مقاله پژوهشی

مجله ماشین‌بینایی و پردازش‌تصویر
سال دهم، شماره سوم، پاییز ۱۴۰۲

شناسایی و ردیابی همزمان اشیا با استفاده از یادگیری عمیق

سید حمید خاتمی[۱] و رمضان هاونگی[۲]

چکیده

شناسایی و ردیابی همزمان اشیا و وسایل نقلیه با استفاده از یادگیری عمیق، به عنوان یک موضوع پژوهشی جدید، در طی سال‌های اخیر بسیار مورد توجه قرار گرفته‌است. این موضوع، یکی از حوزه‌های فعالیت محققان در زمینه هوش مصنوعی و داده کاوی است. برای شناسایی و ردیابی همزمان اشیا و وسایل نقلیه، نیاز است تا الگوریتمی طراحی شود که بتواند از تصاویر و ویدئوها اطلاعات کافی را برای شناسایی اشیا بدست آورد و سپس عملیات ردیابی انجام شود. ردیابی اشیا به‌وسیله تشخیص اشیا، مستلزم آن است که جسم در اولین فریم و در تمام فریم‌های بعدی با موفقیت شناسایی شود. با توجه به پیچیدگی و تنوع اشیا و وسایل نقلیه، استفاده از یادگیری عمیق به عنوان روش اصلی برای شناسایی و ردیابی همزمان اشیا و وسایل نقلیه، تبدیل به یک روش بسیار موثر و قابل قبول در این زمینه شده‌است. یادگیری عمیق به کمک شبکه‌های عصبی، می‌تواند امکان شناسایی و ردیابی همزمان اشیا و وسایل نقلیه را فراهم آورد. بدین‌گونه که با مرتبط‌سازی نتایج حاصل شده از شناسایی اشیا، عملیات ردیابی توسط خط لوله پیشنهادی TPN انجام گرفت و یک شبکه ردیاب(TrackNet) ارائه شد. این شبکه ردیاب، می‌تواند یک شی متحرک و درحال حرکت را که محصور شده‌است، با استفاده از شبکه عصبی کانولوشن(CNN) بهبود یافته، شناسایی کند. دلیل این امر، تشخیص مستقیم لوله‌های باندینگ است. بعلاوه،در این مقاله، چندین شبکه ردیاب برای چالش‌های موجود در مجموعه داده‌های ویدئویی UA-DETRAC که شامل ۱۰ ساعت ویدئو برای ردیابی اشیا و وسایل نقلیه است، مورد آزمایش قرار گرفت . در نهایت، استفاده از یادگیری عمیق و شبکه ردیاب پیشنهادی به عنوان روش اصلی برای شناسایی و ردیابی همزمان اشیا و وسایل نقلیه، باعث می‌شود که این فرایند با سرعت و دقت بالاتری انجام شود؛ به‌طوری که نرخ دقت روش پیشنهادی، ۹۸,۲ درصد است که حداقل ۱۳ درصد، بهبود دقت نسبت به روش‌های قبلی داشته‌است.

کلید واژه ها

شناسایی و ردیابی همزمان اشیا، شناسایی اشیا، ردیابی اشیا، یادگیری عمیق

۱- مقدمه

بینایی رایانه طی سالیان اخیر، پیشرفت‌های زیادی را داشته‌است. بطوری که در بسیاری از پروژه‌های شناسایی و ردیابی اشیا، مورد استفاده قرار می‌گیرد. یک مثال خوب از بینایی رایانه، شبکه‌های اجتماعی مانند فیس‌بوک است که در آن از الگوریتم‌های تشخیص پیشرفته، برای شناسایی افراد استفاده می‌شود[۱]. در واقعیت، یکی از دلایل اصلی گسترش سریع بینایی رایانه، سیستم عامل‌های محاسباتی جدیدی است که همیشه در حال ظهور هستند؛ مانند این موضوع، طیف گسترده‌ای از موضوعات مختلف را شامل می‌شود؛ مانند یادگیری ماشین، پردازش تصویر و سیگنال، پردازش ویدئو و مهندسی کنترل. همچنین بینایی رایانه، این توانایی ارزشمند را می‌دهد که کارهای هوشمندانه‌تری انجام داد و در نتیجه، خدمات بهتری را برای اپراتورهای انسانی خود ارائه می‌دهد. در این دوره، بینایی رایانه به راحتی در برنامه‌های تجاری مانند تصویربرداری پزشکی[۲]، جستجوی تصاویر، رباتیک[۳] و سایر موارد راه یافته است. امروزه دو موضوع اصلی در بینایی رایانه، ردیابی و شناسایی اشیا است. ویژگی‌های استخراج شده از اشیا، مستقیماً در سیستم‌های کنترلی ربات‌ها، وسایل نقلیه هوشمند، پهپادها و

این مقاله در بهمن‌ماه ۱۴۰۱ دریافت، در اردیبهشت‌ماه ۱۴۰۲ بازنگری و در خردادماه پذیرفته شد.

[۱] دانشجوی دکتری الکترونیک، دانشکده برق و کامپیوتر، دانشگاه بیرجند
رایانامه: h.khatami@birjand.ac.ir

[۲] گروه الکترونیک، دانشکده برق و کامپیوتر، دانشگاه بیرجند
رایانامه: havangi@birjand.ac.ir

نویسنده مسئول: رمضان هاونگی

dorl.net/dor/20.1001.1.23831197.1402.10.3.5.7

سیستم‌های نظارتی هوشمند استفاده می‌شوند که این موضوع، امکان کنترل هوشمند و تصمیم‌گیری را فراهم می‌کند. تا همین اواخر، فناوری‌هایی مانند LIDAR [1] (تشخیص و اندازه‌گیری فاصله با استفاده از نور) ، ابزار اصلی برای تشخیص شی بوده‌است[۴].

دلیل استفاده از این فناوری، راحتی در پردازش داده‌ها در زمان واقعی بود. اما پیشرفت‌های اخیر و اکتشافات جدید باعث شد تا الگوریتم‌های ردیابی و همچنین استخراج ویژگی‌ها در سیستم‌های مدرن بینایی رایانه‌ای، عملکرد خوبی از خود نشان دهند.

اصلی‌ترین دلیل این واقعیت، این است که بینایی رایانه، اطلاعاتی را درمورد ظاهر اجسام و همچنین محیط اطراف اجسام جمع‌آوری و فراهم می‌کند. این موضوع به ویژه در سیستم‌های کمکی مثل دستیار راننده، که سیستم LIDAR به دلیل اندازه آن‌ها مناسب نیست، بسیار جذاب است.

بسیاری از سیستم‌ها برای داشتن عملکرد بهتر، نیاز به همراهی چندین سنسور مکمل دارند تا اطلاعات دقیق‌تر و بیشتری را از محیط و اشیا جمع‌آوری کنند. برخی از این سنسورها، سنسورهای دوربین چند منظوره، رادارهای برد کوتاه و برد بلند، سنسورهای اولتراسونیک، GPS و کیلومترشمار است. با استفاده از این سنسورها، عملکرد بهتری را می‌توان شاهد بود که این در دقت سیستم و استخراج اطلاعات کاربرد دارد. به عبارت دیگر، استفاده از سنسورهای مختلف که تحت عنوان مکمل در سیستم استفاده می‌شود، به‌منظور ایجاد یک سیستم قدرتمند است که امکان اجرای یک برنامه واحد در محیط‌های متفاوت را فراهم می‌کند که این موضوع، امری حیاتی است. پیشرفت‌های زیاد در زمینه ردیابی اشیا، موجب شده است تا به عنوان مثال، در زمینه نظارت بر مسیرهای رانندگی با استفاده از تکنیک‌های بینایی ماشین و شناسایی و ردیابی چندین شی بطور همزمان، امروزه شاهد افزایش ایمنی جاده و خودروها باشیم؛ که این موضوع تاثیر مستقیم در کاهش حوادث و همچنین کاهش خطرات انسانی دارد.

ردیابی و تشخیص اشیا، معمولاً بصورت دو عمل جداگانه مورد بحث و بررسی قرار می‌گیرد که انجام این کار بصورت یک فرآیند واحد، هنوز مورد بحث و چالش است. با پیشرفت علم، هنوز در بحث تشخیص اشیا در تصاویر(اجسام ثابت)، به ویژگی‌های ظاهری فضا و مکان مورد نظر بستگی دارد؛ در حالی که در بحث ردیابی اشیا در فیلم‌ها که ثابت نیستند و فریم‌ها به سرعت تغییر می‌کند، علاوه‌بر ویژگی‌های ظاهری مکان مورد نظر، به ویژگی‌های حرکتی در زمان‌های مختلف هم بستگی دارد. شبکه‌های یادگیری عمیق، موجب پیشرفت زیادی در در زمینه تشخیص اشیا در تصاویر دوبعدی شدند؛ مانند شبکه عصبی کانولوشن [2] که پیشرفت‌های زیادی را به‌دنبال داشته است.

همچنین تشخیص برجستگی‌ها در تصاویر و فیلم‌های ۳۶۰ درجه، از نتایج پیشرفت در زمینه شناسایی و ردیابی اشیا است؛ بگونه‌ای که در سالیان اخیر، عکس و فیلم‌های ۳۶۰ درجه، محبوبیت ویژه‌ای را با توسعه و پیشرفت در تکنولوژی واقعیت مجازی(VR)[3] و واقیعت افزوده (AR)[4] ، بدست آورده‌اند. نمایشگرهای HMD[5] یا همان نمایشگرهای "سربند مانند" که روی سر و چشم‌ها قرار می‌گیرد، محبوبیت روزافزونی پیدا کرده‌اند و کمپانی‌های بزرگی همچون Facebook و HTC در حال استفاده از این تکنولوژی هستند. کاربر با استفاده از این هدست‌ها، میتواند تصاویر و فیلم‌های ۳۶۰ درجه‌ای را ببیند و در آن غرق شود.

شکل (۱): نمایشگر سربند یا HMD

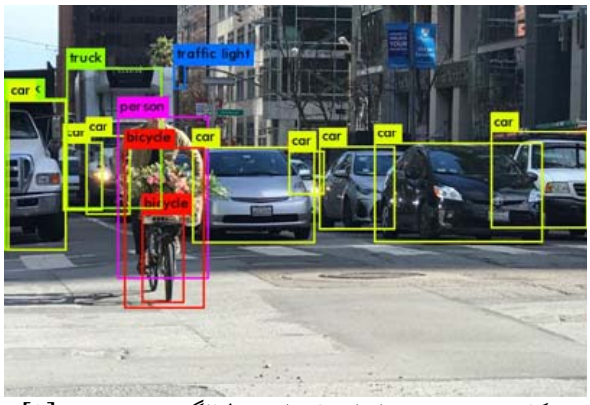
مطالعات علمی نشان می‌دهد که انسان، تمام صحنه‌های یک تصویر را با شدت یکسان مشاهده نمی‌کند. بلکه ذهن انسان، بیشتر به سمت قسمت‌های برجسته و مرتبط تصویر جلب می‌شود. لذا این ویژگی انسان، کاربردهای فراوانی دارد که از این ویژگی در ردیابی و تشخیص اشیای ویدئوها و تصاویر استفاده شده‌است.

راه‌حل‌های مختلفی برای تشخیص و شناسایی اشیا بررسی ارایه شده است. محبوب ترین الگوریتم‌ها عبارتند از الگوریتم YOLO [۵]، الگوریتم R-CNN [۶, ۷] و الگوریتم Deep MultiBox [۸] می باشند.

منظور از YOLO[6] این است که شما فقط یکبار به تصویر نگاه می‌کنید. این الگوریتم، برگرفته شده از قدرت سیستم بینایی انسان است. سیستم بینایی انسان با یک نگاه، می‌تواند اشیا را تشخیص دهد. به عبارت دیگر، YOLO یک الگوریتم شناسایی اشیا یا همان آشکارکننده است که با سرعت بسیار بالا، عمل شناسایی اشیا را انجام می‌دهد که این الگوریتم، مشابه سیستم بینایی انسان طراحی شده‌است. الگوریتم YOLO، دارای سرعت بالا و قدرت محاسباتی بالایی است. سیستم YOLO از روش‌هایی مانند پنجره‌های لغزان[7] و پیشنهادات منطقه[8] خودداری می‌کند و استفاده نمی‌کند. در روش پنجره‌های لغزان برای یافتن اشیا در تصویر، یک

[1] Light detection and ranging
[2] Convolutional neural network
[3] Virtual reality
[4] Augmented reality
[5] Head display mounted
[6] You only look once
[7] Sliding window
[8] Region proposals

قاب را از تمام تصویر عبور می‌دهد و سپس هر یک از این قاب‌ها، مورد ارزیابی قرار می‌گیرد و سپس اشیا شناسایی می‌شوند. به عبارت دیگر، طبقه بند به مناطق مختلف تصویر اعمال می‌شود و سپس شی شناسایی می‌شود. همچنین در روش پیشنهادات منطقه، نواحی در برگیرنده اشیا در تصویر، شناسایی می‌شوند.



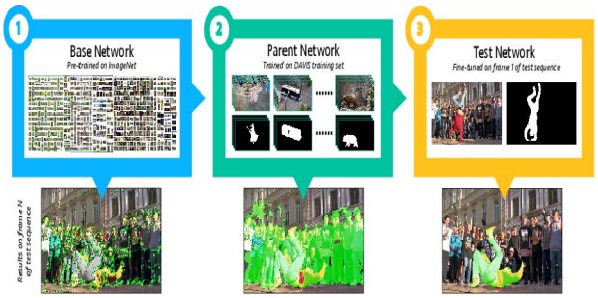
شکل (۲): نحوه شناسایی اشیا توسط الگوریتم YOLO [۵]

همانطور که در شکل (۲) دیده می‌شود، الگوریتم YOLO با یکبار عبور تصویر اصلی، تمام باکس‌های کاندیداهای احتمالی را مشخص می‌کند و به هرکدام، نمره اطمینان داده می‌شود. اگر این نمره اطمینان، کمتر از یک آستانه مشخص باشد، باکس مورد نظر حذف می‌شود. پس از آن، از الگوریتم حذف غیرحداکثرها[1] برای حذف چندین تشخیص از یک شی استفاده می‌شود. به عبارت دیگر، می‌توان گفت که الگوریتم YOLO به مسئله تشخیص اشیا، یک نگاه رگرسیونی دارد که به صورت مستقیم، مختصات باکس و احتمال کلاس‌ها را از پیکسل‌های تصویر به دست آورده است.

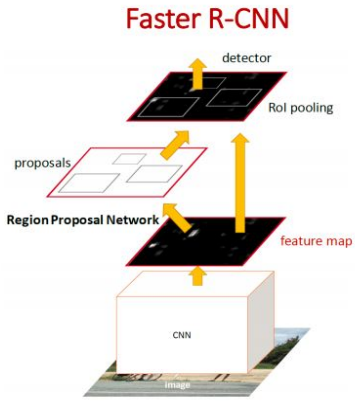
الگوریتم YOLO، به علت موازنه و تعادل خوبی که میان دقت و سرعت دارد، تبدیل به یکی از محبوب‌ترین ردیاب‌ها شده است.[۵]

همچنین R-CNN، یک سیستم تشخیص است که برای یافتن اشیا در تصاویر از پیشنهادات منطقه‌ای استفاده می‌کند. این روش، باکس‌های مرزی بالقوه‌ای را تولید می‌کند(در مجموع، ۲۰۰۰ باکس). سپس یک شبکه کانولوشن، ویژگی‌های این باکس‌های مرزی را استخراج می‌کند. پس از آن، سیستم از یک SVM (ماشین بردار پشتیبانی) استفاده می‌کند که به سادگی تشخیص می‌دهد که هدف، شی است یا نه و اگر شی است، چه نوع شیئی است. سپس با عملکرد حذف غیرحداکثرها، اگر هر شی بیش از یک‌بار شناسایی شده باشد؛ تمام تشخیص‌های تکراری، حذف می‌شوند.

[1] Non-max suppression



شکل (۳): تقسیم‌بندی آبجکت ویدئویی تک شاتی



شکل (۴): الگوریتم R-CNN Faster

در زمینه ردیابی اشیا، چندین الگوریتم در انجمن جهانی یادگیری عمیق ارائه شده‌است. برخی از این الگوریتم‌ها عبارتند از: شبکه-های فول کانولوشن سیامسی[2]، تقسیم‌بندی آبجکت ویدئویی تک شاتی[۸]، الگوریتم SORT[3] [۹] الگوریتم GOTURN[4] [۱۰].

تمام این روش‌ها، شناسایی و ردیابی اشیا را بصورت دو عمل جداگانه مورد بررسی قرار می‌دهند.

در این مقاله، با استفاده از الگوریتم R-CNN Faster و گسترش آن، به نتایج جدیدی دست یافتیم که می‌تواند عمل ردیابی اشیا را با استفاده از شناسایی و تشخیص اشیا، بصورت همزمان انجام دهد. در شکل(۴)، ساختار یک شبکه R-CNN Faster مشاهده می‌شود.

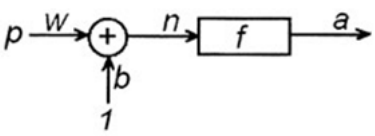
۲- شبکه‌های عصبی و روش یادگیری عمیق

شبکه عصبی، یک شبیه‌سازی از مغز انسان است. درواقع این مغز انسان است که شبکه عصبی طبیعی است و سایر شبکه‌های عصبی، مصنوعی هستند. مهم‌ترین ویژگی‌هایی که مغز انسان قابلیت آن را دارد، عبارت است از: قدرت یادگیری، ذخیره‌سازی اطلاعات، قدرت پیش‌بینی و قدرت محاسبه.

مغز انسان، شبکه گسترده‌ای از نرون‌ها (۱۰[11]) است که برای هر نرون ۱۰[4] اتصال وجود دارد. نرون همان ساده‌ترین واحد ساختاری شبکه عصبی است که خروجی آن، ورودی نرون دیگر است. اصلی‌ترین جزء در شبکه‌های عصبی مصنوعی نیز، نرون‌ها هستند. در این مدل، حرف w وزن(شدت سیناپس)، پارامتر p

[2] Fully-convolutional siamese networks
[3] Simple online realtime tracker
[4] Generic object tracking using regression networks

ورودی، f تابع تحریک(بدنه سلول)، b بایاس و پارامتر a نیز، خروجی است. پارامترهای w و b، پارامترهای قابل تنظیم هستند که با داده‌های آموزشی توسط یک الگوریتم تنظیم می‌شوند.



شکل (۵): مدل ریاضی نرون

چند نمونه از توابع تحریک مرسوم در نرون‌های مصنوعی عبارتند از: تابع تحریک خطی[1] که در شبکه‌های عصبی آدلاین کاربرد دارد، تابع تحریک by signal که معمولاً برای شبکه عصبی پس از انتشار (MLP)[2] استفاده می‌شود، تابع تحریک hard limit که در شبکه‌های عصبی پرسپترون کاربرد دارد، تابع تحریک hard limit متقارن و تابع تحریک شعاعی[3] (RBF) که در شبکه‌های عصبی تابع شعاعی پایه کاربرد دارد.

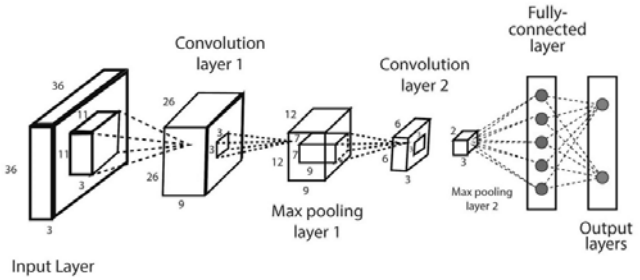
یکی از محبوب‌ترین شبکه‌های عصبی، شبکه عصبی کانولوشن است. این شبکه عصبی، ویژگی‌های آموزش داده‌شده را، به داده‌های ورودی متصل می‌کند و از لایه‌های کانولوشن دو بعدی استفاده می‌کند. این معماری، برای پردازش داده‌های دوبعدی، مانند تصاویر مناسب است. در واقع شبکه عصبی کانولوشن یا همان CNN، نوعی شبکه عصبی است که بصورت تخصصی به پردازش داده‌ها می‌پردازد و در شناسایی و طبقه‌بندی تصاویر، کاربرد زیادی دارد. این شبکه عصبی، دارای یک توپولوژی شبکه مانند است. در شبکه عصبی کانولوشن، بجای ضرب ماتریس‌ها، حداقل در یک لایه از عملیات کانولوشن استفاده می‌شود.

داده‌های سری زمانی، مانند سیگنال‌های صوتی که می‌توان بصورت یک شبکه یک‌بعدیدر فواصل زمانی منظم نمونه برداری کرد را می‌توان با استفاده از CNNهای یک بعدی مدل کرد. همچنین داده‌های تصویری که بصورت شبکه‌ای از پیکسل‌های دوبعدی است، با استفاده از CNNهای دوبعدی مدل می‌شوند.

داده‌های ویدئویی یا سایر داده‌هایی سه بعدی و بیشتر را نیز می‌توان با گسترش ابعاد آن‌ها مدل کرد. در سال‌های اخیر، شبکه عصبی کانولوشن در بسیاری از برنامه‌های دنیای واقعی، به ویژه در زمینه بینایی رایانه، طبقه بندی تصاویر، تشخیص و ردیابی اشیای تصاویر و فیلم‌ها، تقسیم بندی و تشخیص عملکرد انسان، مورد استفاده قرار گرفته‌است و پیشرفت‌های زیادی داشته‌است.

شبکه عصبی کانولوشن نیازی به استخراج دستی ویژگی‌ها ندارد؛ بنابراین نیازی به شناسایی ویژگی‌های مورد استفاده در طبقه‌بندی تصاویر نیست. این شبکه عصبی، با استخراج مستقیم ویژگی‌ها از تصاویر کار می‌کند و ویژگی‌ها از قبل آموزش داده نشده اند. این استخراج ویژگی خودکار، باعث می‌شود که مدل‌های یادگیری عمیق، برای کارهای بینایی رایانه مانند طبقه‌بندی اشیا، بسیار دقیق باشند.

در شکل(۶)، فیلترها با رزولوشن‌های مختلف، به هر تصویر که درحال آموزش است، اعمال می‌شوند و خروجی هر تصویر کانوال شده، به عنوان ورودی لایه بعدی اعمال می‌شود. شبکه عصبی کانولوشن آموزش داده شده با استفاده از ده‌ها و صدها لایه مخفی، ویژگی‌های مختلف یک تصویر را تشخیص دهد. هر لایه مخفی، پیچیدگی ویژگی‌های تصویر آموزش داده شده را افزایش می‌دهد؛ به عنوان مثال، اولین لایه مخفی می‌تواند شیوه تشخیص لبه‌ها را آموزش دهد.



شکل (۶): شبکه عصبی کانولوشن با تعدادی لایه کانولوشن

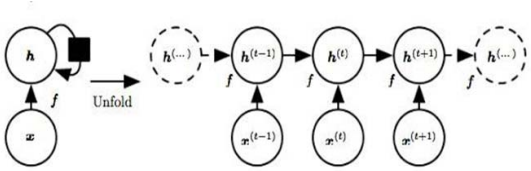
همچنین لایه آخر یاد می‌گیرد که چگونه اشکال سخت‌تر و پیچیده‌تر را تشخیص دهد. شبکه عصبی کانولوشن، از چند ایده اساسی برای بهبود سیستم‌های یادگیری ماشین استفاده می‌کند.

یکی از پر کاربردترین شبکه‌های عصبی در پردازش تصویر، شبکه عصبی بازگشتی است. شبکه عصبی بازگشتی (RNN)[4]، نوعی شبکه عصبی است که در پردازش داده‌های متوالی، مورد استفاده قرار می‌گیرد. به عبارت دیگر، یک نوع توالی زمانی و گراف جهت‌دار وجود دارد و داده‌ها با گذر زمان، انتقال پیدا می‌کنند. شبکه عصبی بازگشتی، می‌تواند یک یا چند فیدبک داشته باشد که این بازگشت‌ها و فیدبک‌ها را، با یک یا چند عنصر تاخیر[5] ایجاد می‌کنند. در شبکه عصبی بازگشتی، مولفه زمان از اهمیت زیادی برخوردار است؛ در حالی که در شبکه عصبی کانولوشن، اهمیت زیادی به مولفه‌های زمانی داده نمی‌شود. در نتیجه، شبکه عصبی بازگشتی، برای داده‌هایی مناسب است که با زمان ارتباط دارند.

در هنگام مدل‌سازی یک سری زمانی، یک شبکه انتقال مستقیم، پارامترهای جداگانه و مجزایی برای هر ورودی دارد. این شبکه عصبی، از حافظه داخلی با هدف پردازش دنباله داده‌های ورودی استفاده می‌کند؛ یعنی ورودی‌های جدید، به ورودی‌های قبلی وابسته است و یک ارتباط دنباله‌ای بین ورودی‌ها وجود دارد که از آن در کاربردهای مختلف، مانند پردازش زبان طبیعی و تشخیص گفتار، استفاده می‌شود. قیمت سهام در بورس، اطلاعات دریافتی و ثبت شده از حسگرهای مختلف و حتی سوابق پزشکی، از

[1] Pure line
[2] Multi layer perceptron
[3] Radial basis function
[4] Recurrent neural network
[5] Delay

کاربردهای شبکه عصبی بازگشتی است که با تغییر زمان، تغییر می‌کند.



شکل (۷): مدل بازشده شبکه عصبی بازگشتی

برخلاف شبکه‌های عصبی قدیمی که توانایی استفاده از ورودی‌های قبلی برای استفاده در ورودی‌های بعدی را ندارند؛ شبکه عصبی بازگشتی، این قابلیت را دارد که ورودی‌های جدید را بر اساس ورودی‌های قبلی پردازش کند. شبکه عصبی بازگشتی، بصورت حلقه‌های تکرارشونده عمل می‌کند که این موضوع باعث ماندگاری اطلاعات می‌شود. همچنین شبکه عصبی بازگشتی با لیست‌ها و دنباله‌ها مرتبط است؛ در نتیجه، زمانی که داده‌های ترتیبی داشته باشیم، باید از شبکه عصبی بازگشتی بهره برد. به عبارت دیگر، مدل‌سازی بر روی یک دنباله یا زنجیره‌ای از بردارها انجام می‌شود.

شبکه عصبی بازگشتی، به همه لایه‌ها، وزن و بایاس یکسانی را می‌دهد. این مسئله، موجب پیچیدگی کمتر در پارامترها می‌شود و سپس از خروجی قبلی به‌عنوان ورودی لایه بعدی استفاده می‌کند. این عملیات باعث می‌شود که خروجی‌های قبلی حفظ شود. قواعد یادگیری شبکه‌های عصبی، برای اصلاح و تغییر وزن‌ها و بایاس‌ها، مورد نیاز است. در حالت کلی، سه نوع روش یادگیری برای شبکه‌های عصبی وجود دارد؛ یادگیری با نظارت، یادگیری بدون نظارت و یادگیری تقویتی.

در بحث یادگیری شبکه‌های عصبی، از یادگیری ماشین و یادگیری عمیق استفاده می‌شود. یادگیری ماشین، زیر شاخه‌ای از تکنیک‌های هوش مصنوعی است که از روش‌های آماری و غیرآماری، در جهت توانمندسازی کامپیوتر برای بهبود خود، با بهره‌گیری از تجربیات گذشته استفاده می‌کند. به عبارت دیگر، یادگیری ماشین این امکان را به سیستم می‌دهد که به‌صورت خودکار، یادگیری داشته‌باشد و پیشرفت کند. در واقع، یادگیری ماشین، توانایی یادگیری مستقل را به ماشین‌ها می‌دهد و این قدرت را به ماشین می‌دهد که بر اساس مشاهده و تجربه و الگوها، آموزش ببینند. هوش مصنوعی و یادگیری ماشین، در عین حال که مجزای از یکدیگر هستند؛ کاملا متصل به یکدیگر هستند. تقلید و شبیه سازی رفتاری مشابه رفتار انسان، از اهداف هوش مصنوعی است؛ درحالی که یادگیری ماشین، بیشتر به جنبه‌های نوشتن نرم‌افزار می‌پردازد و می‌تواند از تجربیات گذشته استفاده کند. به‌عنوان مثال، یادگیری ماشین، نقش بسیار زیادی در پیشرفت سرویس‌های ارائه شده در فضای مجازی دارد.

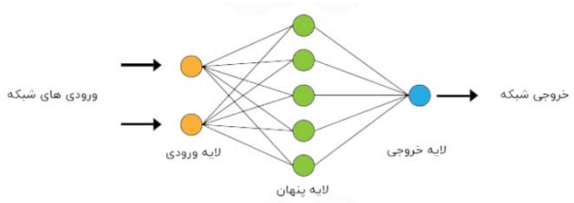
برای شروع یادگیری ماشین، از مشاهدات یا داده‌ها شروع می‌کنند و سپس سیستم از دستورالعمل‌ها و مثال‌ها در جهت رسیدن به یک الگوی مشخص استفاده می‌کند؛ سپس حل مسئله و تصمیم‌گیری آغاز می‌شود. در واقع در یادگیری ماشین، با استفاده از داده‌هایی که وجود دارد؛ مدل را آموزش می‌دهند و سپس از این مدل، برای پیش‌بینی داده‌های جدید استفاده می‌شود. هدف اصلی یادگیری ماشین، تنظیم رفتار سیستم به صورت اتوماتیک و بدون نیاز به دخالت انسان است.

همچنین در یادگیری عمیق، یک مدل کامپیوتری یاد می‌گیرد که کارهای طبقه‌بندی را مستقیماً از روی تصاویر، متن یا صدا انجام دهد. مدل‌های یادگیری عمیق می‌توانند به سطح دقت بالایی برسند و حتی گاهی اوقات از عملکرد انسان بالاتر می‌روند. مدل‌ها با استفاده از مجموعه بزرگی از داده‌های دارای برچسب و همچنین ساختارهای شبکه عصبی که حاوی لایه‌های بسیاری هستند، آموزش می‌بینند. یادگیری عمیق، دقت تشخیص بالایی دارد؛ در نتیجه، به لوازم الکترونیکی مصرفی این امکان را می‌دهد تا انتظارات کاربر را برآورده کند و برای کاربردهای حیاتی مانند خودروی بدون راننده، بسیار حائز اهمیت است. پیشرفت‌های اخیر در یادگیری عمیق تا حدی بهبود یافته است که یادگیری عمیق در برخی از کارها مانند طبقه‌بندی اشیا، در تصاویر، از انسان بهتر عمل می‌کند.

یادگیری عمیق برای اولین بار در سال ۱۹۸۰ نظریه‌پردازی شد. اما در سال‌های اخیر به دو دلیل مورد استفاده زیادی قرار گرفته‌است و مفید نشان داده‌است؛ دلیل اول اینکه، یادگیری عمیق به مقدار زیادی داده دارای برچسب نیاز دارد. به عنوان مثال، توسعه اتومبیل‌های بدون راننده به میلیون‌ها تصویر و هزاران ساعت فیلم نیاز دارد و دلیل دوم اینکه، یادگیری عمیق به قدرت محاسباتی قابل توجهی نیاز دارد.به عنوان مثال، پردازنده‌های گرافیکی با عملکرد بالا، دارای معماری موازی هستند که برای یادگیری عمیق، بسیار کارآمد است. این ترکیب با خوشه‌ها یا رایانش ابری، تیم‌های توسعه را قادر می‌سازد تا زمان آموزش یک شبکه یادگیری عمیق، را از هفته به ساعت یا حتی کمتر کاهش دهند.

بیشتر روش‌های یادگیری عمیق، از معماری شبکه عصبی استفاده می‌کنند. به همین دلیل است که اغلب اوقات به مدل‌های یادگیری عمیق، شبکه‌های عصبی عمیق نیز می‌گویند. اصطلاح "عمیق"، معمولاً به تعداد لایه‌های مخفی در یک شبکه عصبی اشاره دارد. شبکه‌های عصبی سنتی، فقط شامل ۲-۳ لایه پنهان هستند؛ درحالی که شبکه‌های عمیق می‌توانند تا ۱۵۰ لایه داشته باشند. مدل‌های یادگیری عمیق، با استفاده از مجموعه‌های بزرگی از داده‌های دارای برچسب و معماری شبکه عصبی آموزش می‌بینند؛ که این مسئله، بدون نیاز به استخراج ویژگی‌ها و به صورت دستی، ویژگی‌ها را مستقیماً از داده‌ها یاد می‌گیرند.

همانطور که در شکل (۸) می‌بینید، شبکه‌های عصبی از سه لایه ورودی، پنهان و خروجی تشکیل شده‌اند. شبکه عصبی، می‌تواند هزاران لایه پنهان داشته باشد.



شکل (۸): ساختار شبکه عصبی

سه روش رایج برای برای طبقه‌بندی اشیا با استفاده از یادگیری عمیق، عبارت اند از، آموزش از ابتدا [1]، یادگیری انتقالی [2] و استخراج ویژگی‌ها [3].

آموزش و یادگیری انتقالی، در اکثر برنامه‌های یادگیری عمیق، استفاده می‌شود. یک فرآیند که شامل تنظیم دقیق یک مدل آموزش دیده‌است. با یک شبکه عصبی موجود، مانند Alex net یا net Google شروع می‌شود و در داده‌های جدید تنظیم می‌شود. پس از ایجاد برخی تغییرات، می‌توان یک کار جدید مثل دسته‌بندی گربه یا سگ را به جای دسته‌بندی ۱۰۰۰ شی مختلف انجام داد. مزیت بسیار مهم این روش، داده‌های کم آن است(بجای پردازش میلیون‌ها تصویر، از هزاران تصویر استفاده می‌شود)؛ در نتیجه، مدت زمان محاسبه به دقیقه و ساعت کاهش می‌یابد. استفاده از شبکه عصبی برای استخراج ویژگی‌ها، یک رویکرد معقول‌تر و تخصصی‌تر در یادگیری عمیق است. از آن‌جا که تمام لایه‌ها، وظیفه آموزش برخی ویژگی‌ها از تصاویر را دارند، می‌توان در زمان آموزش، این ویژگی‌ها را از شبکه خارج کرد. بنابراین از این ویژگی‌ها می‌توان به‌عنوان ورودی در یک مدل یادگیری ماشین، مانند ماشین‌های بردار پشتیبانی(SVM [4]) استفاده کرد.

۳- شناسایی و ردیابی همزمان با استفاده از یادگیری عمیق

از تشخیص و ردیابی اشیا، به‌عنوان دو عمل مجزا یاد می‌شود. تشخیص اشیا در تصاویر، به ویژگی‌های ظاهری مکانی که در آن قراردارند، بستگی دارد؛ در حالی که ردیابی اشیا در ویدئوها و فیلم‌ها، علاوه بر ویژگی‌های ظاهری مکان مورد نظر، به ویژگی‌های حرکتی در زمان‌های مختلف نیز بستگی دارد.

با استفاده از شبکه‌های یادگیری عمیق، پیشرفت زیادی در زمینه ردیابی اشیا در تصاویر دوبعدی حاصل شده است. ردیابی اشیا با استفاده از تشخیص و شناسایی اشیا، مستلزم آن است که جسم در اولین فریم و تمام فریم‌های بعدی، شناسایی شود و سپس به کمک تطابق دادن تصاویر، ردیابی انجام شود. ردیابی و شناسایی اشیا بطور همزمان، هنوز چالش برانگیز و مشکل است؛ در این مقاله، یک ساختار شبکه‌ای جدید را ارائه شده‌است که می‌تواند با گسترش فریم R-CNN، یک شی متحرک را در یک بخش از ویدئو محصور کند. این مدل، برای تشخیص و ردیابی وسایل نقلیه، آموزش و تست شد و نتایج خوبی را نشان داد.

در این مقاله یک شبکه ردیاب [5] ارائه شده‌است که در شکل (۹)، ساختار کلی این شبکه، نشان داده شده‌است. این شبکه، می‌تواند اشیا متعدد را در ویدئوها به‌طور مشترک با تولید لوله‌های محدود-کننده، شناسایی و ردیابی کند. با استفاده از ویژگی‌های مکانی-زمانی استخراج‌شده توسط یک شبکه عصبی کانولوشن سه‌بعدی، VGG طرح‌های پیشنهادی لوله را ایجاد کرده و آن‌ها را طبقه‌بندی و موقعیت آن‌ها را اصلاح می‌کند.

شبکه ردیاب پیشنهادی شامل سه مرحله است؛ استخراج ویژگی و تبدیل مکانی، شبکه پیشنهاد لوله (TPN [6]) و طبقه‌بندی و پالایش پس از TPN. چندین روش برای انجام پیشنهادات لوله و تنظیم رگرسیون بررسی شده‌است. شبکه ردیاب پیشنهادی در مورد چالش‌های موجود در مجموعه داده ویدئویی UA-DETRAC آموزش داده شد و مورد آزمایش قرار گرفت.

کاری که در این مقاله انجام شده‌است، بر اساس ساختار R-CNN Faster است و از این الگوریتم، الهام گرفته‌است. در این روش، RPN با شبکه پیشنهاد لوله که در یک شبکه کانولوشن سه‌بعدی کار می‌کند، جایگزین شده‌است. سپس از مکانیزم تجمع ROI سه‌بعدی استفاده شد. همچنین تابع هدف، کاهش رگرسیون است که برای در نظر گرفتن تفاوت بین ground truth و مکان‌های کشف شده لوله در تمام فریم‌ها، گسترش یافته‌است که البته با افت طبقه‌بندی لوله، بهینه شده‌است.

منظور از ground truth، همان پاسخ‌های صحیح مبنا است. برای استفاده از هر دو ویژگی مکانی و زمانی، شبکه مبتنی بر شبکه VGG [7] است که از طریق پایگاه داده "Image net" آموزش دیده است و همچنین شبکه C3D (شبکه کانولوشن سه‌بعدی) که از طریق پایگاه داده UCF101 آموزش دیده است و برای طبقه‌بندی استفاده می‌شود.

در این مقاله، یک ویدئو به گروهی از تصاویر(GOP [8]) با طول ثابتT (یعنی ۸ فریم در هربار اجرا) تقسیم شده‌است. سپس فریم-های ویدئویی خام را در هر GOP، به یک ساختار دو سر محور تغذیه می‌کند. سر اول، یک شبکه VGG و سر بعدی، یک شبکه C3D [9] است. در C3D، یک ساختار CNN کانولوشن سه‌بعدی را برای استخراج ویژگی‌های مکانی-زمانی به منظور طبقه‌بندی فیلم‌ها در دسته‌های مختلف پیشنهاد شده‌است. C3D، عملکرد مناسبی را در مقایسه با الگوریتم‌های بینایی رایانه سنتی نشان داده‌است. از

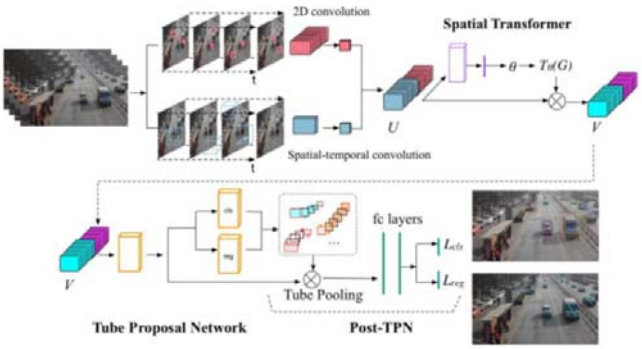
[1] Training from Scratch
[2] Transfer learning
[3] Feature extraction
[4] Support vector machines

[5] Track NET
[6] Tube proposal network
[7] Visual graphics group
[8] Group of pictures
[9] 3D convolutional network

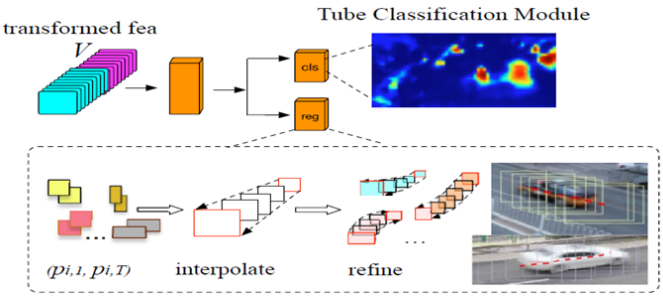
عملکردهای خوب آن، می‌توان به کارایی بالا در مسیرهای متراکم و شلوغ اشاره کرد.

همچنین ثابت شد که ویژگی‌های استخراج شده توسط این شبکه-های C3D، برای تشخیص اشیا در فیلم مناسب است و کارایی خوبی دارد. اگر چه C3D به‌طور خاص برای مکان‌یابی دقیق اشیا آموزش ندیده‌است؛ اما لوکیشن‌هایی که در نقشه‌ها به‌شدت فعال هستند، معمولاً مربوط به اشیا درحال حرکت هستند. بنابراین از ساختار شبکه C3D برای استخراج ویژگی‌ها در TrackNet استفاده شد و وزن‌های شبکه را از طریق آموزش، تصحیح و بروزرسانی شده‌است. همانطورکه در شکل (۹) می‌بینید، ساختار شبکه ردیاب، هردو ویژگی مکانی و زمانی را از GOP ویدئویی تولید می-کند. خروجی نهایی لوله‌های محدودکننده، برای حرکت اجسام داخل این GOP ویدئویی خواهد بود.

.



شکل (۹): ساختار کلی شبکه ردیاب (Track Net)



شکل (۱۰): شبکه پیشنهاد لوله یا TPN

بسیاری از پیشنهادات اولیه لوله را TPN تولید می‌کند. مشابه شبکه پیشنهادی R-CNN Faster، TPN چندین لوله متصل کاندید را در هرمکان پیکسل تولید می‌کند.

یک TPN از دو بخش تشکیل شده‌است. بخش اول، ماژول طبقه‌بندی و بخش دوم، ماژول رگرسیون آفست است. در نقشه حرارتی، مناطقی که اشیا در حال حرکت هستند، مقادیر بالاتر و گرم‌تری دارند. این موضوع در شکل (۱۰) نشان داده شده‌است.

در شکل (۱۱) نمونه امتیازات ۹ شی که نقشه حرارتی آن‌ها در سمت راست تصویر، نشان داده شده‌است. به عنوان مثال، تصویر بالا سمت چپ، مربوط به اندازه ۴۰ و نسبت تصویر ۰.۸۱ است. دو روش کشف-شده برای وایر کردن ماژول رگرسیون آفست لوله وجود دارد.

روش اول، پیش‌بینی مستقیم آفست‌ها برای تمام پیش‌بینی‌ها است و روش دوم، استفاده از روش درون‌یابی خطی است.

شکل (۱۱): نقشه حرارتی

در روش اول، به طور مستقیم، پارامترهای لوله را پیش‌بینی می‌کند. در این ساختار، به طور مستقیم آفست هر فریم تخمین زده می‌شود. با توجه به یک GoP ویدئویی به طولT ، شبکه رگرسیون مستقیماً پارامترهای 4×T را برای هر لوله پیش‌بینی می‌کند.

از آنجایی که کاندیدهای لوله در تمام مکان‌های پیکسل پخش می‌شوند، شبکه رگرسیون با استفاده از یک لایه کانولوشن با 4×T×M نقشه خروجی را پیاده‌سازی می‌کند. در روش دوم، از درون‌یابی خطی آفست‌های جعبه باندینگ از آفست‌ها در ۲ فریم استفاده می‌شود. علی‌رغم این واقعیت که یک شی در داخل یک ویدئو، می‌تواند حرکات خودسرانه و پیچیده‌ای داشته‌باشد؛ حرکات بیشتر اشیا در ویدئوهای دنیای واقعی بسیار نرم است. با توجه به یک دوره زمانی کافی کوتاه، می‌توان مسیر هر گوشه از لوله باندینگ را با یک خط مستقیم تقریب زد. این امر به ویژه برای ویدئوهای ترافیکی حاوی وسایل‌نقلیه در حال حرکت، بشدت خود را نشان می‌دهد. با این مشاهدات، شبکه رگرسیون به جای تعیین آفست موقعیت‌های گوشه‌ها در همه فریم‌ها، فقط آفست‌ها را در فریم‌های ابتدایی و انتهایی تخمین می‌زند و به صورت خطی، آفست‌ها را در فریم‌های دیگر درون‌یابی می‌کند.

روش دوم(درون‌یابی خطی) را برای پیشنهادات لوله در مرحله TPN، به منظور صرفه‌جویی و حفظ پارامترها و محدودکردن حرکات نرم و رهایی از روش اول در مرحله post-TPN، برای اصلاح بیشتر مکان‌ها استفاده شد.

همانطورکه در شکل (۹) نشان داده شده‌است، شبکه پیشنهادی لوله، پیشنهادات لوله زیادی را تولید می‌کند که موقعیت آن‌ها توسط لوله‌های کاندید اصلی و جابجایی‌های پیش‌بینی شده، تعیین می‌شود. لوله‌های پیشنهادی با امتیازات بالایی از مرحله دوم طبقه‌بندی و رگرسیون، عبور خواهد کرد. در این مرحله، پیشنهادات لوله به طبقات مختلف، مانند ماشین، اتوبوس، ون و غیره برای مجموعه داده UA-DETRAC، طبقه‌بندی می‌شوند. مجموعه داده UA-DETRAC، یک مجموعه داده در دانشگاه

آلبانی است که برای ارزیابی سیستم‌های MOT[1] و آشکارسازها استفاده می‌شود. این مجموعه داده، شامل ۱۰۰ ویدئوی چالشی است که از صحنه‌های ترافیکی تهیه شده‌است. در این ویدئوها، بیش از ۱۴۰ هزار فریم با اطلاعات زیادی از جمله میزان روشنایی، نوع وسایل نقلیه، نسبت برش و کادرهای محدودکننده وسایل نقلیه، موجود می‌باشد که برای شناسایی و ردیابی اشیا استفاده می‌شود. سپس موقعیت تنظیم شده برای لوله نیز اصلاح خواهد شد. به‌جای استفاده از ویژگی‌های ترکیب‌شده از مجاورت در نقشه ویژگی مانند TPN، ویژگی‌های خاص برای نواحی لوله پیشنهادی با استفاده از ادغام لوله ترکیب می‌شوند.

در لوله‌های pooling، امکان توصیف نواحی پیشنهادی مختلف توسط بردارهای ویژگی که ابعادی یکسان دارند، از طریق ادغام ROI فراهم می‌شود. در این مورد، یک لوله پیشنهاد می‌شود که متشکل از جعبه‌های محدودکننده در فریم‌های مختلف است که در اندازه‌ها و مکان‌های مختلف، متفاوت هستند. به‌جای استفاده از همان نقشه ویژگی، اجتماع همه جعبه‌های مرزی در یک لوله پیشنهادی، یافت می‌شود و ویژگی‌هایی که منطقه مورد نظر را پوشش می‌دهند، از نقشه‌های ویژگی تبدیل‌شده، استخراج می‌شوند.

اغلب دیتاست‌های تشخیص اشیا، تصاویر دوبعدی هستند؛ مانند Image Net، PASCAL VOC [۱۱]، Microsoft COCO [۱۲] و غیره. در چالش ILSVRC 2015 ، ImageNet [۱۳] و ظیفه VID را با ۳۰ دسته برای جلب توجه در تشخیص اشیا در ویدیوها معرفی کرد. با این حال، بسیاری از ویدئوها تنها شامل اشیا غالب بسیار کمی هستند. در حالی که در دنیای واقعی، تشخیص اشیا چندگانه و ردیابی اشیا چندگانه باید به‌طور همزمان مورد توجه قرارگیرند. به‌عنوان مثال، در تجزیه و تحلیل ترافیک و رانندگی خودکار، تشخیص و ردیابی دقیق وسایل نقلیه به‌خصوص در منطقه شهری شلوغ، یک چالش بزرگ باقی می‌ماند. کارهای موجود بر روی ردیابی خودرو در مناطق شهری که از روش‌های سنتی مانند مدل‌سازی پس‌زمینه یا خوشه‌بندی نقاط مشخصه استفاده می‌کنند، عملکرد محدودی دارند.

به منظور ارزیابی مدل، هم در تشخیص و هم در ردیابی اشیا متعدد، از مجموعه دادهUA-DETRAC استفاده شده‌است[۱۴]. این روش به معرفی معیاری برای تشخیص و ردیابی اشیا می‌پردازد، که از سکانس‌های ویدیویی چالش‌برانگیز گرفته‌شده از صحنه‌های واقعی با زوایای دید مختلف تشکیل شده‌است. در این مقاله، مجموعه داده را به ۴۵ ویدئوی آموزشی و ۱۵ ویدئوی آزمایشی تقسیم شدهاست و اطمینان حاصل‌شد که هم آموزش و هم آزمایش، تمام ویوها و میدان‌های دید مختلف دوربین را پوشش می‌دهند.

محدوده طول فیلم از حدود ۷۰۰ فریم تا ۲۵۰۰ فریم است. این مجموعه داده، انواع مختلفی از آب و هواهای مختلف مانند آفتابی، ابری، بارانی و شبانه را پوشش می‌دهد. همچنین از مجموعه داده نمونه‌برداری نشد تا مطمئن شویم که مجموعه آموزش و آزمایش، هر کدام شامل نمونه‌هایی است که تحت شرایط آب و هوایی متفاوت گرفته شده‌است. با این حال، معلوم شد که مدل آموزش‌دیده نسبت به شرایط آب و هوایی مختلف، به جز چند فیلم شبانه که شرایط نوری بسیار متفاوتی نسبت به بقیه دارند، بسیار قوی است. علاوه بر این، دوربین‌های مورد استفاده برای ثبت ویدئوهای شبانه نیز در شرایط نور کم، مشکلات خارج از فوکوس داشتند و به تصاویر تار منجر شدند که ممکن است باعث افت عملکرد شده باشند شبکه ردیاب پیشنهادی، در شرایط نوری متفاوت مقاوم است و می‌تواند اندازه خودروهای کوچک و بزرگ را با نسبت‌های تصویر متفاوت پوشش دهد.

همچنین لوله‌های باندینگ پراکنده‌تر را برای وسایل نقلیه سریع و لوله‌های باندینگ متراکم‌تر را برای وسایل نقلیه کندتر یا وسایل نقلیه‌ای که دورتر هستند، تولید می‌کند.

۴- نتایج

در طول آموزش، ۸ فریم ویدیوی متوالی به طور تصادفی از مجموعه آموزشی انتخاب می‌شوند. در این مقاله، ابتدا شاخه VGG را به تنهایی تحت فریم R-CNN Faster، با استفاده از کل مجموعه آموزشی به‌عنوان یک نقشه حرارتی تنظیم می‌شود. سپس TrackNet پیشنهادی با شاخه VGG آموزش‌دیده است. همچنین آخرین لایه کانولوشن در ستون C3D تنظیم می‌شود. نرخ یادگیری اولیه، 0.001 بود و ۱۰ بار بعد از k10 تکرار، کاهش یافت. لازم به ذکر است که در این مقاله از یک سخت‌افزار با پردازنده core i7 نسل ۷، رم ۱۶ گیگابایتی DDR4 ، کارت گرافیک Geforce 940 mx و هارد SSD استفاده شده‌است.

همچنین از بهینه‌ساز Adam استفاده شده است که نیاز به حافظه زیادی ندارد و لحاظ محاسباتی، بهینه است و به آسانی پیاده‌سازی می‌شود. بهینه ساز Adam، مدل را برای ۵۰ هزار تکرار آموزش داده‌است و شبکه ردیاب را با استفاده از COCO API استاندارد ارزیابی کرده‌است. برخی از نتایج ردیابی و تشخیص بصری در شکل (۱۲) نشان‌داده شده‌است.

در پیاده‌سازی فعلی، در طول آموزش، ۸ فریم ویدیویی متوالی به طور تصادفی از مجموعه آموزشی انتخاب می‌شوند. با این حال، با توجه به این واقعیت که وسایل نقلیه مختلف دارای سرعت‌های متفاوتی در ویوهای مختلف هستند(به عنوان مثال، حرکت بزرگ‌تر، هنگامی که وسایل نقلیه به دوربین نزدیک‌تر هستند و حرکت کوچک‌تر، هنگامی که دور هستند)، آموزش فقط با "سرعت‌های اصلی" کافی نیست.

در این مقاله، یک ترفند آموزشی پیشنهاد شد؛ اینکه صرف‌نظر از فریم، برای این‌که شبکه ردیاب آموزش داده‌شود. در طول آموزش،

[1] Multiple object tracking

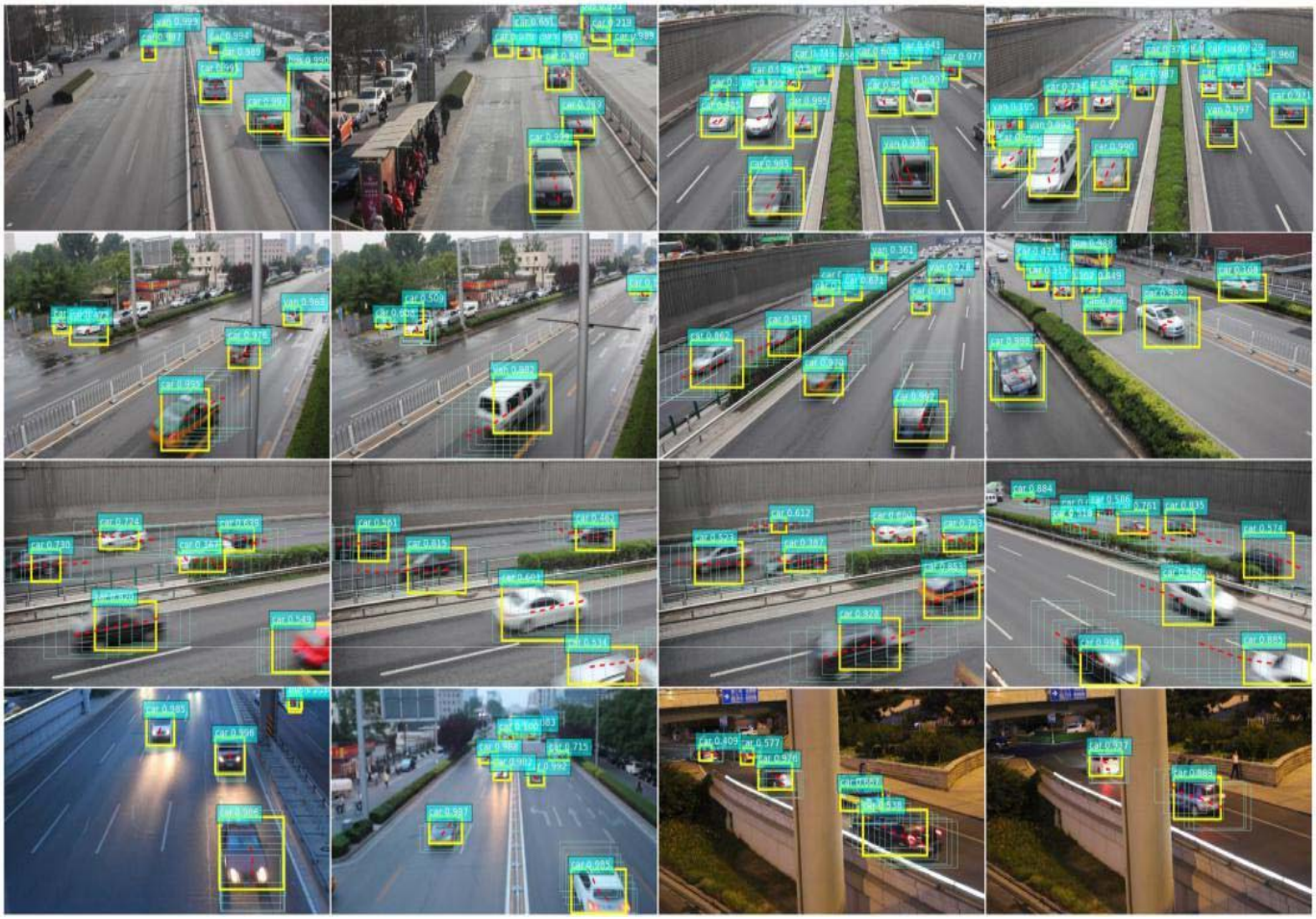
قبل از آوردن یک GoP جدید، یک ضریب پرش که شامل یک عدد صحیح تصادفی s انتخاب می‌شود(در بازه [0,5]). سپس به جای واکشی یک GoP بلند ۸ فریم متوالی، یک GoP نمونه‌گیری شده و فریم‌های بعدی، واکشی می‌شوند. برای مثال وقتی S=1 است، فریم ۰ ، ۲، ۴، ۶، ۸، ۱۰، ۱۲، ۱۴ به جای فریم ۰ ، ۱، ۲، ۳، ۴، ۵، ۶، ۷ انتخاب می‌شوند. وقتی که S=4 است، فریم ۱۳، ۱۷، ۲۱، ۲۵، ۲۹، ۳۳، ۳۷، ۴۱ به جای فریم ۱۳، ۱۴، ۱۵، ۱۶، ۱۷، ۱۸، ۱۹، ۲۰ انتخاب می‌شوند. با انتخاب تصادفی عامل پرش در طول آموزش، شبکه نسبت به سرعت‌های مختلف قوی‌تر است. برخی از نتایج ارزیابی در جدول(۱) و جدول(۲) نشان‌داده شده‌است.

R-CNN Faster، به‌عنوان خط اصلی تعیین شد. همچنین کل مدل R-CNN Faster را با VGG به‌عنوان شبکه پایه، روی همان مجموعه داده آموزشی برای ۷۰ هزار تکرار ، تنظیم گردید. پس از تنظیم دقیق، R-CNN Faster، به نرخ فراخوانی بسیار بالایی در مجموعه داده UA-DETRAC برای تشخیص وسایل نقلیه دست یافت. همچنین شبکه ردیاب با استفاده از دو مجموعه معیار ارزیابی شد؛ مجموعه اول، عملکرد تشخیص شی در هر فریم را با استفاده ازCOCO API بررسی می‌کند و مجموعه دوم، عملکرد ردیابی شی را با استفاده از معیارهای MOT بررسی می‌کند.

به منظور درک نقش‌هایی که اجزای اصلی طراحی را بر عهده دارند، انواع مختلف TrackNet آموزش و آزمایش می‌شوند. در جدول(۱) و (۲)، مقایسه بین TrackNet بدون ترانسفورمر و TrackNetبدون اتصال ویژگی VGG و TrackNet بدون ترانسفورمر یا VGG در حالت "پیش‌بینی همه" یا حالت درون‌یابی خطی(LP) در طول TPN و و *TrackNet که ورژن کامل‌تری است، نشان داده شده‌است. مجموعه داده‌های آموزشی و آزمایشی را بر اساس زوایای‌دید به نمای‌چپ، نمای‌راست و نمای‌جلو تقسیم شد و عملکردها را هنگام آموزش و با توجه به مجموعه داده، سطوح اختلاف برای زوایای دید مختلف متفاوت است.

به عنوان مثال، مجموعه داده فرعی نمای جلویی، ساده‌ترین است. مدل پیشنهادی پس از اینکه داده‌های آموزشی را با تغییر افقی افزایش داد، بهبود یافت. مشاهده می‌شود که مدل پیشنهادی با توجه به داده‌های آموزشی بیشتر و یا استفاده از دیگر ترفندهای افزایش داده، حتی بهتر عمل می‌کند.

همچنین استفاده از آموزش"فریم‌های در حال پرش"، توانست به افزایش عملکرد کمک کند. به منظور مقایسه شبکه ردیاب با مبنای تشخیص دوبعدی، تمام لوله‌های محدوده تولید شده توسط شبکه ردیاب در نظر گرفته‌شدند و تمامی جعبه‌های محدوده کننده در هر فریم ارزیابی شد.



شکل (۱۲): نمونه‌هایی از لوله‌های باندینگ پیش‌بینی شده

در جدول(۱)، نرخ‌های میانگین دقت(AP%)، تحت تنظیمات مختلف گزارش شده‌است (یعنی آستانه‌های IoU و ناحیه جعبه باندینگ). تمام متغیرهای شبکه ردیاب گزارش‌شده در اینجا از ۳۰۰ پیشنهاد بالا در طول آزمایش استفاده کردند؛ به جز موردی که با ۲۰۰۰ پیشنهاد نشان داده شده‌است.

جدول (۱): نتایج تشخیص انواع مختلف TrackNtet بر روی مجموعه داده UA – DETRAC (ارزیابی با استفاده از COCO API).

					AP @ IOU			AP area		
	T_0(G)	VGG	C3D	LP	0.1:1	0.1	0.5	S	M	L
ابعاد ۱۲۸										
فقط سرC3D (بدون VGG و ترانسفورمر و پیش-بینی همه)			✓	✓	7.31	28.25	3.55	2.74	4.89	14.96
فقط سرC3D و W/L (بدون VGG و بدون ترانسفورمر و درون‌یابی)			✓	✓	19.67	42.50	21.54	12.55	16.45	26.31
Track Net (بدون ترانسفورمر)		✓	✓	✓	30.90	64.66	39.63	9.93	27.71	40.01
TrackNet* فقط نمای چپ	✓	✓	✓	✓	28.28	63.25	36.38	**13.97**	24.40	37.20
TrackNet* فقط نمای راست	✓	✓	✓	✓	26.74	57.52	33.30	5.0	21.96	33.08
TrackNet* فقط نمای جلو	✓	✓	✓	✓	34.40	66.15	49.40	8.83	34.70	42.27
TrackNet*	✓	✓	✓	✓	31.53	64.96	41.38	11.05	28.39	41.29
TrackNet* (آزمایش در طی ۲۰۰۰ ROI)	✓	✓	✓	✓	32.58	70.11	40.59	12.73	29.47	40.33
TrackNet* (آموزش W/flipped)	✓	✓	✓	✓	37.04	73.48	49.19	11.55	33.96	44.09
TrackNet* (آموزش w/skip)	✓	✓	✓	✓	**37.47**	**73.85**	**50.73**	11.31	**34.90**	**44.29**
افزایش ابعاد از ۱۲۸ به ۵۱۲										
Squash VGG 512 No C3D		✓		✓	35.27	71.01	46.71	11.97	31.74	42.54
Squash VGG 512 No C3D (آموزش w/skip)		✓		✓	39.79	**74.65**	57.53	19.94	36.44	45.99
Squash VGG 512 C3D 512		✓	✓	✓	40.12	73.95	56.73	**24.25**	36.81	**47.27**
Squash VGG 512 C3D 512 (آموزش w/skip)		✓	✓	✓	**40.45**	74.16	**58.78**	23.85	**37.72**	46.72

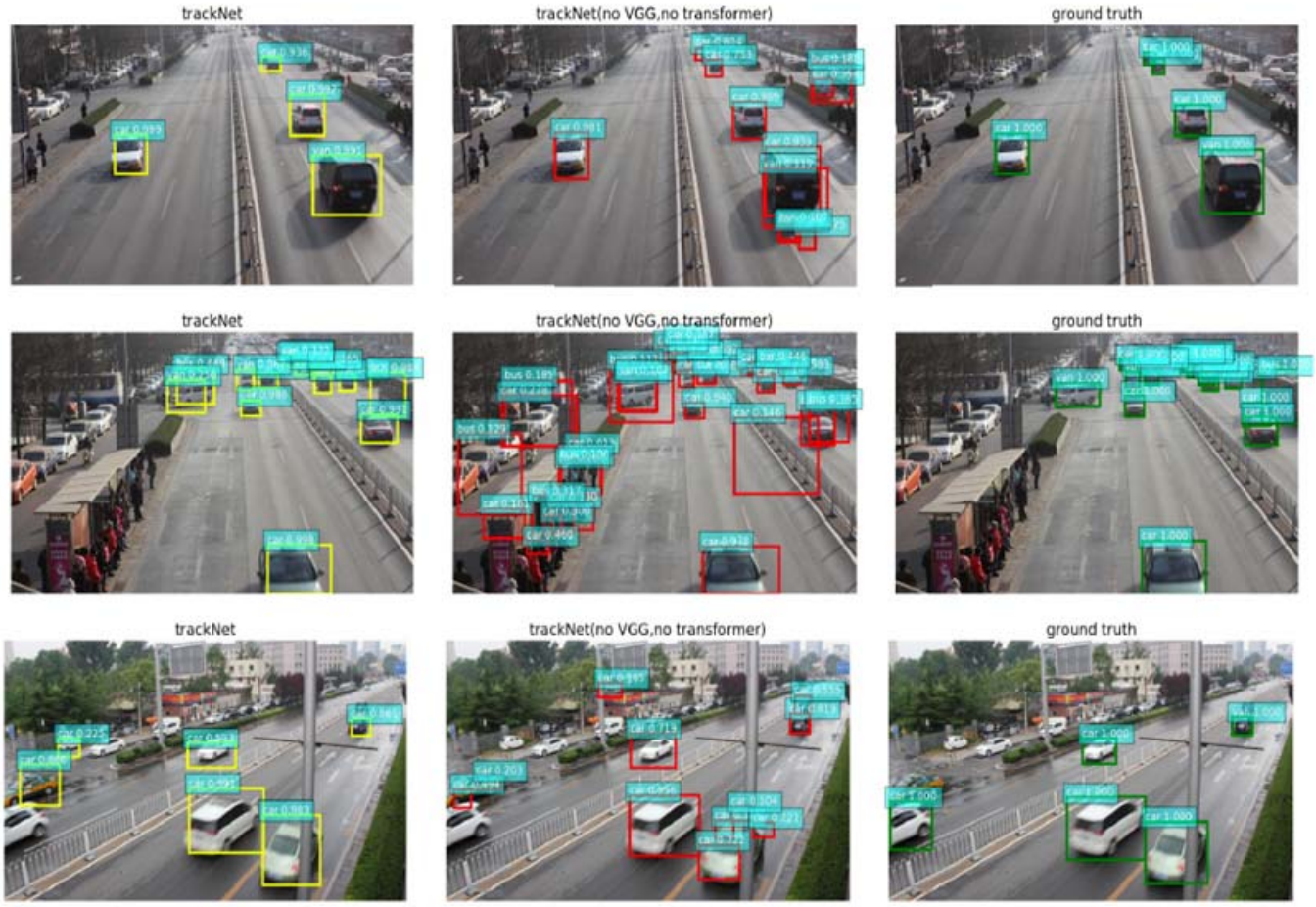
از جدول (۱) و جدول (۲)، مشاهده می شود که عملکردها پس از اتصال VGG تقویت شده‌است. درون‌یابی خطی (LP)، به راحتی به عنوان تنظیم یکنواختی ضمنی در طول مرحله TPN عمل کرده و عملکرد را حتی با پارامترهای کم‌تر، بهبود بخشیده‌است. با توجه به مجموعه داده، سطوح اختلاف برای زوایای دید مختلف متفاوت است. به عنوان مثال، مجموعه داده فرعی نمای جلویی، ساده‌ترین است. مدل پیشنهادی پس از اینکه داده‌های آموزشی با تغییر افقی افزایش داده‌شد، بهبود یافت. همچنین استفاده از آموزش"فریم‌های در حال پرش"، به افزایش عملکرد مدل،کمک شد. همچنین در جدول(۲)، میانگین نرخ‌های فراخوانی(AR%)، تحت تنظیمات مختلف گزارش شده‌است(یعنی آستانه‌های IoU و آستانه‌های حداکثر تشخیص در هر تصویر). از جدول(۱) و (۲) مشخص است که عملکردها پس از اتصال VGG تقویت شده‌است. درون‌یابی خطی (LP)، به راحتی به عنوان تنظیم یکنواختی ضمنی در طول مرحله TPN عمل کرده و عملکرد را حتی با پارامترهای کم‌تر، بهبود بخشیده‌است. جدول (۱) و (۲)، میانگین نرخ‌های دقت(AP) و میانگین فراخوانی(AR) را برای شرایط ارزیابی مختلف نشان می‌دهد. معیار برای برچسب زدن تشخیص به عنوان یک تطابق

جدول (۲): نرخ‌های میانگین فراخوانی(AR%) تحت تنظیمات مختلف

					AR @ IOU		AR num max detection			
	T_0(G)	VGG	C3D	LP	0.1:1	0.1	0.5	1	10	100
ابعاد ۱۲۸										
فقط سرC3D (بدون VGG و ترانسفورمر و پیش‌بینی همه)			✓	✓	43.59	27.64	10.46	8.34	13.34	13.76
فقط سرC3D و W/L (بدون VGG و بدون ترانسفورمر و درون‌یابی)			✓	✓	55.59	49.41	36.62	15.27	25.88	26.43
Track Net (بدون ترانسفورمر)		✓	✓	✓	71.47	66.92	48.64	25.31	36.46	36.85
TrackNet* فقط نمای چپ	✓	✓	✓	✓	68.50	64.18	45.53	26.88	35.08	35.08
TrackNet* فقط نمای راست	✓	✓	✓	✓	69.92	63.10	45.86	23.14	35.30	35.66
TrackNet* فقط نمای جلو	✓	✓	✓	✓	69.98	67.72	55.50	26.37	38.58	39.08
TrackNet*	✓	✓	✓	✓	71.50	67.01	49.61	26.11	36.95	37.26
TrackNet* (آزمایش در طی ۲۰۰۰ ROI)	✓	✓	✓	✓	**80.97**	74.23	51.24	27.89	39.71	40.14
TrackNet* (آموزش W/flipped)	✓	✓	✓	✓	78.83	75.81	57.10	28.85	41.81	42.34
TrackNet* (آموزش w/skip)	✓	✓	✓	✓	79.44	**76.45**	**58.29**	**29.41**	**42.25**	**42.80**
افزایش ابعاد از ۱۲۸ به ۵۱۲										
Squash VGG 512 No C3D		✓		✓	77.35	74.02	56.07	28.54	40.96	41.31
Squash VGG 512 No C3D (آموزش w/skip)		✓		✓	**81.48**	**78.82**	64.11	30.42	44.76	45.58
Squash VGG 512 C3D 512		✓	✓	✓	80.81	78.35	63.77	30.05	44.83	45.58
Squash VGG 512 C3D 512 (آموزش w/skip)		✓	✓	✓	80.59	78.47	**65.53**	**30.47**	**45.25**	**46.06**

از آنجا که شبکه ردیاب در حال حاضر، TrackNet های مختلفی را تولید می‌کند که در جدول (۱) و جدول (۲) نشان داده شده‌است و با تنظیمات مختلف، آزمایش شده‌است؛ هیچ ارتباطی نباید در یک GoP ،فریم‌های مختلف وجود داشته‌باشد لوله‌های پیشنهادی، به سادگی براساس 3D-IoT U در سراسر GoPها به هم متصل می‌شوند تا ردیابی‌های طولانی‌تر را فرموله کنند. پس از ارتباط و متصل شدن مسیرهای ایجاد شده، از R-CNN Faster به همراه الگوریتم SORT که یک ردیاب ساده، سریع و آنلاین است در کنار TrackNet، با استفاده از معیارهای ردیابی اشیا چندگانه، مقایسه می‌شوند. این عملکردها، تحت آستانه Th=0.8 و Th=1.0 در جدول (۳) گزارش شده‌است. در این جدول، معیار Recall ، نسبت داده‌های درست دسته‌بندی شده به تعداد کل داده‌هایی است که باید دسته‌بندی شوند. هرچه این معیار بزرگتر باشد، نشان‌دهنده این است که تعداد داده‌های کمتری، اشتباهاً دسته‌بندی شده است. عملکرد این معیار، فقط در کنار معیار دقت[1] قابل دفاع است؛ یعنی این معیار به تنهایی، نباید مورد استفاده قرار گیرد. همچنین منظور از معیار دقت، دقت ردیابی است که از نسبت تعداد پیش‌بینی‌های درست انجام شده، به تعداد کل پیش‌بینی‌ها بدست می‌آید. این معیار هم هرچه بیشتر باشد، بهتر است و نشان-دهنده دقت بالای ارزیابی است. معیار بعدی در جدول(۳)، معیار false positive rateاست که نشان‌دهنده نسبت پاسخ‌های مثبت غلط، به کل پاسخ‌های مثبت است. منظور از مثبت غلط، پاسخ‌هایی است که در واقعیت اشتباه است؛ اما در ارزیابی، در دسته پاسخ‌های صحیح قرار گرفته‌است. معیارهای MOTA و MOTP ، بترتیب بر پایه صحت و دقت در ردیابی‌های همزمان چند شی است.

[1] Precision



شکل (۱۳): مقایسه نتایج

صحت(A=Accuracy)، نشان‌دهنده تعداد پیش‌بینی‌های صحیح انجام شده بر تعداد کل پیش‌بینی های انجام‌شده است. دقت، نسبت تعداد پیش‌بینی‌های درست انجام شده، به تعداد کل پیش‌بینی‌ها است. در شکل(۱۳)، مقایسه نتایج در شرایط مختلف را مشاهده می‌کنید.در شکل(۱۴) مشاهده می‌شود که لوله باندینگ با هشت جعبه باندینگ متوالی به رنگ آبی، برای هشت فریم از GoP نشان‌داده شده‌است. جعبه باندینگ به رنگ زرد و مسیر حرکت مرکز جرم به رنگ قرمز است.

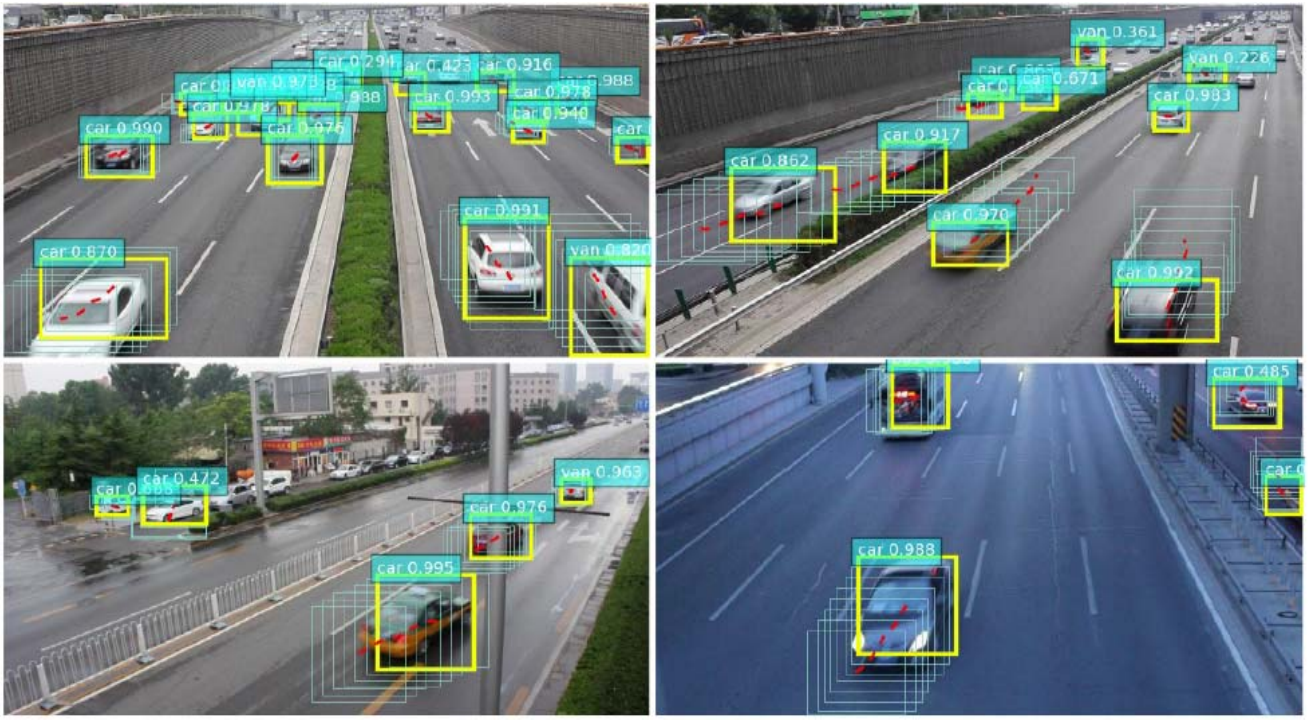
در جدول (۴) روش پیشنهادی این مقاله با سایر روش‌ها مقایسه شده است. همان‌طور که در جدول (۴) مشاهده می شود، روش پیشنهادی در نرخ دقت، نرخ فراخوانی، مدت زمان پردازش و تعداد فریم پردازش‌شده در هر ثانیه، بهبود قابل توجهی را داشته است. به طوری که دقت دقت این روش، ۹۸,۲ درصد و نرخ فراخوانی آن ۸۸,۶ درصد است. همچنین، مدت زمان پردازش، ۲۹ میلی‌ثانیه و تعداد فریم‌های پردازش‌شده، ۳۶ فریم در ثانیه است. در نتیجه، عملکرد روش پیشنهادی نسبت به سایر روشها بهبود قابل توجهی دارد.

جدول (۴): مقایسه روش پیشنهادی با سایر روشها

روش	نرخ دقت	نرخ فراخوانی	زمان پردازش هر فریم (ms)	تعداد فریم پردازش شده در یک ثانیه
[۱۵]	٪85	٪65	40	25
[۱۹]	٪84	٪72	29	34
[20]	٪64	٪53	333	3
[۲۱]	٪75	٪68	38	26
[۲۲]	٪63	٪51	35	28
پیشنهادی	**٪98.2**	**٪88.6**	**29**	**36**

جدول (۳): نتایج ردیابی با استفاده از معیارهای MOT

	Recall		Precisin		False Positive Rate		MOTA		MOTP	
	0.8	1.0	0.8	1.0	0.8	1.0	0.8	1.0	0.8	1.0
فقط نمای چپ	62.1	71.3	56.6	58.0	0.76	0.39	42.3	60.2	60.7	39.9
فقط نمای راست	60.4	71.7	52.5	68.0	0.88	0.10	57.7	69.0	61.7	35.2
فقط نمای جلو	70.7	72.6	80.0	82.7	0.66	0.29	67.4	71.1	63.9	50.0
TrackNet*	**87.4**	**88.6**	**97.3**	**98.2**	**0.27**	**0.09**	**76.9**	**88.9**	**92.1**	**84.2**



شکل (۱۴): نتیجه با لوله باندینگ

۵-نتیجه‌گیری

در این مقاله، یک شبکه ردیاب ارائه شد و نشان داده شد که می-تواند اشیا متعدد را در ویدئوها به‌طور مشترک با تولید لوله‌های محدودکننده، شناسایی و ردیابی کند. با استفاده از ویژگی‌های مکانی-زمانی استخراج‌شده توسط یک شبکه عصبی کانولوشن سه‌بعدی، VGG طرح‌های پیشنهادی لوله را ایجاد کرده و آن‌ها را طبقه‌بندی و موقعیت آن‌ها را اصلاح کرد. TrackNet شامل سه مرحله است؛ استخراج ویژگی و تبدیل مکانی، شبکه پیشنهاد لوله (TPN) و طبقه‌بندی و پالایش پس از TPN. همچنین چندین روش برای انجام پیشنهادات لوله و تنظیم رگرسیون بررسی شده‌است. شبکه ردیاب پیشنهادی با مجموعه داده ویدئویی UA-DETRAC آموزش داده شد و مورد آزمایش قرار گرفت.روش پیشنهادی پیشنهادی نرخ دقت و نرخ فراخوانی بالاتری همچنین مدت زمان پردازش کمتری در مقایسه با روشهای دیگر دارد . نرخ دقت روش پیشنهادی، ۹۸,۲ درصد است که از نظر دقت بهبود قابل ملاحظه-ای نسبت به سایر روش‌ها دارد.

۵- مراجع

[1] Wu, C.-J., et al. Machine learning at facebook: Understanding inference at the edge. in 2019 IEEE international symposium on high performance computer architecture (HPCA). 2019. IEEE.

[2] Suzuki, K., Overview of deep learning in medical imaging. Radiological physics and technology, 2017. 10(3): p. 257-273.

[3] Steffens, C., P.L.J. Drews, and S.S. Botelho. Deep learning based exposure correction for image exposure correction with application in computer vision for robotics. in 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE). 2018. IEEE.

[4] Wang, H., et al., Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle. Robotics and Autonomous Systems, 2017. 88 :p. 71-78.

[5] Redmon, J. and A. Farhadi, Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[6] Ren, S., et al., Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015. 28.

[7] Uijlings, J.R., et al., Selective search for object recognition. International journal of computer vision, 2013. 104(2): p. 154-171.

[8] Liu, W., et al. Ssd: Single shot multibox detector. in European conference on computer vision. ۲۰۱۶. Springer.

[9] Carnec, M., P. Le Callet, and D. Barba. Visual features for image quality assessment with reduced reference. in

IEEE International Conference on Image Processing 2005. 2005. IEEE.

[10] Wang, C., et al. Deep-lk for efficient adaptive object tracking. in 2018 IEEE International Conference on Robotics and Automation (ICRA). 2018. IEEE.

[11] Redmon, J. and A. Farhadi. YOLO9000: better, faster, stronger. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[12] Tran, D., et al. Learning spatiotemporal features with 3d convolutional networks. in Proceedings of the IEEE international conference on computer vision. 2015.

[13] Shou, Z., D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[14] Everingham, M., et al., The pascal visual object classes (voc) challenge. International journal of computer vision, 2010. 88(2): p. 303-338.

[15] Lin, T.-Y., et al. Microsoft coco: Common objects in context. in European conference on computer vision. 2014. Springer.

[16] Deng, J., et al. Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition. 2009. Ieee.

[17] Tang, S., et al., Object localization based on proposal fusion. IEEE Transactions on Multimedia, 2017. 19(9): p. 2105-2116.

[18] He, S., et al. "Visual tracking via locality sensitive histograms". in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. 2013. IEEE.

[19] Zhang, K., L. Zhang, and M.-H. Yang, "Realtime compressive tracking", in Computer Vision–ECCV 2012. Springer. p. 864-877

[20] Zhao, T. and R. Nevatia. "Tracking multiple humans in crowded environment". in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on IEEE.

[21] Hare, S., A. Saffari, and P.H. Torr. "Struck: Structured output tracking with kernels". In Computer Vision (ICCV), 2011 IEEE International Conference on. 2011. IEEE.

[22] Kalal, Z., J. Matas, and K. Mikolajczyk. Pn learning: "Bootstrapping binary classifiers by structural constraints". in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010. IEEE.

سید حمید خاتمی در سال ۱۳۹۸ مدرک کارشناسی و در سال ۱۴۰۱ مدرک کارشناسی‌ارشد خود را در رشته مهندسی برق، گرایش الکترونیک از دانشگاه بیرجند اخذ نمود. وی در حال حاضر، دانشجوی دکتری الکترونیک در دانشگاه بیرجند می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان پردازش تصویر و ویدئو، پردازش تصاویر پزشکی، یادگیری ماشین و یادگیری عمیق است.

رمضان هاونگی کارشناسی ارشد و دکترای خود را در رشته مهندسی برق کنترل بهترتیب در سالهای 1382و 1391از دانشگاه صنعتی خواجه نصیرالدین طوسی دریافت کرد. وی هم اکنون دانشیار گروه الکترونیک پردیس مهندسی دانشگاه بیرجند است. زمینه پژوهشی مورد علاقه ایشان ناوبری اینرسی و تلفیقی، تئوری تخمین، داده‌کاوی، الگوریتمهای تقریبی و محاسبات نرم است.