطقه یاد می گیرد که هر پیشنهاد را در یکی از کلاس های پیش زمینه یا پس زمینه طبقه بندی کند.

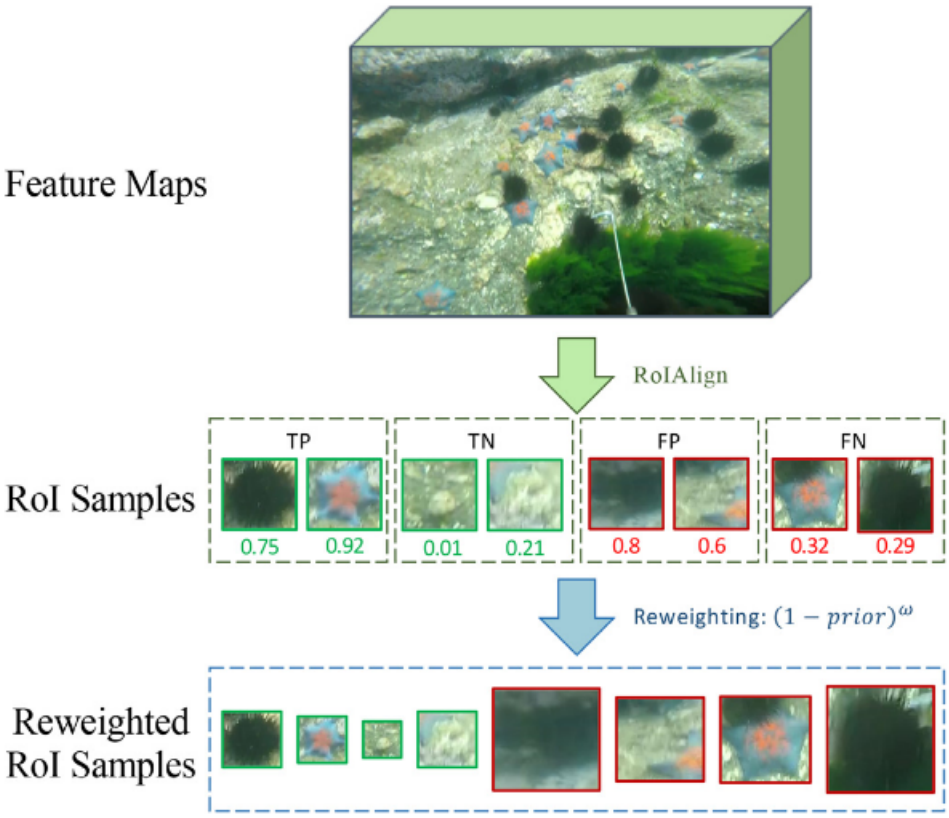
۳-۴-۴ وزن‌دهی مجدد

در آشکارساز دو مرحله ای، مرحله دوم پیش بینی هایی را انجام می دهد که مستقل از مرحله اول است. در نتیجه، امتیاز پایین برای یک نمونه با کیفیت بالا در مرحله اول، بر نتیجه تشخیص نهایی تأثیری نخواهد داشت. با این حال، در خط لوله احتمالی دو مرحله‌ای، زمانی که ماژول تشخیص اشیا به اشتباه یک احتمال کم برای یک پیشنهاد مثبت با کیفیت بالا ایجاد می‌کند، به سختی می‌توان آن را به‌عنوان یک پیش‌بینی با اطمینان بالا در نظر گرفت.

بنابراین، ما یک استراتژی نمونه برداری نرم به نام افزایش وزن مجدد را پیشنهاد می کنیم، برای نمونه k، وزن طبقه بندی آن عبارت است از:

$$\begin{cases} W_k = (1 - p_k)^w, & k \in \mathcal{F}, \\ W_k = p_k^w, & k \in \mathcal{B}, \end{cases}$$

زمانی‌که w از صفر بزرگتر است پارامتر تقویت کننده است.F مجموعه نمونه های اشیا و B نشان دهنده مجموعه نمونه های پس زمینه است.



شکل ۳-۱۰: نمونه‌ای از وزن‌دهی مجدد برای تقویت پیشنهاد [۶]

۳-۵ توجه چند مقیاسی در شبکه عصبی کانولوشنی مبتنی بر منطقه

درک صحنه سنجش از دور برای استخراج اطلاعات ارزشمند از تصاویر با وضوح بالا، از جمله تشخیص و طبقه بندی اشیا، بسیار مهم است. روش‌های سنتی تشخیص اشیاء در مدیریت مقیاس‌ها، جهت‌گیری‌ها و پس‌زمینه‌های پیچیده موجود در داده‌های سنجش از دور با چالش‌هایی مواجه هستند.

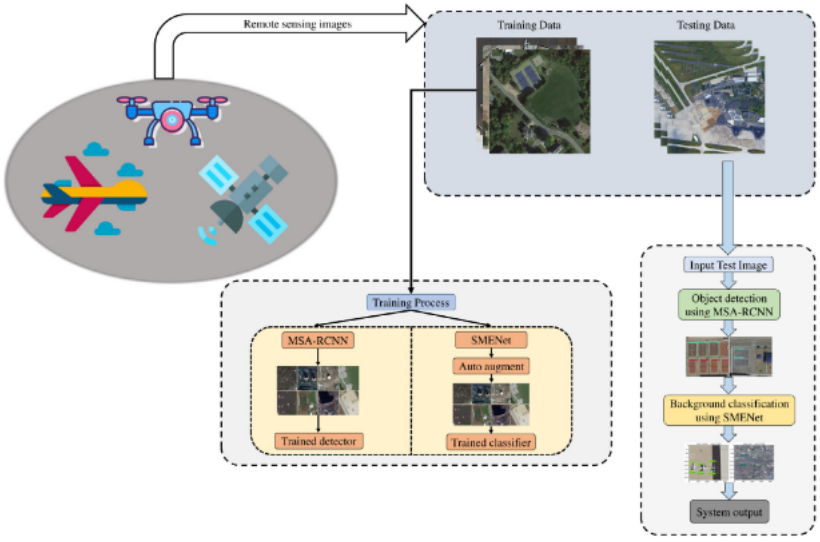
تصویری از سناریوی کاربردی درک صحنه سنجش از دور با استفاده از سیستم تشخیص شی مبتنی بر شبکه عصبی کانولوشن مبتنی بر منطق چندمقیاسی [۹] در زیر نشان داده شده است:

[۹] Multiscale Attention R-CNN (MSA R-CNN)

شکل ۳-۱۱: کاربرد درک صحنه در سنجش از دور

تصاویر سنجش از دور حاوی اطلاعات مفید زیادی از جمله ویژگی‌های جسم و پس‌زمینه هستند. علاوه بر این، استفاده از تشخیص اشیاء در تصاویر سنجش از راه دور به دلیل پس‌زمینه‌های پیچیده، واریانس وضوح و جهت‌گیری متفاوت که بر عملکرد سیستم درک صحنه سنجش از دور تأثیر می‌گذارد، چالش‌های متعددی را ایجاد می‌کند.

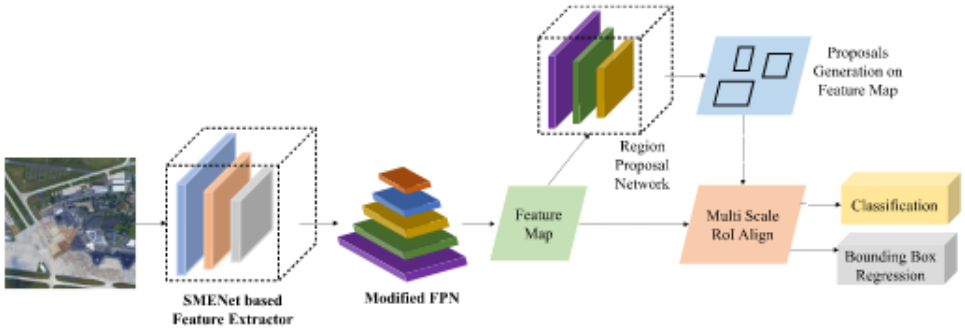
شکل زیر معماری کلی سیستم درک صحنه سنجش از دور پیشنهادی را نشان می‌دهد که می‌تواند به طور موثر اجسام را شناسایی کرده و اطلاعات پس‌زمینه یک تصویر معین را استخراج کند.



شکل ۳-۱۲: ساختار کلی سیستم تشخیص شی برای درک تصاویر صحنه سنجش از دور [۷]

روش‌های تشخیص شی دو مرحله‌ای مانند شعک منطقه سریع‌تر به دلیل عملکرد بالاتر با استنتاج

سریع پیش‌بینی، مدل‌های بسیار محبوبی هستند. شعک منطقه چندمقیاسی از ساختار مدل شعک منطقه سریعتر پیروی می‌کند، مدل ابتدا ویژگی ها را از تصویر با استفاده از شعک استخراج می‌کند و سپس ویژگی ها در تشخیص شی پردازش می‌شوند تا شی و پس زمینه را به همراه آنها طبقه بندی کند. ساختار شبکه شعک منطقه چندمقیاسی را می توان در شکل زیر مشاهده کرد.



شکل ۳-۱۳: مدل تشخیص شی شعک منطقه چندمقیاسی [۷]

شعک منطقه چندمقیاسی برای تشخیص شیء و درک صحنه در تصاویر سنجش از دور است و از سه بخش اصلی تشکیل شده است:

- **SMENet:** یک شبکه ویژگی‌گیری چند مقیاس که از دو شبکه VGG16 تشکیل شده است. یکی از شبکه‌ها برای استخراج ویژگی‌های جزئی و دیگری برای استخراج ویژگی‌های کلی استفاده می‌شود.
- **ADIL:** یک اتصال جانبی داخلی پویا که برای بهبود انتقال اطلاعات بین طبقات مختلف ویژگی در شبکه ویژگی‌های هرمی [۱۰] استفاده می‌شود.
- **DLAM:** یک ماژول توجه توزیع‌شده سبک‌وزن که برای بهبود پردازش اطلاعات ویژگی در شبکه FPN استفاده می‌شود.

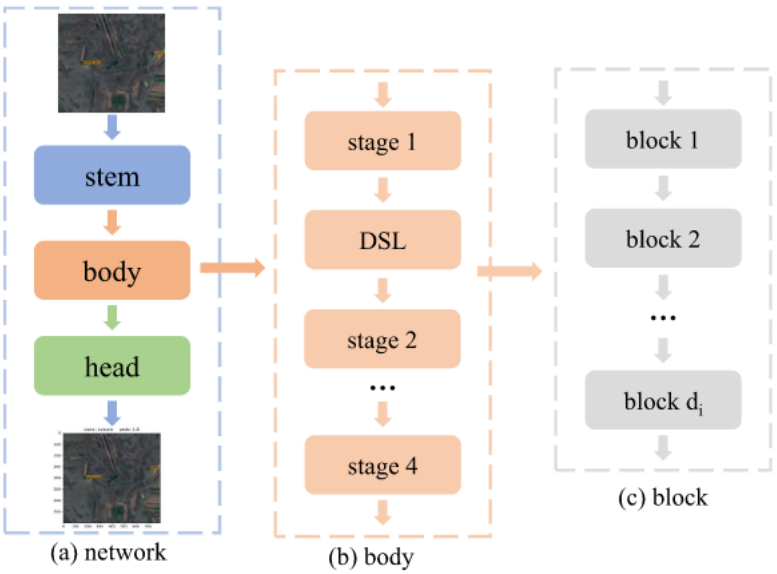
در ادامه، به بررسی هر یک از بخش‌های اصلی شعک منطقه چندمقیاسی می‌پردازیم.

[۱۰] Feature Pyramid Networks (FPN)

۳-۵-۱ مؤلفه استخراج ویژگی

شناخت مؤثر اشیاء متراکم و کم رنگ با استفاده از مدل‌های تشخیص شیء اغلب به دلیل بازنمایی ویژگی های پیچیده ، می تواند چالش برانگیز باشد ، که شامل انواع رنگ ها و اشکال در زمینه های متنوع است. چارچوب استخراج ویژگی معمولی به طور معمول پیش بینی ها را بر اساس لایه نهایی خصوصیات استخراج شده انجام می دهد. با این حال ، توانایی شبکه در تشخیص اشیاء با تعداد پیکسل های موجود در تصویر منبع محدود است که یک نقشه ویژگی واحد می تواند نشان دهد. برای حفظ یک میدان پذیرش کافی ، از نمونه‌گیری پایین [۱۱] استفاده می شود [۷].

شکل زیر معماری مدل SMENet مورد استفاده در این مطالعه را برای استخراج ویژگی ها و طبقه بندی پس زمینه در تصاویر سنجش از راه دور شناسایی شده نشان می دهد. مدل پیشنهادی شامل سه مؤلفه اصلی است: ساقه ، بدن و سر. هر یک از این مؤلفه ها نقش مهمی در پردازش داده های ورودی و تولید خروجی معنی دار برای مراحل بعدی دارند.



شکل ۳-۱۴: ساختار کلی مدل SMENet برای استخراج ویژگی از تصاویر [۷]

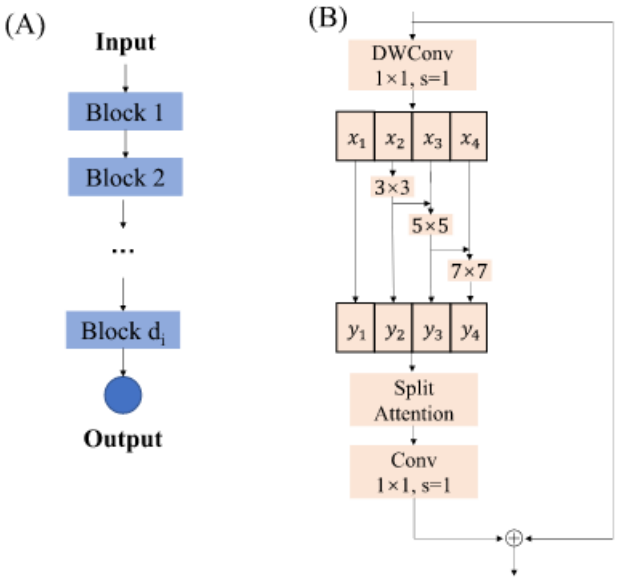
بدنه مدل SMENet شامل چهار مرحله است که هر مرحله شامل تعداد متغیر بلوک ها است. این بلوک ها به عنوان واحدهای اساسی ساختمان مدل عمل می کنند و استخراج ویژگی ها را از داده های ورودی تسهیل می کنند. هر بلوک از دو عنصر اصلی تشکیل شده است: یک ماژول کانولوشن و یک ماژول توجه.

[۱۱] Downsampling

ماژول کانولوشن وظیفه اعمال عملیات پیچیدن را در داده های ورودی دارد و از این طریق استخراج ویژگی های مکانی را قادر می سازد. در مقابل ، ماژول توجه با تأکید بر اطلاعات مهم و سرکوب جزئیات بی ربط ، بر پالایش ویژگی های استخراج شده متمرکز است.

مدل SMENet می تواند به طور مؤثر بازنمایی های ویژگی پیچیده را بیاموزد و به طور مؤثر زمینه ها را در تصاویر سنجش از راه دور طبقه بندی کند. سازگاری مدل پیشنهادی، با تعداد متفاوتی از بلوک ها و مراحل، اجازه می دهد تا متناسب با الزامات خاص کاربردی تنظیم شود، در نتیجه عملکرد بهینه را در کارهای مختلف سنجش از راه دور تضمین می کند.

شکل زیر معماری اجزای ساقه و بدن SMENet را نشان می دهد. SMENet از چندین عنصر بلوک تشکیل شده است، این روش پردازش چند مقیاس امکان استخراج ویژگی ها را در مقیاس های مختلف فراهم می کند. این امر به ویژه در برنامه های سنجش از دور مفید است [۷].

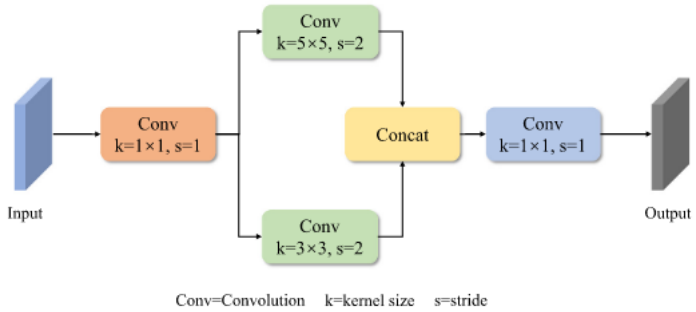


شکل ۳-۱۵: ساختار مرحله فردی و بخش بلوک SMENet [۷]

۳-۵-۲ لایه نمونه‌گیری پایین

شکل زیر روش نمونه‌گیری پایین به کار رفته توسط SMENet بین هر مرحله برای حفظ اطلاعات ویژگی غنی ضمن کاهش ابعاد نقشه ویژگی ورودی را نشان می دهد. این رویکرد برای بهبود عملکرد کلی مدل با حفظ اطلاعات مهم فضایی و به حداقل رساندن از دست دادن جزئیات خوب که اغلب در طی تکنیک

های پایین آمدن معمولی رخ می دهد ، طراحی شده است [۷].



شکل ۳-۱۶: ساختار لایه نمونه‌گیری پایین [۷]

۳-۵-۳ شبکه پیشنهادی منطقه

ویژگی‌های استخراج‌شده از تصاویر ورودی توسط استخراج‌کننده ویژگی، از طریق شبکه پیشنهادی منطقه منتقل می‌شوند، که از آموزش انتها به انتها [۱۲] استفاده می‌کند تا اشیاء درون تصاویر را به همراه جعبه‌های مرزی مربوطه آنها پیشنهاد کند. این فرآیند به طور موثر پیشنهادات منطقه ای را ایجاد می کند که به طور بالقوه حاوی اشیاء مورد علاقه هستند، و امکان طبقه بندی و کارهای رگرسیون بعدی را فراهم می کند.

۳-۵-۴ انتخاب منطقه موردنظر به صورت چند مقیاسی

تکنیک تراز چند مقیاس منطقه مورد علاقه، نقشه های ویژگی های نمونه‌گیری از بلا [۱۳] را انجام می دهد و منطقه مورد علاقه با اندازه های مختلف را برای هر شیء نامزد تولید می کند. این روش امکان استخراج مؤثر از ویژگی ها در مقیاس های مختلف را فراهم می کند و نمایش جامع تری از اشیاء موجود در تصاویر را فراهم می کند.

۳-۵-۵ لایه پیش بینی نهایی

لایه پیش بینی نهایی نقش مهمی در طبقه بندی اشیا و در تولید پیشنهادات رگرسیون نهایی در جعبه مرزی با استفاده از یک لایه کاملاً متصل دارد. این لایه برای سنتز نتایج از مراحل پردازش قبلی طراحی

[۱۲] End-to-end
[۱۳] Upsampled

شده است که منجر به تشخیص دقیق شی و محلی سازی در تصاویر می شود. به طور سنتی، از Cross-Entropy به عنوان یک تابع ضرر در لایه پیش بینی استفاده می شود.

۳-۵-۶ شبکه ویژگی‌های هرمی اصلاح شده

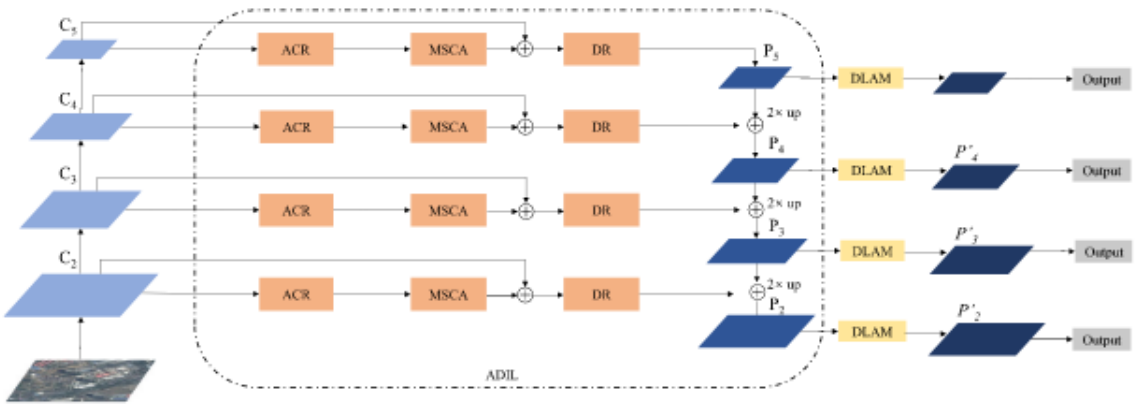
شبکه ویژگی‌های هرمی نقش مهمی در استخراج ویژگی ها از تصاویر ورودی ایفا می کند و به طور گسترده در روش های تشخیص اشیا استفاده می شود. شویه برای افزایش دقت و سرعت آشکارسازهای شی معرفی شد و راه حل کارآمدتری برای وظایف تشخیص اشیا ارائه کرد. در سال های اخیر، شویه در معماری های شعک ادغام شده است تا ویژگی های غنی را از تصاویر ورودی استخراج کند.

با این حال، آشکارسازهای شی مبتنی بر شویه موجود اغلب در صحنه‌های سنجش از دور به دلیل پیچیدگی ذاتی اشیاء با پس‌زمینه‌های مختلف، ویژگی‌های چند مقیاسی و از دست دادن اطلاعات در اتصال جانبی داخلی، عملکرد ضعیفی دارند.

برای پرداختن به این موضوع، این مطالعه ماژول اتصال ADIL را به معماری شویه معرفی می کند. روش ADIL به طور خاص کاهش کانال تطبیقی، تجمع بافت چند مقیاسی و مسیریابی پویا را برای کاهش از دست دادن اطلاعات در اتصال جانبی داخلی، یک مسئله رایج در روش‌های موجود، ترکیب می‌کند. این رویکرد کل نگر تضمین می‌کند که مدل ما نه تنها جزئیات پیچیده را در مقیاس‌های مختلف ثبت می‌کند، بلکه با چالش‌های منحصربه‌فرد ناشی از تصاویر سنجش از دور سازگار می‌شود و در نتیجه عملکرد تشخیص و تشخیص شی برتر را به همراه دارد [۷].

علاوه بر ماژول ADIL، مکانیسم توجه، که معمولاً در مدل‌های یادگیری عمیق برای استخراج ویژگی‌های ضروری از تصاویر استفاده می‌شود، نیز بررسی می‌شود. در نتیجه، یک ماژول جدید DLAM را پیشنهاد می‌کند [۷].

شکل زیر ساختار کلی شویه اصلاح شده را نشان می دهد. ماژول ADIL برای جایگزینی اتصال جانبی داخلی سنتی قبل از به دست آوردن ویژگی های هرمی گنجانده شده است. متعاقباً، ماژول توجه پیشنهادی به ویژگی‌های خروجی اضافه می‌شود تا نقشه‌های ویژگی پیشرفته را به دست آورد. ترکیب ماژول ADIL و ماژول DLAM به شویه اصلاح شده اجازه می دهد تا تصاویر پیچیده اشیاء را در صحنه های سنجش از راه دور بهتر ثبت کند که منجر به بهبود عملکرد تشخیص اشیا می شود.



شکل ۳-۱۷: ساختار شویه اصلاح شده [۷]

۳-۶ نقشه فعالسازی کلاس در شبکه عصبی کانولوشنی مبتنی بر منطقه

تشخیص شیء یک زمینه مهم در پردازش تصویر است که این مدل به شناسایی اشیا در تصاویر و فیلم‌ها می‌پردازد. روش‌های تشخیص شیء سنتی معمولاً از دو مرحله جداگانه برای تولید پیشنهادات و پیش‌بینی کلاس استفاده می‌کنند. در مرحله پیشنهاد، مناطق احتمالی وجود شیء در تصویر شناسایی می‌شوند. در مرحله پیش‌بینی کلاس، کلاس هر پیشنهاد تعیین می‌شود.

نقشه‌های فعال‌سازی کلاس [۱۴] نقشه‌هایی از تصویر هستند که نشان می‌دهند کدام مناطق از تصویر برای طبقه‌بندی یک شیء خاص مهم هستند. آنها معمولاً با استفاده از شبکه‌های عصبی کانولوشنی تولید می‌شوند.

برای تولید نفک که مخفف نقشه‌های فعالیت کلاس است، ابتدا یک شبکه شعک برای طبقه‌بندی شیء آموزش داده می‌شود. سپس، ویژگی‌های خروجی شبکه شعک برای یک شیء خاص گرفته می‌شوند و با استفاده از یک تابع فعال‌سازی، مانند تابع سیگموید، به یک نقشه مقیاس‌بندی می‌شوند.

مقدار هر پیکسل در نفک نشان‌دهنده این است که چقدر احتمال دارد آن پیکسل برای طبقه‌بندی شیء خاص مهم باشد. پیکسل‌هایی با مقادیر بالاتر در نفک نشان‌دهنده مناطقی از تصویر هستند که برای طبقه‌بندی شیء خاص مهم‌تر هستند.

شعک منطقه نفک یک روش تشخیص شیء انتهابه انتها است که نیازی به مراحل جداگانه برای تولید

[۱۴] Class Activation Maps (CAM)

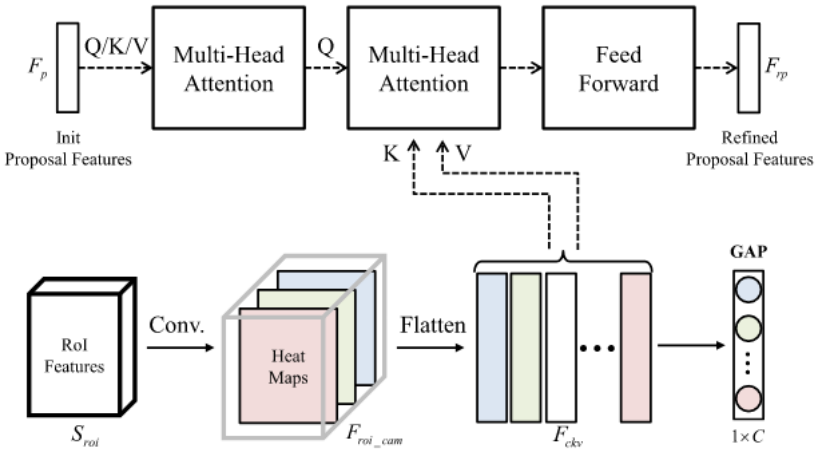
پیشنهادات و پیش‌بینی کلاس ندارد. شعک منطقه نفک از نفک برای تولید نقشه‌های توجهی استفاده می‌کند که می‌توانند به طور موثری برای شناسایی مناطقی از تصویر که یک شیء در آن قرار دارد استفاده شوند [۸].

معماری کلی شعک منطقه نفک از شبکه‌های عصبی کانولوشنی مبتنی بر منطقه تنک [۱۵] پیروی می کند. شعک منطقه تنک یک روش تشخیص شیء انتهابه انتها است که از نفک استفاده می کند. این روش بر خلاف روش های سنتی تشخیص شیء که از پیشنهادات متراکم استفاده می کنند، از پیشنهادات کم تراکم استفاده می کند. پیشنهادات کم تراکم پیشنهادات بسیار کمتری نسبت به پیشنهادات متراکم دارند. این به شعک منطقه تنک اجازه می دهد تا کارآمدتر باشد و دقت بهتری نسبت به روش های سنتی داشته باشد [۸].

شعک منطقه نفک یک روش تشخیص شیء انتهابه انتها است که از نفک برای تولید نقشه‌های توجهی استفاده می‌کند. معماری شعک منطقه نفک همانطور که در شکل زیر نشان داده شده استاز دو ماژول نفک استفاده می‌کند:

۳-۶-۱ ماژول E-CAM

این ماژول از یک رمزگذار ترانسفورمر [۱۶] برای ترکیب ویژگی‌های پیشنهاد و اطلاعات توجه در نفک ها استفاده می‌کند و این ماژول اجازه می‌دهد تا توجهی را در سطح تعبیه‌شده ایجاد کند که به تشخیص شیء کمک می‌کند.



شکل ۳-۱۸: ماژول E-CAM [۸]

[۱۵] Sparse R-CNN

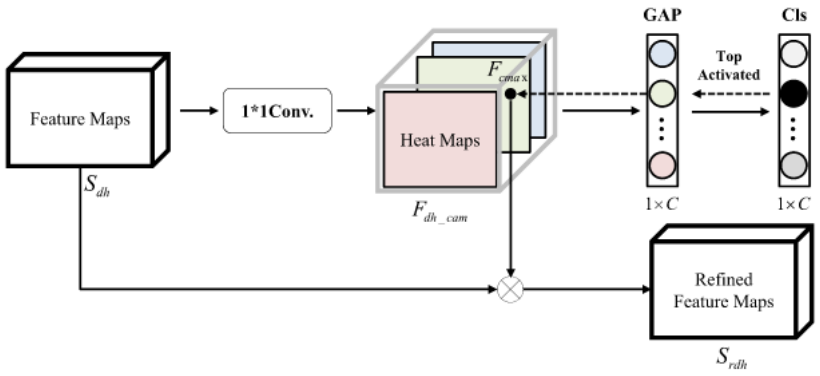
[۱۶] Transformer Encoder

ماژول E-CAM ویژگی پیشنهادی و نفک ها را به عنوان ورودی می گیرد. هدف این ماژول خروجی کردن ویژگی پیشنهادی تصفیه شده است.ماژول E-CAM از یک رمزگذار ترانسفورماتور کامل اما کم عمق برای اعمال عملیات توجه بر روی ویژگی پیشنهادی استفاده می کند. از آنجایی که رمزگذار ترانسفورماتور فقط ورودی های متوالی را می پذیرد، ویژگی پیشنهادی و نفک باید قبل از ارسال به ترانسفورماتور به بردارهای ویژگی تبدیل شوند. سپس، خروجی به یک لایه جمع‌بندی میانگین جهانی [۱۷] و سپس یک لایه softmax وارد می‌شود و جفت های کلید-مقدار را تشکیل می شود.

به طور خاص، از طریق معماری رمزگذار ترانسفورماتور کم عمق که ویژگی های پیشنهادی را به عنوان پرس و جو و ورودی های متوالی را به عنوان جفت های کلید-مقدار دریافت می کند، می توانیم ویژگی های پیشنهادی اصلاح شده را تولید کنیم [۸].

۳-۶-۲ ماژول S-CAM

این ماژول از نفک برای تولید یک نقشه توجه در سطح فضایی استفاده می‌کند. این نقشه توجه سپس با ویژگی‌های تصویر اصلی ضرب می‌شود تا یک نقشه توجه ترکیبی ایجاد شود که به تشخیص شیء کمک می‌کند.



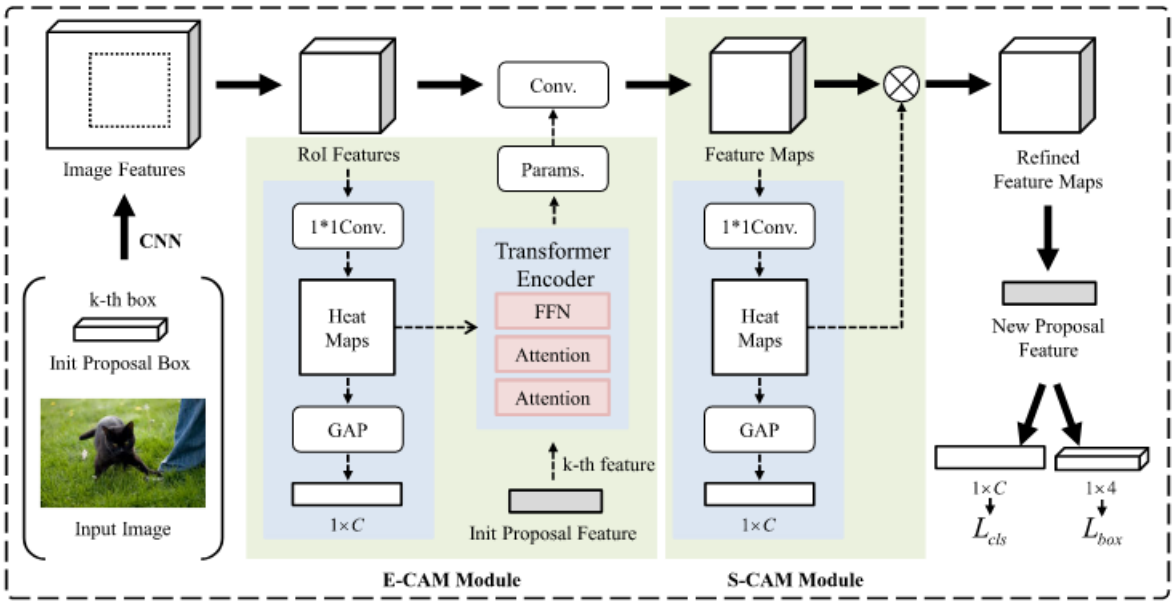
شکل ۳-۱۹: ماژول S-CAM [۸]

اطلاعات توجه فضایی می تواند در نفک پنهان شود. بنابراین، ما از ماژول S-CAM برای بهبود نقشه‌های ویژگی استفاده می‌کنیم، که با ضرب نقشه‌های ویژگی با نقشه توجه فعال شده بالا توسط نفک‌ها، توجه سطح فضایی را ایجاد می‌کند.

ماژول S-CAM در شکل نشان داده شده است. ورودی ماژول S-CAM نقشه های ویژگی است. با

[۱۷] Global Average Pooling (GAP)

توجه به نقشه‌های ویژگی یک لایه کانولوشن با اندازه ۱*۱ اضافه می‌کنیم. سپس، خروجی به یک لایه جمع‌بندی میانگین جهانی و سپس یک لایه softmax وارد می شود. نقشه فعالسازی کلاس با بیشترین احتمال را به عنوان وزن توجه انتخاب می کنیم. سپس، می‌توانیم نقشه‌های ویژگی تصفیه شده را با ضرب نقشه‌های ویژگی با وزن‌های توجه به صورت زیر بدست آوریم [۸].



شکل ۳-۲۰: مدل تشخیص شی شعک منطقه نفک [۸]

از مزایای شعک منطقه نفک می‌توان به موارد زیر اشاره کرد:

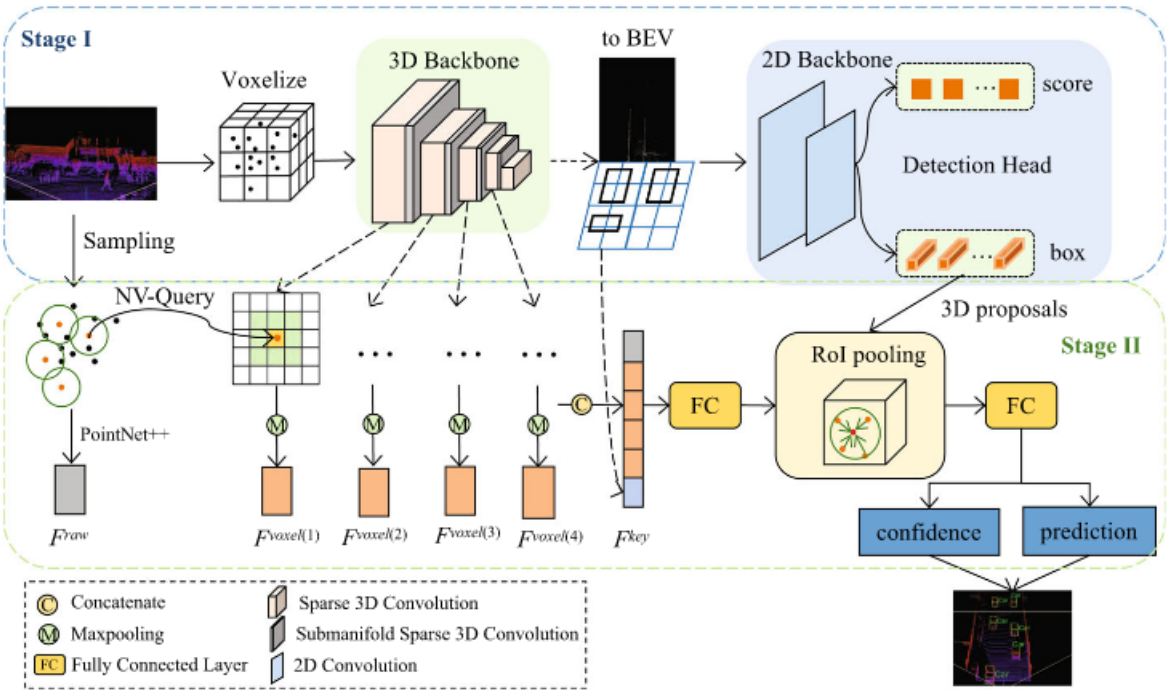
- **کارایی:** یک روش تشخیص شیء انتهابه انتها است که نیازی به مراحل جداگانه برای تولید پیشنهادات و پیش‌بینی کلاس ندارد.
- **دقیق بودن:** از نفک برای تولید نقشه‌های توجهی استفاده می‌کند که می‌توانند به طور موثری برای شناسایی مناطقی از تصویر که یک شیء در آن قرار دارد استفاده شوند.
- **قابلیت تعمیم:** در آزمایش‌ها روی داده‌های چالش‌برانگیز نشان داده است که می‌تواند به طور موثری روی داده‌های جدید تعمیم یابد.

۳-۷ تجمیع ویژگی ها بر اساس همسایگان برای تشخیص اشیاء سه بعدی

مدل NV2P-RCNN یک چارچوب دو مرحله ای برای تشخیص اشیاء سه بعدی در رانندگی خودکار است. این چارچوب بر روی ویژگی های پیکسل های همسایه در یک قاب نقطه ای متمرکز است.

- در مرحله اول، قاب نقطه ای به پیکسل های سه بعدی با اندازه ثابت تبدیل می شود. سپس، ویژگی های پیکسل با استفاده از یک شبکه عصبی سه بعدی با مکانیزم باقی مانده استخراج می شود. این شبکه عصبی ویژگی های مهمی از اشیاء سه بعدی را از پیکسل های اولیه استخراج می کند [۹].

- در مرحله دوم، ویژگی های پیکسل از پیکسل های همسایه برای هر پیکسل استخراج می شود. سپس، این ویژگی ها با ویژگی های پیکسل اصلی آن پیکسل ادغام می شوند. این ادغام ویژگی ها اطلاعات مهمی را از صحنه ارائه می دهد که می تواند به بهبود دقت تشخیص کمک کند [۹].



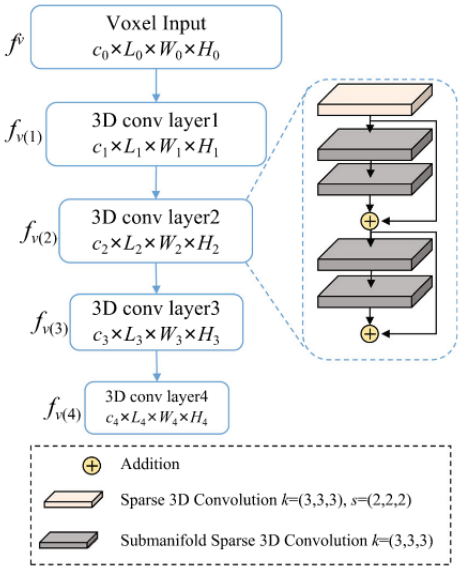
شکل ۳-۲۱: مدل تشخیص شی NV2P-RCNN [۹]

در ادامه توضیحات بیشتری درباره مراحل این شبکه و ساختار آن ارائه شده است.

۳-۷-۱ استخراج ویژگی

در این مرحله از شعک های سه بعدی برای یادگیری ویژگی‌های معنایی سطح بالا می‌دهیم و با توجه به پراکندگی حجم ویژگی‌ها، استفاده از کانولوشن پراکنده سه بعدی برای استخراج ویژگی ها می تواند در مقایسه با شعک سنتی به کارایی بالاتر و عملکرد بهتری دست یابد [۹].

همانطور که در شکل زیر نشان داده شده است، دو بلوک باقیمانده در هر لایه پیچشی پراکنده برای استخراج ویژگی ها استفاده می شود. گام پیچیدگی سه بعدی پراکنده در هر لایه (۲،۲،۲) است، به جز لایه اول که گام آن (۱،۱،۱) است. اندازه هسته تمام کانولوشن های سه بعدی (۳،۳،۳) است.



شکل ۳-۲۲: استخراج یژگی در مدل NV2P-RCNN [۹]

۳-۷-۲ تجمیع ویژگی‌ها

در این بخش ابتدا برخی از نقاط به عنوان نقاط کلیدی، نمونه برداری می شوند و سپس ویژگی‌ها در نقاط کلیدی و همسایگان را تجمیع می‌کنیم. در نهایت، از این ویژگی‌های استخراج شده برای اصلاح پیشنهادات و تولید نتایج پیش بینی پس از ادغام ناحیه مورد توجه استفاده می شود.

در مرحله استخراج نقاط کلیدی، ماژول PointNet++ برای استخراج ویژگی های ابر نقطه خام استفاده می شود. کل فرآیند از سه بخش تشکیل شده است:

۱. نمونه برداری از نکات کلیدی. از دورترین نقطه نمونه [۱۸] برای نمونه برداری از n نقطه کلیدی

[۱۸] Farthest Point Sampling (FPS)

استفاده می کنیم.

۲. گروه بندی. نقطه کلیدی را به عنوان مرکز در نظر بگیرید، یک توپ با شعاع مشخص به عنوان همسایگی آن رسم کنید، و نقاط در همان محله یک گروه هستند.

۳. رمزگذاری ویژگی. با استفاده از پرسپترون چندلایه [۱۹] ویژگی های نقاط یک گروه را استخراج می کند.

ویژگی های استخراج شده برای هر نقطه را و همسایگانش را از طریق میانگین یا بیشینه مقدار تجمیع استخراج می‌کنیم، در نهایت، ویژگی‌های همه شبکه‌ها به عنوان ویژگی کلی ناحیه مورد توجه به هم متصل می‌شوند. پس از عبور از یک لایه کاملاً متصل برای کاهش ابعاد، آنها به شاخه طبقه بندی و شاخه رگرسیون تغذیه می شوند تا دسته ناحیه مورد توجه و اطمینان را پیش بینی کنند.
ویژگی های ادغام شده سپس به یک شبکه عصبی برای طبقه بندی اشیاء سه بعدی و تخمین مختصات آنها تغذیه می شوند. این شبکه عصبی از یک مکانیزم یادگیری خودکار برای بهبود عملکرد تشخیص استفاده می کند.
این چارچوب مزایای زیر را نسبت به سایر روش های تشخیص اشیاء سه بعدی دارد:

- استفاده از ویژگی های پیکسل همسایه برای هر پیکسل، اطلاعات مهمی را از صحنه ارائه می دهد که می تواند به بهبود دقت تشخیص کمک کند.
- استفاده از یک شبکه عصبی برای ادغام ویژگی های پیکسل، یک فرآیند یادگیری خودکار را فراهم می کند که می تواند به بهبود عملکرد تشخیص کمک کند.

[۱۹] Multi Layer Perceptron

فصل چهارم

جمع‌بندی و نتیجه‌گیری

۴-۱ جمع‌بندی و نتیجه‌گیری

در این گزارش، ما به بررسی توسعه‌های الگوریتم‌های تشخیص شیء شعک منطقه نظیر، شعک منطقه سریع، شعک منطقه سریع‌تر، شعک منطقه تقویت‌شده، شعک منطقه چندمقیاسی، شعک منطقه نفک و NV2P-RCNN پرداختیم.

شعک منطقه اولین الگوریتم تشخیص شیء مبتنی بر شبکه‌های عصبی کانولوشنی بود. این الگوریتم از یک ردیاب شیء برای پیدا کردن مناطق احتمالی شیء در تصویر و سپس از یک شبکه عصبی کانولوشنی برای طبقه‌بندی هر منطقه استفاده می‌کند.

شعک منطقه سریع یک بهبود قابل توجه نسبت به شعک منطقه بود. این الگوریتم از یک ردیاب شیء یک مرحله‌ای استفاده می‌کند که زمان پردازش را به میزان قابل توجهی کاهش می‌دهد [۳].

شعک منطقه سریع‌تر یک پیشرفت بیشتر نسبت به شعک منطقه سریع بود. این الگوریتم از یک ردیاب شیء و یک شبکه عصبی کانولوشنی در یک مرحله استفاده می‌کند که زمان پردازش را حتی بیشتر کاهش می‌دهد [۴].

روش	زمان تست هر تصویر	سرعت‌بخشی
شعک منطقه	۵۰ ثانیه	x۱
شعک منطقه سریع	۲ ثانیه	x۲۵
شعک منطقه سریع‌تر	۲.۰ ثانیه	x۲۵۰

جدول ۴-۱: مقایسه مدل شعک منطقه سریع و سریع‌تر [۵]

مدل تقویت شده شعک منطقه یک الگوریتم تشخیص شیء مبتنی بر تقویت است. این الگوریتم از یک سری از الگوریتم‌های شعک منطقه برای بهبود دقت تشخیص استفاده می‌کند، دو ایده مورد استفاده قرار گرفته است. ایده اول برای حرکت دادن پنجره روی کل تصویر برای پیدا کردن نواحی پیشنهادی و ایده دوم برای پیدا کردن چند شی که دارای مرکز یکسانی در یک خانه هستند [۶].

روش	AP	AP50
شعک منطقه سریع	۴۴.۵	۸۰.۹
شعک منطقه سریع‌تر	۴۵.۹	۸۱.۶
شعک منطقه تقویت‌شده	۵۰.۷	۸۲.۷

جدول ۴-۲: مقایسه مدل شعک منطقه تقویت‌شده با شعک منطقه سریع و سریع‌تر [۶]

در جدول بالا AP مخفف میانگین صحت است و معیاری است برای ارزیابی عملکرد مدل‌های تشخیص شیء استفاده می‌شود که میانگین تعداد پیش‌بینی‌های صحیح به تعداد کل شی‌های موجود در تصویر است.AP50 به معنای میانگین صحت در فاصله ۵۰ درصد است. یعنی فاصله میان مرکز پیش‌بینی و مرکز شیء واقعی کمتر از ۵۰ درصد طول شیء واقعی است. AP50 یک معیار معمول برای ارزیابی عملکرد مدل‌های تشخیص شیء است، زیرا حساسیت نسبتاً بالایی دارد.

شعک منطقه چند مقیاسی یک الگوریتم تشخیص شیء مبتنی بر یادگیری چند وظیفه‌ای است. این الگوریتم از یک شبکه عصبی کانولوشنی برای انجام چندین کار همزمان، از جمله تشخیص شیء، ردیابی شیء و تشخیص چهره، استفاده می‌کند [۷].

شعک منطقه نفک یک الگوریتم تشخیص شیء مبتنی بر نقشه‌های توجه است. این الگوریتم از نقشه‌های توجه برای شناسایی مناطق مهم در تصویر برای تشخیص شیء استفاده می‌کند [۸].

NV2P-RCNN یک الگوریتم تشخیص شیء مبتنی بر شبکه‌های عصبی کانولوشنی و شبکه‌های عصبی پیکسل است. این الگوریتم از یک شبکه عصبی کانولوشنی برای شناسایی مناطق احتمالی شیء در تصویر و سپس از یک شبکه عصبی پیکسل برای طبقه‌بندی هر منطقه استفاده می‌کند [۹].

در مجموع، الگوریتم‌های تشخیص شیء شعک منطقه پیشرفت‌های قابل توجهی در دقت و کارایی تشخیص شیء ایجاد کرده‌اند. این الگوریتم‌ها برای طیف گسترده‌ای از کاربردها، از جمله تشخیص اجسام در تصاویر ویدئویی گرفته تا تشخیص سرطان در تصاویر پزشکی، استفاده می‌شوند.

پیشنهاداتی برای تحقیقات آینده در زمینه الگوریتم‌های تشخیص شیء شعک منطقه به شرح زیر است:

- بهبود دقت تشخیص شیء با استفاده از الگوریتم‌های یادگیری عمیق جدید
- کاهش زمان پردازش شعک منطقه با استفاده از تکنیک‌های بهینه‌سازی
- گسترش شعک منطقه به تشخیص شیء چند کلاسه
- توسعه الگوریتم‌های تشخیص شیء شعک منطقه برای کاربردهای خاص، مانند تشخیص چهره یا تشخیص اجسام در تصاویر پزشکی

با ادامه تحقیقات در این زمینه، انتظار می‌رود که الگوریتم‌های تشخیص شیء شعک منطقه به عنوان یک ابزار قدرتمند برای تشخیص شیء در کاربردهای مختلف تبدیل شوند.

منابع و مراجع

[1] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. ”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.

[2] Uijlings, Jasper RR, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. ”Selective search for object recognition.” International journal of computer vision 104 (2013): 154-171.

[3] Girshick, Ross. ”Fast r-cnn.” In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

[4] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. ”Faster r-cnn: Towards real-time object detection with region proposal networks.” Advances in neural information processing systems 28 (2015).

[5] Hmidani, O., and EM Ismaili Alaoui. ”A comprehensive survey of the R-CNN family for object detection.” In 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet), pp. 1-6. IEEE, 2022.

[6] Song, Pinhao, Pengteng Li, Linhui Dai, Tao Wang, and Zhan Chen. ”Boosting R-CNN: Reweighting R-CNN samples by RPN’s error for underwater object detection.” Neurocomputing 530 (2023): 150-164.

[7] Sagar, ASM Sharifuzzaman, Yu Chen, YaKun Xie, and Hyung Seok Kim. ”MSA R-CNN: A comprehensive approach to remote sensing object detection and scene understanding.” Expert Systems with Applications 241 (2024): 122788.

[8] Zhang, Shengchuan, Songlin Yu, Haixin Ding, Jie Hu, and Liujuan Cao. ”CAM R-CNN: End-to-End Object Detection with Class Activation Maps.” Neural Processing Letters 55, no. 8 (2023): 10483-10499.

[9] Huo, Weile, Tao Jing, and Shuang Ren. ”NV2P-RCNN: Feature Aggregation Based on Voxel Neighborhood for 3D Object Detection.” Neural Processing Letters (2023): 1-21.