این مقاله در قالب نهایی آن در مجله
ماشین‌بینایی و پردازش‌تصویر چاپ خواهد شد

تاریخ انتشار برخط: اسفندماه ۱۴۰۲

مقاله مروری

http://jmvip.sinaweb.net
www.ismvip.ir

مروری بر روش‌های تشخیص جسم مبتنی بر شبکه‌های عصبی پیچشی: بررسی چالش‌ها و راهکارهای جدید

محمدحسین حمزه‌نژادی[۱]، حدیث محسنی[۲]

چکیده

تشخیص جسم یک مسئله مهم و کاربردی در زمینه بینایی ماشین است که شامل پیش‌بینی موقعیت اجسام در تصویر و دسته‌بندی آن‌ها براساس دسته‌های مشخص‌شده است. برای اینکه بتوان اجسام را با موقعیت در تصویر پیدا و شناسایی کرد، استخراج ویژگی‌های مناسب که قدرت تشخیص بین اجسام مشابه و مختلف را داشته باشند یک امر ضروری است. در سال‌های اخیر، شبکه‌های عصبی پیچشی (CNN) به دلیل توانایی خود در یادگیری خودکار ویژگی‌های بصری، به عنوان یک راهکار موثر برای تشخیص جسم مطرح شده‌اند. این مقاله سعی دارد که به اهمیت و چالش‌های پیش رو در این زمینه پرداخته و پس از معرفی مجموعه‌داده‌های مشهور، شبکه‌ها و مدل‌های مطرح تشخیص جسم را معرفی کند. سپس با معرفی معیارهای ارزیابی متداول در این حوزه، مدل‌های معرفی شده را مورد ارزیابی قرار دهد و راهکارها و نوآوری‌های اخیر در این زمینه را معرفی کند.

کلید واژه ها

تشخیص جسم، شبکه‌های عصبی پیچشی، بینایی ماشین

۱- مقدمه

انسان‌ها از سنین کم شروع به آموختن خصوصیات و ویژگی‌های های اجسام مختلف می‌کنند و می‌توانند آن‌ها را به راحتی تشخیص دهند. اما برخلاف انسان‌ها، آموزش ویژگی‌های اجسام متعلق به دسته‌های متفاوت و مکان‌یابی آن‌ها برای سیستم‌های بینایی ماشین یک امر چالش برانگیز است. در واقع، تشخیص و شناسایی جسم یکی از مسائل مهم در زمینه بینایی ماشین است و عبارت است از مشخص کردن مکان دقیق اجسام در یک تصویر (با استفاده از جعبه‌های مشخص کننده جسم) و شناسایي و دسته‌بندی آن‌ها. امروزه مسئله‌ی تشخیص جسم در حوزه‌های متفاوتی کاربرد دارد که از جمله‌ی آن‌ها می‌توانبه شناسایی تومورها و عارضه‌ها در حوزه‌ی تصاویر پزشکی [۱]، خودروهای خودران [۲]، امنیت و شناسایی تهدیدها [۳] اشاره کرد.

در حال حاضر، به دلیل در دسترس بودن مقدار زیادی داده، پردازنده‌های گرافیکی سریع‌تر و الگوریتم‌های بهتر، امکان استفاده از سیستم‌های بینایی ماشین برای شناسایی و طبق‌بندی اجسام در تصویر و با دقت بالا وجود دارد. با مکان‌یابی و شناسایی اجسام در تصویر، امکان تحقق اهدافی مانند شمارش تعداد اجسام در یک تصویر، ردیابی مکان دقیق و برچسب‌گذاری آنها میسر است. از طرف دیگر، یکی از چالش‌های موجود در طراحی سیستم‌های تشخیص جسم مبتنی بر الگوریتم‌های بینایی ماشین، مقاومت و کارآیی این سیستم‌ها در برابر تغییرات گوناگون نظیر چرخش، انتقال و تغییر در مقیاس است.

این مقاله در تیرماه ۱۴۰۲ دریافت، در مهرماه بازنگری و در آبان‌ماه پذیرفته شد.

[۱] کارشناس ارشد مهندسی کامپیوتر گرایش نرم افزار، دانشگاه شهید باهنر کرمان
رایانامه: mh_hamzenejad@eng.uk.ac.ir
[۲] دانشیار گروه هوش مصنوعی، بخش مهندسی کامپیوتر دانشگاه شهید باهنر کرمان
رایانامه: hmohseni@uk.ac.ir

نویسنده مسئول: حدیث محسنی

طی چندین دهه تحقیق در راستای هدف تشخیص و شناسایی جسم در تصویر، راهکارهای متعدد و متنوعی ارائه شده است که از ابتدایی‌ترین و موفق‌ترین آن‌ها می‌توان به روش‌هایی مانند ویولا جونز [۴]، HOG [۵] و DPM [۶] اشاره کرد. این روش‌ها با استخراج ویژگی‌های طراحی شده توسط محققان به طور دستی و پنجره‌های متحرک کار می‌کردند و به همین دلیل سرعت کم و عملکرد ضعیف در تشخیص اجسام در تصاویر پیچیده داشتند. با اشباع شدن راهکارهای کلاسیک و معرفی یادگیری عمیق و شبکه‌های عصبی پیچشی (CNN) ، چشم‌انداز ادراک بصری تغییر کرد. استفاده از CNN در مدل‌های دسته‌بندی تصاویر مانند AlexNet [۷] الهام بخش تحقیقات بیشتر در مورد کاربرد آن در بینایی ماشین شد و به توسعه و معرفی مدل‌های تشخیص جسم مبتنی بر یادگیری عمیق و شبکه‌هاي عصبي پيچشي هم‌چون RCNN [۸] و YOLO [۹] منجر شد.

شبکه‌ي عصبي پيچشي نوعی معماری شبکه برای الگوریتم‌هاییادگیری عمیق است و به طور خاص برای پردازش تصویر و کارهایی که شامل پردازش داده‌های پیکسلی است استفاده می‌شود. این شبکه‌ها با یادگیری ویژگی‌های ظاهری داده‌های ورودی، رابطه‌ی فضایی بین پیکسل‌ها را حفظ می‌کنند و ویژگی‌ها در کل تصویر یادگیری می‌شوند. این قابلیت در CNN ها باعث می‌شود که این شبکه‌ها گزینه‌ی مناسبی برای کارهای بینایی ماشین و مدل‌های تشخیص جسم باشند و عملکرد بهتری نسبت به راهکارهای تشخیص جسم کلاسیک داشته باشند.

با توجه به کارآيي روش‌هاي يادگيري عميق و رشد روزافزون استفاده از این روش‌ها در کاربردهاي مختلف در بینایی ماشین، در این مقاله با نگاهی تکاملی به مقالات مروری [۱۰] و [۱۱]، به مروری جامع بر راهکارها و مدل‌های تشخیص جسم مبتنی بر یادگیری عمیق و تکامل‌آن‌هامی‌پردازیم و عملکرد مدل‌های مختلف را مورد ارزیابی و تحلیل قرار می‌دهیم. براي اين منظور، در بخش ۲ مسئله و چالش‌های این زمینه مورد بحث قرار می‌گیرد، در بخش ۳ مجموعه داده‌های مطرح تشخیص جسم مرور می‌شوند و در بخش ۴ به توضیح شبکه‌های استخراج ویژگی در مدل‌های تشخیص جسم می‌پردازیم. سپس در بخش ۵ انواع مدل‌های تشخیص جسم مبتنی بر CNN و نحوه‌ي تکامل آن‌ها توضیح داده می‌شود. بخش‌های۶ و ۷ به ترتیب راهکارهای افزایش سرعت و نوآوری‌های اخیر در زمینه تشخیص جسم را مرور می‌کنند و در بخش ۸ تحقیقات اخیر در زمینه تشخیص جسم بررسی شده اند. بخش ۹ به بررسی عملکرد مدل‌های پایه مطرح در این زمینه می‌پردازد، بخش۱۰ به افق پیش رو براي بهبود روش‌هاي تشخیص جسم می‌پردازد و زمینه‌های باز برای ادامه‌ی تحقیق در این زمینه را مطرح می‌کند. در نهایت، نتیجه‌گیری و جمع‌بندی نهایی در مورد موضوع تشخیص و شناسایی اجسام انجام می‌شود.

۲-بیان مسئله

برخلاف دسته‌بندی تصاویر که در آن تنها یک برچسب برای هر تصویر محاسبه و تعیین می‌شود، در مسائل مربوط به تشخیص جسم، هدف تشخیص همه‌ی اجسام موجود در یک تصویر است، به طوری که مکان اجسام در تصویر با جعبه‌های مشخص کننده‌ی جسم نمایش داده شود و همه‌ی جعبه‌ها با توجه به دسته‌های تعریف شده دسته‌بندی شوند. در نتیجه، مسئله تشخیص جسم معمولاً یک مسئله آموزش نظارت شده[۱] محسوب می‌شود.

با وجود پیشرفت‌های قابل توجه در این زمینه، همچنان چالش‌های قابل توجهی در حل مسائل تشخیص جسم در دنیای واقعی وجود دارد. یکی از چالش‌های شاخص برای مدل‌های تشخیص جسم، نیاز به داده‌های آموزشی برچسب گذاری شده به تعداد زیاد و دارای تنوع کافی برای آموزش مناسب است. در صورت زیاد بودن تعداد دسته‌های مجموعه‌داده، نیاز شبکه به داده‌هایی با برچسب گذاری دقیق افزایش می‌یابد.

یکی دیگر از چالش‌ها در این زمینه تفاوت ظاهری نمونه‌های متعلق به یک دسته است که به دلیل ذات اجسام یا انسداد با اجسام دیگر و زاویه و شرایط روشنایی متفاوت به وجود می‌آید. این تفاوت در ظاهر اجسام متعلق به یک دسته‌ی یکسان، استخراج ویژگی‌های اجسام، مکان‌یابی و شناسایی آنها را با مشکل مواجه می‌کند. برای چالش مکان‌یابی اجسام استفاده از تکنیک‌های پیشرفته، مانند روش‌های مبتنی بر جعبه‌های محوری و مکانیسم توجه می‌توانند جزو راهکارهای احتمالی باشند.

همین‌طور باید توجه داشته که در مسائل دنیای واقعی، سرعت و بهینه بودن مدل بسیار اهمیت دارد. مدل‌های امروزی برای تولید نتایج به منابع محاسباتی بالایی نیاز دارند. با رایج شدن دستگاه‌های قابل حمل و روبات‌ها، توسعه‌ی مدل‌های تشخیص جسم کارآمد و ارزیابی آن‌ها از جهت پیچیدگی محاسباتی و سرعت ضروری است. در مواجهه با چالش سرعت می‌توان از طراحی مدل‌ها بر پایه‌ی معماری مدل‌های سبک وزن، تکنیک‌های فشرده‌سازی مدل و ارتقای سخت‌افزاری بهره برد.

از سویی دیگر، در سال‌های اخیر راهکارهای تشخیص جسم، خصوصاً مدل‌های مبتنی بر شبکه های عصبی پیچشی، سعی در حل چالش‌های گوناگون در این زمینه داشته‌اند. در این مقاله تلاش شده است که این راهکارها مورد بررسی قرار گیرند.

۳-مجموعه داده‌ها

همانطورکه در بخش قبل اشاره شد، در آموزش مدل‌های تشخیص جسم مبتنی بر CNN داده‌های استفاده شده برای آموزش شبکه بسیار مهم هستند. از زمان معرفی راهکارهای گوناگون برای تشخیص جسم، مجموعه‌داده‌های مختلفی برای نیازهای مختلف

[۱] Supervised

منتشر شده‌اند. به همین دلیل در این بخش به معرفی چندین مجموعه‌داده‌ی مطرح در زمینه‌ی تشخیص جسم می‌پردازیم.

۳-۱- مجموعه‌داده‌های عمومی

تعدادی از مجموعه‌داده‌های متداول در زمینه‌ی تشخیص جسم برای کاربردهای عمومی به وجود آمده‌اند. این مجموعه‌داده‌ها معمولاً دارای تصاویری از دسته‌های متعدد و متداول برگرفته از دنیای واقعی و زندگی روزمره هستند (مانند انسان‌ها، حیوانات،خودرو و...). همچنین اجسام موجود در تصاویر آن‌ها به دلیل تنوع دسته‌ها دارای ابعاد بسیار متفاوتی هستند که خود این موضوع یکی از چالش‌های موجود در تشخیص جسم است و باعث شده که مجموعه‌داده‌های تشخیص جسم عمومی به مجموعه‌داده‌های اصلی برای ارزیابی کلی مدل‌های تشخیص جسم تبدیل شوند.

یکی از اولین مجموعه‌داده‌های مطرح در تشخیص جسم مجموعه‌داده PASCAL VOCاست [۱۲]. این مجموعه‌داده دارای ۲۰ دسته از انواع اجسام متداول است و دارای ۱۱ هزار تصویر آموزشی و ۲۷ هزار جسم برچسب‌گذاری‌شده است. این مجموعه‌داده به غیر از تشخیص جسم دارای برچسب گذاری برای کاربردهایی مانند قطعه‌بندی تصویر[۱] و تشخیص فعالیت[۲] نیز هست. شکل (۱-الف) نمونه‌ای از یک تصویر برچسب گذاری شده از این مجموعه‌داده را نشان می‌دهد.

مجموعه‌داده MS-COCOیکی از چالش برانگیزترین مجموعه‌داده‌های موجود در زمینه‌ی تشخیص جسم است که در سال ۲۰۱۵ معرفی شد [۱۳]. این مجموعه‌داده دارای ۹۱ دسته از اجسام متداول است و بیش از ۲ میلیون نمونه تصویر دارد که به طور میانگین۵/۳ دسته در هر تصویر موجود است. علاوه بر این، این مجموعه داده شامل ۲۷/۷ نمونه در هر تصویر است که بیشتر از سایر مجموعه‌داده‌های مشابه است. MS-COCO شامل تصاویری از زوایای مختلف نیز می‌شود. شکل (۱-ب) یک تصویر از این مجموعه‌داده را نشان می‌دهد.

مجموعه داده Open-Images از ۲/۹ میلیون تصویر تشکیل شده است که با دسته‌بندی تصاویر، جعبه‌های مشخص کننده‌ی اجسام و اطلاعات تقسیم‌بندی برچسب‌گذاری شده‌اند. این مجموعه‌داده در سال ۲۰۱۷ معرفی شد و تاکنون شش بار به‌روزرسانی شده‌است. برای تشخیص اجسام، Open-Images دارای ۱۶ میلیون جعبه برای ۶۰۰ دسته در ۹/۱ میلیون تصویر است که آن را به بزرگترین مجموعه داده تشخیص جسم تبدیل می‌کند.
سازندگان آن برای انتخاب تصاویر پیچیده و متنوع دقت بسیاری داشتند به‌طوری‌که‌به‌طور میانگین در هر تصویر ۸/۳ دسته وجود

[۱]Image Segmentation
[۲]Activity Recognition

جدول (۱) : مقایسه مجموعه‌داده‌های مطرح عمومی در زمینه‌ی تشخیص جسم

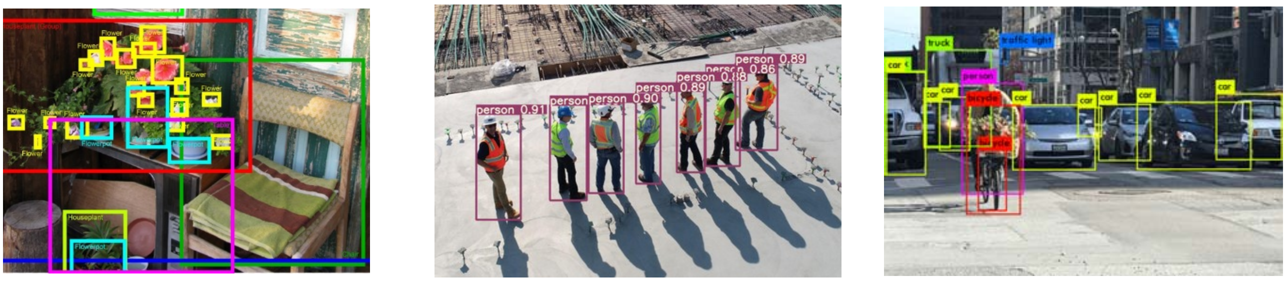
مجموعه‌داده	زمینه	دسته‌ها	تصاویر
[۱۲] PASCAL VOC	عمومی	۲۰	۵۷۱۷
[۱۳] MS-COCO	عمومی	۸۰	۱۱۸۲۸۷
[۱۴]Open Images	عمومی	۶۰۰	۱۷۴۳۰۴۲
[۱۵] CheXpert	پزشکی	۱۴	۲۲۴۳۱۶
[۱۶] MURA	پزشکی	۷	۴۰۰۰۰
[۱۷] DOTA	سنجش از دور	۱۸	۱۱۲۶۸
[۱۸] VisDrone	سنجش از دور	۱۰	۸۶۲۹
[۱۹] BDD100K	رانندگی خودکار	۱۰	۱۰۰۰۰۰

دارد.شکل (۱-ج) یک تصویر نمونه از این مجموعه‌داده را نشان می‌دهد.

جدول (۱) مشخصات مجموعه‌داده‌های ذکرشدهرا نشان می‌دهد. نکته قابل توجه در بررسی مجموعه‌داده‌ها این است که تعداد تصاویر برای دسته‌های مختلف به‌طور قابل توجهی متفاوت است. در اکثر این مجموعه‌داده‌ها، چند دسته‌ی پرتکرار وجود دارد و تعداد تصاویر در دیگر دسته‌ها افت بسیار چشمگیری در مقایسه با دسته‌های پرتکرار دارد. برای مثال، در مجموعه‌داده PASCAL VOCحدود ۱۳۷۷۵ تصویر وجود دارد که حاوی دسته "انسان" و ۲۸۲۹ تصویر وجود دارد که حاوی دسته "خودرو" هستند. تعداد تصاویر برای۱۸ کلاس باقیمانده در این مجموعه‌داده تقریباً به صورت خطی کاهش یافته و در نهایت به تنها ۵۵ تصویر دارای دسته‌ی "گوسفند" می‌رسد. به‌طور مشابه، برای مجموعه داده MS-COCO، دسته‌ی "انسان" دارای ۲۶۲۴۶۵ تصویر است و دسته‌ی بعدی یا همان دسته‌ی "خودرو" دارای۴۳۸۶۷ تصویر است. این روند نزولی برای تعداد تصاویر در دیگر دسته‌ها ادامه می‌یابد، تا جایی که تنها ۱۹۸ تصویر وجود دارد که حاوی اجسام با دسته‌ی "سشوار" باشند. این موضوع در مجموعه داده‌ی Open-Images نیز مشاهده می‌شود که در آن دسته‌ی "انسان" با ۳۷۸۰۷۷ تصویر بیشترین فراوانی را دارد و دسته‌ی "کاغذبر" تنها ۳ تصویر دارد.

این عدم تعادل در تعداد تصاویر در دسته‌های مختلف باعث ایجاد سوگیری[۳] در روند آموزش هر مدل تشخیص جسم می‌شود و هر مدل تشخیص جسم که بر روی این مجموعه تصاویر آموزش داده شود، به احتمال زیاد عملکرد تشخیصی بهتری را بر روی دسته‌های با تعداد تصاویر بیشتر نشان می‌دهد، در حالی که ممکن است در تشخیص اجسام در دسته‌های با تعداد تصاویر محدود، عملکرد ضعیف‌تری داشته باشد. بنابراین عدم تعادل در تعداد داده در دسته‌های مختلف یک چالش مهم برای تحقیقات بیشتر در این زمینه است.

[۳]Bias



الف　　ب　　ج

شکل ۱: مجموعه‌داده‌های الف) PASCAL-VOC [۱۲]، ب) COCO [۱۳] و ج)Open Images [۱۴].

۳-۲- مجموعه‌داده‌های تخصصی

تعدادی از مجموعه داده‌های تشخیص جسم به صورت تخصصی برای آموزش و ارزیابی مدل‌های تشخیص جسم در یک کاربرد خاص گردآوری شده‌اند. برای مثال، در حوزه پزشکی، مجموعه داده‌ی CheXpert [۱۵] یکی از مجموعه‌داده‌های مطرح است که شامل ۲۲۴۳۱۶ رادیوگرافی قفسه سینه از ۶۵۲۴۰ بیمار است و دارای ۱۴ دسته از گزارش‌های پزشکی بر روی این تصاویر است.

یکی دیگر از مجموعه‌داده‌های مطرح در این زمینه، مجموعه داده‌یMURA (رادیوگرافی اسکلتی عضلانی) است [۱۶] که حاوی بیش از ۴۰۰۰۰ تصویر اشعه ایکس استخوان با ۷ دسته از وضعیت‌های مختلف اسکلتی و عضلانی است که آن را برای توسعه‌ی مدل‌هایی برای تشخیص ناهنجاری‌ها در تصاویر رادیوگرافی ارزشمند می‌کند.

برای آموزش مدل‌های مخصوص خودروهای خودران، مجموعه‌داده BDD100K [۱۹] از مجموعه‌داده‌های مهم در این زمینه است و یک مجموعه داده رانندگی مستقل بسیار محبوب محسوب می‌شود. این مجموعه‌داده شامل ۱۰۰۰۰۰ ویدئو است و شامل برچسب گذاری برای اجسام موجود در جاده و خیابان است.

در زمینه سنجش از دور، مجموعه داده‌یDOTA، (مجموعه داده در مقیاس بزرگ برای تشخیص اشیا در تصاویر هوایی)[۱۷]یکی از شاخص‌ترین مجموعه‌داده‌های متشکل از تصاویر ماهواره‌ای است. DOTA شامل ۲ میلیون نمونه جسم در تصاویر هوایی ماهواره‌ای با تمرکز بر تشخیص اجسام کوچک در صحنه‌های پیچیده‌ی شهری و روستایی است. این مجموعه‌داده برای توسعه الگوریتم‌های تشخیص جسم برای تجزیه و تحلیل تصاویر هوایی، از جمله برنامه‌های کاربردی در کشاورزی، برنامه‌ریزی شهری و نظارت بر محیط‌زیست کاربرد دارد. همچنین مجموعه‌داده [۱۸VisDrone]یک مجموعه‌داده برای سنجش از دور است که از بیش از ۸۰۰۰ هزار تصویر هوایی گرفته شده با پهپاد تشکیل شده‌است و دارای ۱۰ دسته از اجسام شهری مانند خودرو و عابران پیاده است.

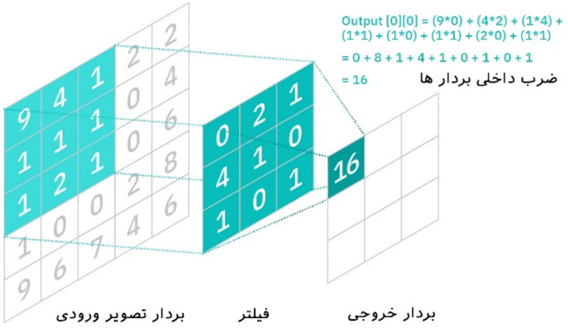
ذکر این نکته ضروری است که مشکل عدم تعادل در تصاویر در دسته‌های مختلف در مجموعه‌داده‌های تخصصی نیز وجود دارد. با این وجود، وجود مجموعه داده‌های تخصصی متنوع به محققان و متخصصان این امکان را می‌دهد تا الگوریتم‌ها و مدل‌های تشخیص جسم را برای کاربردهای مختلف توسعه دهند.

۴- گام اول در تشخیص جسم: استخراج ویژگی

در مدل‌های تشخیص جسم، بخش استخراج ویژگی (که به Backbone معروف است) اولین بخش یک مدل پس از ورودی است. استخراج ویژگی به فرآیند تجزیه و تحلیل یک تصویر ورودی و استخراج ویژگی‌های معنی دار از آن اشاره دارد که می‌تواند برای شناسایی و مکان یابی اجسام درون تصویر استفاده شود. این ویژگی‌ها معمولاً شامل لبه‌ها، گوشه‌ها و سایر الگوهایی است که مشخصه‌ی اجسام در تصویر است. شکل (۲) جایگاه بخش استخراج ویژگی در یک مدل تشخیص جسم را نشان می‌دهد. در مدل‌های تشخیص جسم بر پایه یادگیری عمیق، شبکه پیچشی یا همان CNN بخشی از شبکه‌ی اصلی است که وظیفه استخراج ویژگی‌ها از تصویر ورودی را بر عهده دارد. CNN معمولاً از چندین لایه پیچشی، ادغام و توابع فعال‌سازی تشکیل شده است که به تدریج ابعاد فضایی ورودی را کاهش داده و عمق نقشه‌های ویژگی را افزایش می‌دهند.به عبارت دیگر، فرآیند استخراج ویژگی در CNN شامل عبور تصویر ورودی از یک سری لایه‌های پیچشی است (شکل ۳) که هر لایه، ورودی را با مجموعه‌ای از فیلترهای قابل یادگیری برای تولید مجموعه‌ای از نقشه‌های ویژگی پردازش می‌کند.این نقشه‌های ویژگی، وجود و مکان الگوها یا ویژگی‌های خاصی مانند لبه‌ها، گوشه‌ها و ... را در تصویر ثبت می‌کنند.

بعد از هر یک یا چند لایه کانولوشن، معمولاً یک لایه‌ی ادغام برای کاهش ابعاد فضایی نقشه‌های ویژگی و استخراج ویژگی‌های انتزاعی‌تر از تصویر ورودی اعمال می‌شود. در نهایت، ویژگی‌های استخراج‌شده از یک یا چند لایه‌ی کاملاً متصل در شبکه عبور می‌کنند تا خروجی نهایی شبکه تولید شود که در کاربرد تشخیص جسم ، این خروجی معمولاً مجموعه‌ای از جعبه‌های مشخص‌کننده موقعیت جسم و احتمالات دسته‌های مختلف برای هر جسم شناسایی‌شده است. در نتیجه بخش استخراج ویژگی نقش مهمی در تعیین دقت و سرعت مدل‌های تشخیص جسم ایفا می‌کند.

شکل (۲): ساختار کلی مدل‌های تشخیص جسم مبتنی برCNN.



شکل (۳): عمل پیچش در لایه‌های پیچشی [۲۰].

۴-۱- AlexNet

شبکه AlexNet [۷] یک معماری مبتنی بر شبکه عصبی پیچشی برای طبقه بندی تصاویر است که برنده چالش تشخیص تصویری در مقیاس بزرگ (ILSVRC) بوده است. شبکه‌ی AlexNet از هشت لایه قابل یادگیری،یعنی پنج لایه پیچشی و سه لایه کاملاً متصل تشکیل شده است (شکل (۴)).آخرین لایه کاملاً متصل به یک طبقه‌بندی کننده N طرفه متصل است. این شبکه از چندین هسته‌ی پیچشی در سراسر شبکه برای به دست آوردن ویژگی‌های تصویر استفاده می‌کند. همچنین از عملگر dropout و تابع فعال‌سازی ReLU به ترتیب برای منظم‌سازی و همگرایی سریع‌تر آموزش استفاده می‌کند. در واقع، شبکه‌های عصبی پیچشی با به‌وجود آمدن AlexNet جان تازه‌ای گرفتند و به رویکردی رایج در پردازش داده‌های تصویری تبدیل شدند. این شبکه می‌تواند ویژگی‌ها را به‌طور خودکار از پیکسل‌های خام یاد بگیرد و به دلیل استفاده از تکنیکdropout بیش برازش را کاهش می‌دهد.

۴-۲- VGG

سیمونیان و همکارانش در [۲۲]و با دیدگاهی متفاوت نسبت به Alexnet ، تأثیرات عمق شبکه را بر روی دقت آن مورد بررسی قرار دادند. آنها شبکه یVGG[۱] را پیشنهاد کردند که ایده‌ی اصلی این شبکه استفاده از فیلترهای پیچشی کوچک برای ساخت شبکه‌هایی با عمق‌های مختلف بود. این کار پارامترهای شبکه را به شدت کاهش می‌دهد و در نتیجه، شبکه زودتر همگرا می‌شود. تحقیق انجام شده برای شبکه VGG نشان داد که چگونه معماری شبکه عمیق (۱۶-۱۹ لایه)می‌تواند برای انجام دسته‌بندی و محلی سازی با دقت بالاتر استفاده شود. در واقع، VGG با افزودن پشته‌ای از لایه‌های پیچشی به سه لایه‌ی کاملاً متصل و به دنبال آن یک لایهsoftmaxایجاد شده‌است که تعداد لایه‌های پیچشی می‌تواند از ۸ تا ۱۶ متغیر باشد.

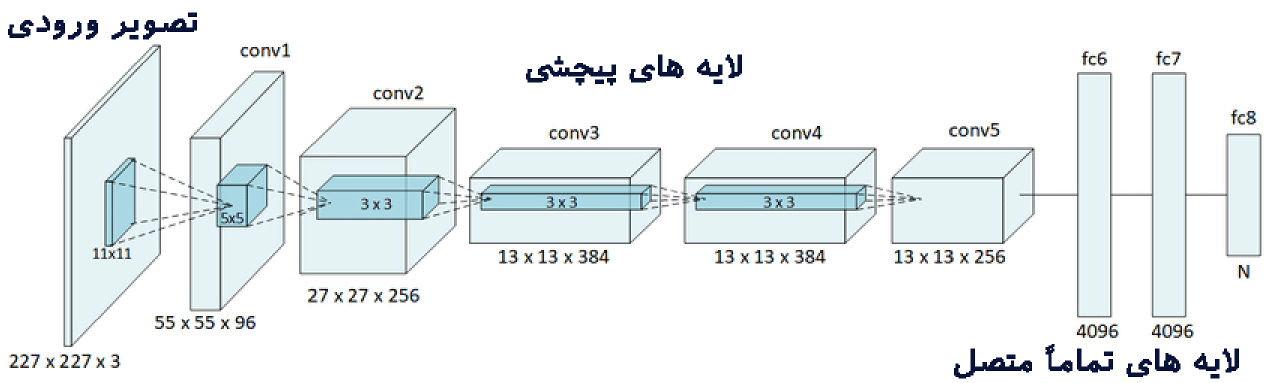
به دلیل اینکهVGG را می توان به راحتی با کتابخانه های معروف یادگیری ماشینی پیاده سازی کرد و آموزش داد، این شبکه خیلی زود به یکی از پرکاربردترین شبکه‌های استخراج ویژگی برای طبقه‌بندی تصویر و مدل‌های تشخیص جسم تبدیل شد. شکل (۵) معماری شبکه VGG را نشان می‌دهد.

۴-۳-GoogLeNet

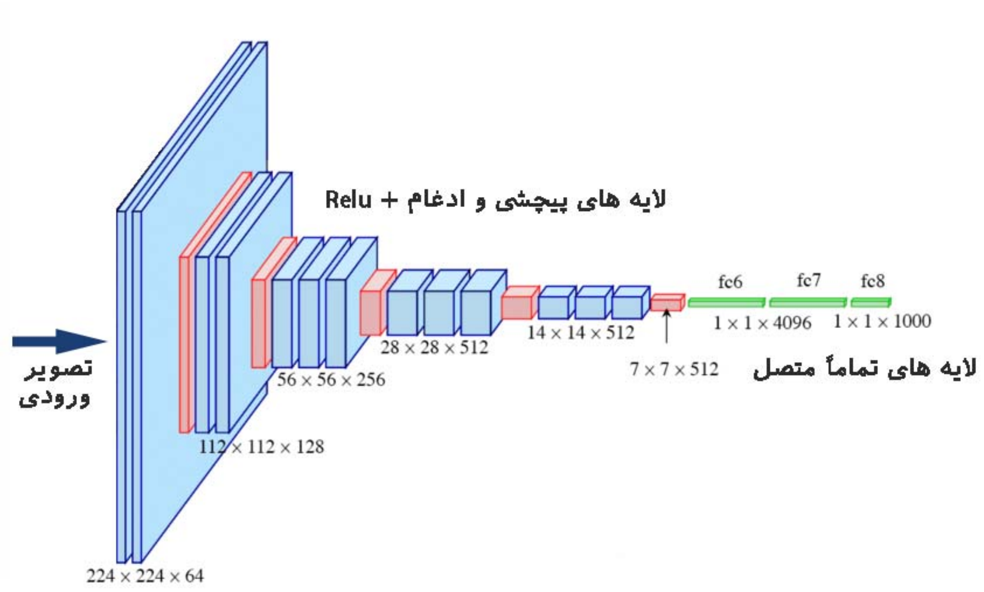
با وجود اینکه شبکه‌های استخراج ویژگی به سمت شبکه‌های سریعتر و دقیق‌تر پیش می‌رفتند، امکان استفاده از آن‌ها در برنامه‌های کاربردی دنیای واقعی به دلیل نیاز به منابع محاسباتی زیاد، به صورت گسترده میسر نبود. در واقع، زمانی که شبکه‌ها برای عملکرد بهتر از افزایش تعداد لایه‌ها و مقیاس استفاده می‌کنند، هزینه‌ی محاسبات در آن‌ها به طور تصاعدی افزایش می‌یابد.در تحقیق انجام شده در [۲۳] که منجر به معرفی شبکه GoogLeNet شد، هدر رفتن محاسبات در شبکه دلیل اصلی این مشکل بیان شده‌است. مدل‌های بزرگتر دارای تعداد زیادی پارامتر هستند و تمایل دارند داده‌ها را بیش از حد برازش دهند. در نتیجه، برای حل این مشکلات، GoogLeNet با استفاده از معماری متصل به صورت محلی به‌جای معماری کاملاً متصل پیشنهاد شده‌است.

بنابراین GoogLeNet یک شبکه‌ی عمیق ۲۲ لایه است که با توالی چندین ماژول Inception ساخته شده است.همان‌طور که در شکل (۶) مشخص است،ماژول‌هایInception شبکه‌هایی هستند که دارای فیلترهای پیچشی با چند اندازه‌ی مختلف در یک سطح هستند. نقشه‌های ویژگی ورودی از این فیلترها عبور می‌کنند و به لایه‌ی بعدی متصل می‌شوند. این شبکه همچنین دارای طبقه‌بندی‌کننده‌های کمکی در لایه‌های میانی برای کمک به تنظیم و انتشار گرادیان است. GoogLeNet نشان داد که چگونه استفاده کارآمد از بلوک‌های محاسباتی می‌تواند در عین پیچیدگی محاسباتی کمتر، دارای عملکرد همتراز با شبکه‌های با پارامتر زیاد باشد.

[۱] Visual Geometry Group



شکل (۴): ساختار کلی AlexNet [۲۱].



شکل (۵): معماری شبکه استخراج ویژگی VGG [۲۲].

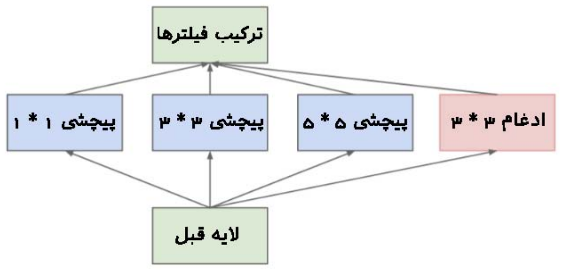
اگر چه شبکه GoogLeNet سریعتر از سایر مدل‌ها در زمان معرفی خودش بود، نسخه‌های به‌روزرسانی‌شده از ماژول Inception نیز عملکرد آن را بهبود بخشیدند.

۴-۴-ResNet

در [۲۴] کیمینگ و همکارانش شبکه استخراج ویژگیResNet را معرفی کردند و نشان دادند که چگونه با افزایش عمق در شبکه‌های عصبی پیچشی، دقت در این شبکه‌ها ابتدا اشباع شده و سپس به سرعت کاهش می‌یابد. آنها استفاده از بلوک‌های Residual را برای اتصال لایه‌های پیچشی و افزایش بهینگی پیشنهاد کردند، این لایه‌هایک اتصال پرش بین لایه‌های پیچشی اضافه می‌کنند. این اتصال تنها بین ورودی و خروجی لایه‌های پیچشی اضافه می‌شود و پارامتر یا پیچیدگی محاسباتی اضافی به شبکه اضافه نمی‌کند.شبکه ResNet معمولاً ۳۴ لایه دارد و اساساً دارای فیلترهای پیچشی بزرگ (۷×۷) است و به دنبال آنها دارای ماژول‌هایBottleneck و در نهایت یک لایه کاملاً متصل است. لایه‌هایBottleneck دارای۳ لایه پیچشی (۳×۱، ۱×۱، ۳×۳) هستند.

همچنین نویسندگان ResNet ثابت کردند که شبکه‌ی۱۶ لایه VGG پیچیدگی محاسباتی بالاتر و دقت کمتری نسبت به شبکه‌ی پیچشی عمیق ResNet با ۱۰۱ و ۱۵۲ لایه دارد. در مقاله‌ای دیگر، نویسندگان ResnetV2 [۲۵] را پیشنهاد کردند که از نرمال‌سازی دسته‌ای و لایه ReLU در بلوک‌ها استفاده می‌کند، قدرت تعمیم بیشتری دارد و آموزش آن آسان‌تر است.

به دلیل اینکه این شبکه می تواند توابع پیچیده و غیر خطی را با پارامترهای کمتر یاد بگیرد، شبکه ResNet به طور گسترده‌ای به عنوان شبکه استخراج ویژگی در مدل‌های تشخیص جسم استفاده می‌شود و اصول اصلی آن الهام بخش ظهور شبکه‌های دیگر بوده‌است. شکل (۷) معماری کلی ResNet را نشان می‌دهد.



شکل (۶): روند کلی ماژول Inception در شبکه GoogLeNet

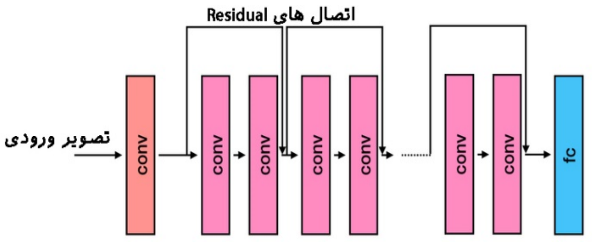
۴-۵- EfficientNet

در تحقیق انجام شده در [۲۶]، نویسندگان به طور سیستماتیک افزایش مقیاس شبکه و اثرات آن بر عملکرد مدل‌های پیچشی را مورد مطالعه قرار داده و نشان دادند که چگونه تغییر پارامترهای یک شبکه مانند عمق، عرض و اندازه‌ی ورودی می‌تواند بر دقت آن تأثیرگذار باشد. گرچه افزایش عمق یک شبکه می‌تواند به یادگیری ویژگی‌های غنی‌تر و پیچیده‌تر کمک کند، اما از طرف دیگر می‌تواند منجر به پدیده‌ی ناپدید شدن گرادیان شود. به‌طور مشابه، مقیاس‌بندی عرض شبکه، یادگیری ویژگی‌های زیردسته‌ها را آسان‌تر می‌کند، اما در دستیابی به ویژگی‌های سطح بالا مشکل دارد. راهکار پیشنهادی در EfficientNet استفاده از یک ضریب ترکیبی است که می‌تواند هر سه پارامتر عمق، عرض و اندازه ورودی شبکه را به طور هماهنگ مقیاس‌بندی کند. هر یک از این سه پارامتر دارای یک ثابت مرتبط است که با انجام جستجو بین مقادیر مختلف بر رویک شبکه پایه محاسبه می‌شود. این راهکار در مجموع دقت مدل و میزان محاسبات را بهبود می‌بخشد. در نتیجه، EfficientNetیک شبکه ساده و کارآمد است که با وجود ابعاد کوچکتر، در زمان ارائه از نظر دقت و سرعت بهتر از مدل‌های موجود بوده است . این شبکه‌ی استخراج ویژگی با افزایش چشمگیر در بهره‌وری، رویکرد جدیدی را در طراحی شبکه‌های کارآمد ارائه کرد.

۴-۶- CSPNet

اگرچه شبکه‌های عصبی پیچشی ذکر شده نتایج خوبی در دستیابی به دقت بالا در کاربردهای بینایی کامپیوتری از خود نشان دادند، با این حال، اکثر این شبکه‌ها نیاز به منابع محاسباتی زیاد دارند. در [۲۷] ونگ و همکارانش نشان دادند که محاسبات استنتاجی سنگین را می‌توان با کاهش اطلاعات گرادیان تکراری در شبکه کاهش داد. آنها CSPNet را پیشنهاد کردند که مسیرهای مختلفی را برای جریان گرادیان در شبکه ایجاد می‌کند. CSPNet نقشه‌های ویژگی را در لایه پایه به دو قسمت جدا می‌کند. یک قسمت از بلوک شبکه پیچشی جزئی عبور داده می‌شود، در حالی که قسمت دیگر با خروجی‌های آن در مرحله بعدی ترکیب می‌شود.

این ایده باعث کاهش تعداد پارامترها، افزایش استفاده از واحدهای محاسباتی و کاهش استفاده از حافظه می‌شود، پیاده سازی آن آسان است و قابلیت اجرا در معماری‌های دیگر را نیز دارد.تحقیقات نشان داده است که استفاده از CSPNet در شبکه‌های دیگر محاسبات را از ۱۰ تا ۲۰ درصد کاهش می‌دهد، در حالی که دقت ثابت مانده یا بهبود می‌یابد. همچنین هزینه‌ی حافظه و گلوگاه محاسباتی نیز با این روش به میزان قابل توجهی کاهش می‌یابد[۲۷].



شکل (۷): معماری کلی شبکه ResNet

از آنجایی که CSPNet می تواند صحنه های بزرگ و پیچیده را پردازش کند،این شبکه در بسیاری از مدل‌های تشخیص جسم پیشرفته مورد استفاده قرار گرفته‌است، در حالی که برای دستگاه‌های موبایل و قابل حمل با امکانات پردازشی محدود هم از این شبکه استفاده می‌شود.

جدول (۲) شبکه‌های بررسی شده برای استخراج ویژگی را به ترتیب پیدایش و برحسب معیارهای مختلف مانند تعداد لایه‌ها و پارامترهای کلی شبکه، دقت و میزان محاسبات مورد نیاز نشان می‌دهد. با بررسی این جدول مشخص می‌شود که اگرچه شبکه‌ی GoogLeNet سریعترین شبکه به لحاظ سرعت است و تعداد پارامترهای آن کمتر است، اما شبکه‌هایی مانند EfficientNet در کنار داشتن حجم محاسبات قابل قبول، دقت بسیار بالاتری دارند. پس از بررسی شبکه‌های متداول در استخراج ویژگی از تصاویر، در قسمت بعد، با فرض در دست داشتن ویژگی‌های استخراج شده از تصاویر، مدل‌های تشخیص جسم مبتنی بر شبکه‌های CNN دسته‌بندی و مورد بررسی قرار می‌گیرند.

۵- بررسی مدل‌های تشخیص جسم مبتنی بر شبکه‌های پیچشی

پیشرفت روز افزون شبکه‌هایCNN منجر به افزایش تحقیقات در این زمینه و معرفی مدل‌های تشخیص جسم مبتنی بر این شبکه‌ها شده‌است. مدل‌های تشخیص جسم مبتنی بر CNN ، برخلاف راهکارهای کلاسیک، ویژگی‌های اجسام را به طور نظارت شده یاد می‌گیرند و تعمیم پذیری بیشتری دارند. مدل‌های تشخیص جسم بر پایهCNN به دو دسته‌ی کلید و مرحله‌ای و یک مرحله‌ایتقسیم می‌شوند. مدل‌های دو مرحله‌ای ابتدا موقعیت اجسام را محاسبه کرده و سپس در مرحله دوم دسته‌ی مربوط به اجسام یافت شده را مشخص می‌کنند، در حالی که مدل‌هاییک مرحله‌ای این دو کار را

جدول ۲: مقایسه شبکه‌های استخراج ویژگی

مدل	سال	تعداد لایه‌ها	تعداد پارامترها (میلیون)	دقت Top-1	BFLOPs
AlexNet [۷]	۲۰۱۲	۷	۶۲/۴	% ۶۳/۳	۱/۵
VGG-16 [۲۲]	۲۰۱۴	۱۶	۱۳۸/۴	% ۷۳	۱۵/۵
GoogLeNet [۲۳]	۲۰۱۴	۲۲	۶/۷	% ۷۰/۲	۱/۶
ResNet-50 [۲۴]	۲۰۱۵	۵۰	۲۵/۶	% ۷۶	۸,۳
CSPResNet-50 [۲۷]	۲۰۱۹	۵۹	۲۰/۵	% ۷۸/۲	۹,۷
EfficientNet [۲۶]	۲۰۱۹	۱۶۰	۱۹	% ۸۳	۴/۲

در یک مرحله و به صورت توام انجام می‌دهند. شکل (۸) دسته‌بندی کلی راهکارهای تشخیص جسم و مدل‌های مهم در این زمینه را نشان می‌دهد. همچنین در شکل (۹) روند ظهور و پیشرفت روش‌ها بر اساس سال و رویکرد روش نشان‌داده شده‌است.

۵-۱- مدل‌های دو مرحله‌ای

همان‌طور که پیشتر ذکر شد، در این دسته از مدل‌ها ابتدا موقعیت اجسام در تصویر مشخص شده و سپس در مرحله دوم دسته بندی مربوط به اجسام یافت شده انجام می‌شود.

۵-۱-۱- RCNN

در سال ۲۰۱۴ یکی از نخستین مدل‌های تشخیص جسم مبتنی بر شبکه‌های پیچشی توسط گیرشیک و همکارانش معرفی شد. آن‌ها با ارائه راهکاری برای یافتن نواحی دارای جسم در تصویر با استفاده از ویژگی‌هایCNN، شبکه عصبی پیچشی ناحیه محور (RCNN) را ارائه کردند [۸].

شکل(۱۰) ساختار کلی مدل RCNN را نشان می‌دهد.این مدل در مرحله‌ی نخست با استخراج مجموعه‌ای از ناحیه‌های پیشنهادی (جعبه‌های کاندید مشخص کننده جسم) و با جستجوی انتخابی شروع می‌شود. سپس هر پیشنهاد به یک تصویر با اندازه ثابت تغییر مقیاس داده و به یک شبکه CNN برای استخراج ویژگی‌ها داده می‌شود. در نهایت در مرحله‌ی دوم، طبقه‌بندی کننده ماشین بردار پشتیبانی (SVM) ، برای تشخیص دسته‌بندی‌های اجسام استفاده می‌شود. مدل RCNN نسبت به راهکارهای تشخیص جسم کلاسیک مزایای بسیاری دارد. این مدل می تواند در مجموعه داده‌ها و وظایف مختلف تشخیص جسم به دقت بالایی دست یابد و می‌توانند با استفاده از الگوریتم جستجوی انتخابی برای تولید نواحی پیشنهادی، اجسام با اندازه‌ها و اشکال مختلف را پردازش کند.اگرچه الگوریتم RCNN در زمان خود نسبت به روش‌های سنتی پیشرفت بسیاری داشت، اما مشکلاتی هم داشت. یکی از اصلی‌ترین معایب آن، محاسبات اضافی برای استخراج ویژگی‌ها در تعداد زیادی از ناحیه‌های پیشنهادی با همپوشانی زیاد در یک تصویر است که باعث کاهش سرعت تشخیص می‌شود. در سال‌های بعد، مدل‌های جدیدی بر پایهRCNN با عملکرد بهبود یافته ارائه شدند.

۵-۱-۲- SPPNet

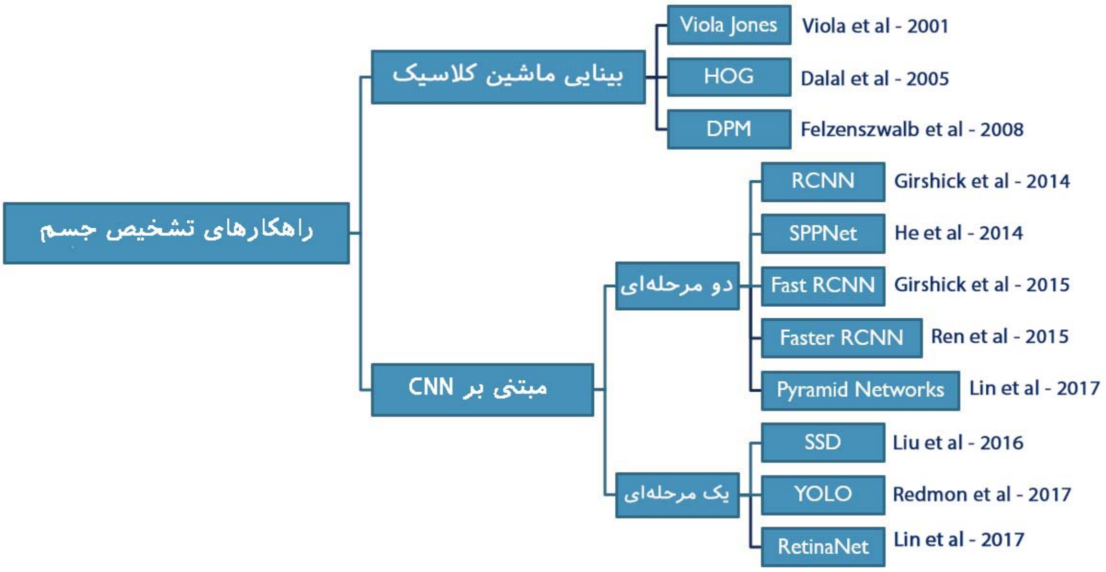
در سال ۲۰۱۴، شبکه ادغام هرمی فضایی (SPPNet) توسط هی و همکارانش معرفی شد [۲۸]. معماریSPPNet در شکل (۱۱) نشان داده شده است. اگرچه مدل‌های CNN قبلی برای پردازش، نیاز به تصاویر با اندازه ثابت داشتند، SPPNet یک لایه ادغام هرمی فضایی (SPP) را معرفی کرد که به مدل CNN اجازه می‌دهد تا بدون توجه به اندازه نواحی پیشنهادی و بدون تغییر اندازه تصویر، توالی‌ای با طول ثابت تولید کند. در هنگام انجام تشخیص جسم با استفاده از SPPNet ، نقشه‌های ویژگی فقط یک بار در کل تصویر محاسبه می‌شوند و برای مناطق دلخواه، توالی‌هایی با طول ثابت با استفاده از آشکارسازهای آموزش‌دیده تولید می‌شوند که از محاسبه ویژگی‌های پیچشی اجتناب می‌کنند. SPPNet عملکرد بسیار سریعتری نسبت به RCNN دارد، بدون اینکه دقت تشخیص کاهش یابد. اما این مدل هنوز دارای مشکلاتی است که از جمله‌ی آنها می‌توان به آموزش چند مرحله‌ای و نادیده گرفتن لایه‌های قبلی در تنظیم لایه‌های کاملاً متصل اشاره کرد.

۵-۱-۳- Fast-RCNN

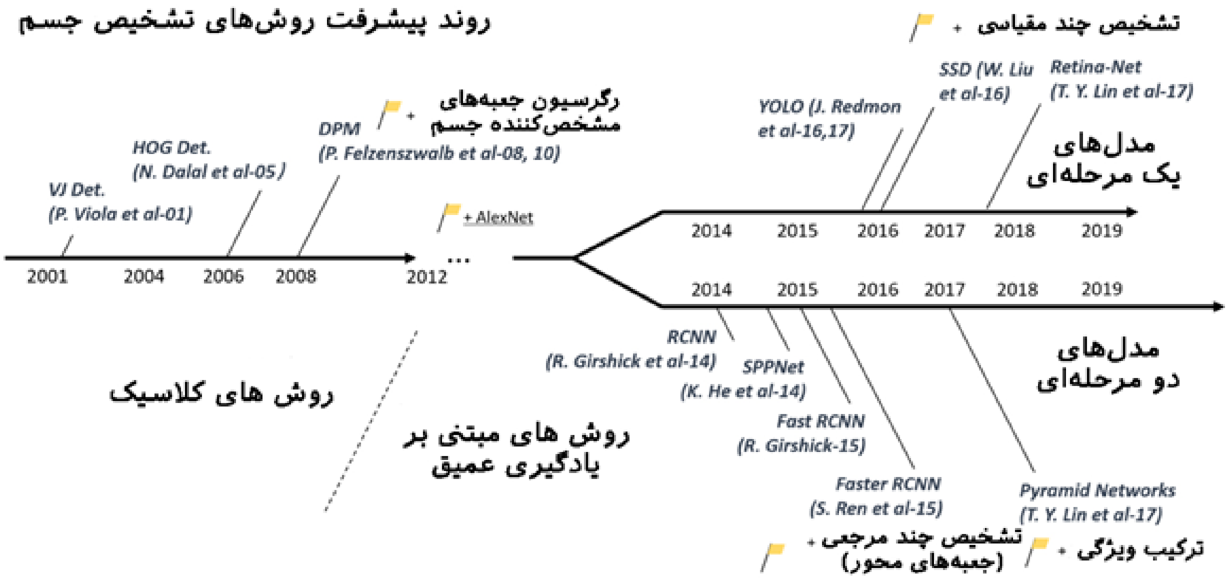
در سال ۲۰۱۵، گیرشیک مدل Fast-RCNN را معرفی کرد که ادغام و بهبودی بر RCNN و SPPNet است [۲۹]. این مدل امکان آموزش همزمان یک آشکارساز و رگرسیون برای تعیین جعبه محدود کننده جسم را با استفاده از یک شبکه یکسان فراهم می‌کند. گرچه این مدل از ادغام مزایایRCNN و SPPNet بهره می‌برد، اما سرعت تشخیص در شبکه کماکان توسط فرآیند تشخیص ناحیه‌های پیشنهادی محدود است. با این حال، Fast-RCNN دقت بهتری نسبت به RCNN داشته و بسیار سریعتر از آن عمل می‌کند.

۵-۱-۴- Faster-RCNN

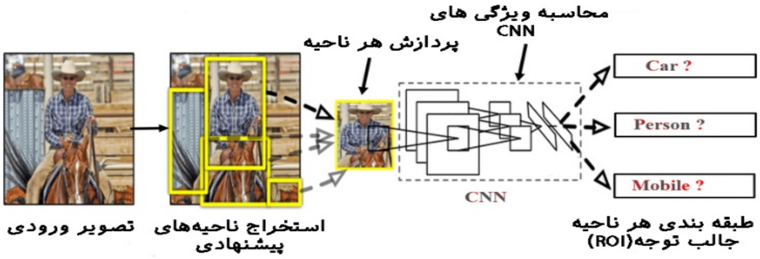
اندکی پس از معرفیFast-RCNN ، مدل تشخیص جسم Faster-RCNN توسط رن و همکاران پیشنهاد شد [۳۰]. این مدل اولین مدل تشخیص جسم بر اساس CNN بود که دارای سرعت تشخیص نزدیک به بی‌درنگ بود. ایده اصلی Faster-RCNN، معرفی شبکه پیشنهاد منطقه (RPN) است که یافتن ناحیه‌های پیشنهادی با هزینه محاسباتی کم را امکان‌پذیر می‌کند.



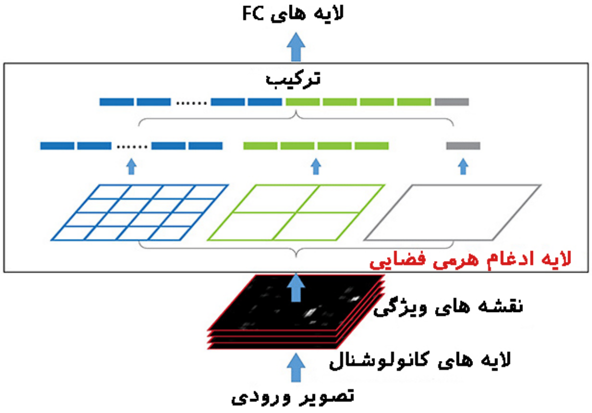
شکل (۸): دسته‌بندی کلی راهکارهای تشخیص جسم و مدل‌های مهم در این زمینه



شکل (۹): روند ظهور و پیشرفت روش‌های تشخیص جسم بر اساس سال و رویکرد روش



شکل (۱۰): ساختار کلی مدل تشخیص جسمRCNN



شکل (۱۱): ساختار SPPNet [۲۸].

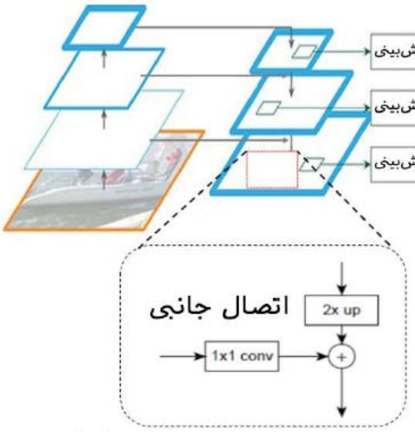
RPNیک شبکه پیچشی است که به طور همزمان مرزهای جسم و احتمال جسم بودن را در هر موقعیت پیش‌بینی می‌کند.مزیت اصلیFaster-RCNN نسبت به Fast-RCNN این است که به دلیل استفاده از RPN این مدل می تواند نواحی پیشنهادی را به صورت انتها به انتها با شبکه تشخیص بیاموزد، که دقت و پایداری مدل را بهبود می بخشد.اگرچه Faster-RCNN از گلوگاه سرعت Fast-RCNN و RCNN عبور می‌کند، اما در مراحل تشخیص ویژگی و تعیین جعبه برای اجسام دارای افزونگی محاسباتی است.

در Faster-RCNN، با افزودن طبقه‌بندی‌کننده دامنه سطح تصویر و سطح نمونه و مؤلفه‌های خطای سازگاری، مشکل جبران دامنه ناشی از توزیع ناسازگار بین نمونه‌های آموزشی و نمونه‌های واقعی حل شده است.همچنین، شبکه RPN با استفاده از آموزش چند مقیاسی بهبود پیدا کرده است.

در تحقیق دیگری [۳۲]، مدلی جدید به نام Sparse-RCNNبرای شناسایی اجسام در تصاویر معرفی شده است. این مدل یک روش کاملاً پراکنده برای شناسایی اجسام در تصاویر است که از محاسبات اضافی برای یافتن اجسام کاندیدا و تخصیص برچسب‌های چند به یک کاملاً اجتناب می‌کند. در این روش، یک مجموعه‌ی تنک و ثابت از پیشنهادهای جعبه‌های اجسام برای انجام طبقه‌بندی و مکان‌یابی اجسام یادگرفته می‌شود. این مدل با کاهش محاسبات اضافی و استفاده از روش‌های بهینه‌سازی جدید به دقت و سرعت مناسبی نسبت به مدل‌های مشابه دست یافته است در این مدل پیش‌بینی نهایی بدون NMS انجام می‌شود.

۵-۱-۵ -FPN

در سال ۲۰۱۷، لین و همکاران [۳۳] شبکه هرمی ویژگی (FPN) را معرفی کردند. پیش از FPN ، بیشتر مدل‌های تشخیص جسم مبتنی بر یادگیری عمیق، تشخیص را فقط در لایه آخر شبکه انجام می‌دادند. اگرچه ویژگی‌های لایه‌های عمیق‌تریکCNN برای انجام دسته‌بندی مفید است، اما برای یافتن مکان اجسام در تصویر مفید نیست. برای این منظور، در FPNیک معماری بالا به پایین با اتصالات جانبی برای استخراج ویژگی‌های معنایی سطح بالا در همه مقیاس‌ها توسعه داده شده‌است. شکل (۱۲) ساختار کلی این شبکه را نشان می‌دهد. از آنجایی که یکCNN به طور طبیعی یک هرم ویژگی را از طریق انتشار به جلو تشکیل می‌دهد،FPN پیشرفت‌های بزرگی را در شناسایی اجسام با مقیاس‌های مختلف نشان می‌دهد. مزیتFPN این است که می تواند عملکرد تشخیص اجسام را که شامل اجسام در مقیاس های مختلف می شود، بهبود بخشد و با شبکه‌های استخراج ویژگی دیگر به خوبی ترکیب شود. در حال حاضر، FPN و شبکه‌های هرمی به یک بخش اساسی برای بسیاری از مدل‌های تشخیص جسم نوین مبتنی بر یادگیری عمیق تبدیل شده‌اند.



شکل (۱۲): معماری شبکه FPN [۳۳].

۲-۵- مدل‌های یک مرحله‌ای

مدل‌های تشخیص جسم یک مرحله‌ای بر خلاف مدل‌های دو مرحله‌ای، بدون نیاز به مرحله‌ی پیشنهاد ناحیه‌های پیش‌بینی و جعبه‌های اجسام، مستقیماً مکان‌یابی و دسته‌بندی اجسام را در یک مرحله انجام می‌کنند. این مدل‌ها معمولاً سریع‌تر از مدل‌های دو مرحله‌ای هستند و برای کاربردهای نیازمند به سرعت بالا و بی‌درنگ، مانند رانندگی خودکار و رباتیک، مناسب‌تر هستند.

۱-۲-۵-SSD

در سال ۲۰۱۵، لیو و همکاران آشکارساز چند جعبه ای تک‌شات (SSD)[۱]را معرفی کردند [۳۴]. مدل SSD یکی از مدل‌هایمعروف برای تشخیص جسم در تصاویر است که با استفاده از یکCNN آموزش داده می‌شود.

این مدل به صورت همزمان و با یکبار اجرا، می‌تواند اجسام را در تصاویر شناسایی کند و با تعیین مکان و اندازه آن‌ها، ابعاد بیشتری از اطلاعات مورد نیاز برای تشخیص جسم را فراهم کند.

[۱] Single Shot Multibox Detector

در این مدل، از مفهوم جعبه مرجع استفاده می‌شود که به عنوان نماینده‌ی مکان و اندازه‌ی اجسام است و برای تشخیص اجسام به وسیله شبکه‌ی عصبی، خروجی شبکه با این جعبه‌ها مقایسه می‌شود (شکل ۱۳)). به طور خاص، روش SSD با داشتن سرعت و دقت بالا در برنامه‌هایی که نیاز به تشخیص اجسام به صورت بی‌درنگ دارند کاربرد دارد.مزیت اصلیSSD نسبت به مدل‌های پیشین این است که این مدل به اندازه تکنیک‌های کندتر که دارای فرایند نواحی پیشنهادی هستند، دقیق است.

از زمان ارائه مدل SSD ، بسیاری از محققان این مدل را به عنوان یکی از بهترین مدل‌های تشخیص جسم در دنیای پردازش تصویر شناخته‌اند و به دنبال بهبود و ارتقاء آن هستند. در این راستا، پژوهش‌های فراوانی برای بهبود مدل SSD ارائه شده‌است. به عنوان مثال، در مقاله [۳۵]، یک شبکه بهبود یافته بر اساس مدل SSD معرفی شده است که از یک ساختار ترکیبی ویژگی‌های چندلایه برای افزایش اطلاعات معنایی ویژگی‌های کم عمق استفاده می‌کند. علاوه بر این، در این شبکه از یک ماژول توجه[۱] و مسیره برای نمایش اطلاعات ویژگی استفاده شده است که با استفاده از مکانیزم توجه، تاثیر نویز پس‌زمینه را کاهش می‌دهد و ویژگی‌های کلیدی را برجسته می‌کند. همچنین، تابع هزینه برای کاهش عدم تعادل بین نمونه‌های مثبت و منفی بهینه شده است. نتایج ارائه شده بر روی مجموعه‌داده‌های تصاویر ماهواره‌ای، کارایی این روش مبتنی بر SSD را به خوبی نشان می‌دهد.

همچنین در مقاله [۳۶] در سال ۲۰۲۲ یک مدل بر پایه‌یSSD بهبود یافته برای تشخیص وسیله‌نقلیه در صحنه‌های ترافیکی پیشنهاد شده‌است. در این روش شبکه MobileNet به عنوان شبکه استخراج ویژگی برایSSD انتخاب شده که عملکرد بی‌درنگ بودن الگوریتم را بهبود می‌بخشد. هم‌چنین برای بهبود دقت تشخیص، تکنیک‌های توجه کانال و واپیچش[۲] استفاده شده‌است.

RetinaNet-۵-۲-۲

در سال ۲۰۱۷ مدل یک مرحله‌ای RetinaNet معرفی شد [۳۷] که یکی از مدل‌های پرکاربرد برای تشخیص اجسام در تصاویر است. RetinaNet با استفاده از شبکه‌های عصبی پیچشی و الهام گرفتن از ساختار شبکه‌هایی از قبیلSSD و Faster-RCNN ، بهبودهای قابل توجهی را در دقت تشخیص اجسام داشته است. RetinaNet برای رفع مشکل کاهش گرادیان در شبکه‌های عصبی با شبکه‌های پیچشی مبتنی بر هرم ویژگی و خطای کانونی[۳] کار می‌کند. هرم ویژگی به این معناست که از تصاویر با اندازه‌های مختلف برای آموزش شبکه استفاده می‌شود تا شبکه بتواند اجسام با اندازه‌های مختلف را به خوبی تشخیص دهد.همچنین، خطای کانونی به شبکه کمک می‌کند تا به دسته‌های کم تکرار بیشتر توجه کند. این خلاقیت‌های قابل توجه که در معماری این مدل در نظر گرفته شده‌اند، باعث می‌شوند دقت متوسط تشخیص اجسام افزایش یابد.

CenterNet-۵-۲-۳

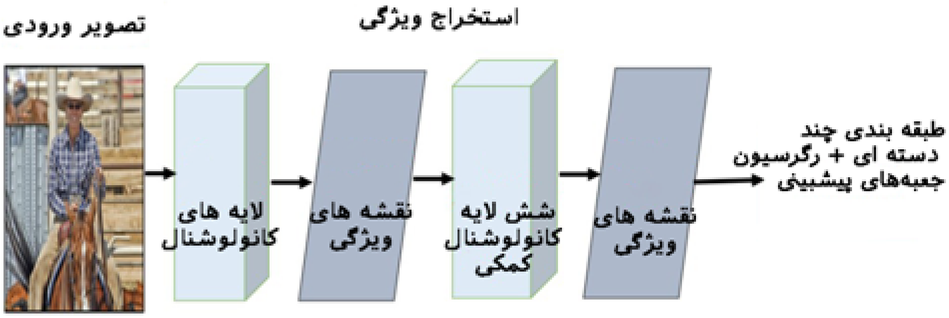
مدل‌های تشخیص جسم مبتنی بر CNN معمولاً از جعبه‌های محور[۴] استفاده می‌کنند. این جعبه‌ها با اندازه و موقعیت‌های مختلف و از پیش تعیین شده هستند و به عنوان مرجعی برای پیش‌بینی موقعیت و اندازه اجسام در تصویر استفاده می‌شوند. در واقع، برای هر ناحیه از تصویر، چندین جعبه‌ی محور پیش‌تعیین شده تعریف می‌شود و برای شناسایی اجسام، جعبه‌ای که بیشترین هم‌پوشانی با هر جسم را دارد، به عنوان جعبه نهایی آن جسم انتخاب می‌شود. یکی از مشکلات اصلی استفاده از جعبه‌های محور تعیین اندازه و نسبت ابعاد مناسب برای آن‌ها و تعداد آن‌ها است، زیرا عدم مطابقت اندازه جعبه‌های محور با اجسام هدف باعث کاهش دقت مدل می‌شود.

الگوریتم تشخیص جسم CenterNet [۳۸] که در سال ۲۰۱۹ معرفی شد، یک مدل بدون جعبه محور است که قدم مهمی در راستای رفع مشکلات جعبه‌های محور برداشت.

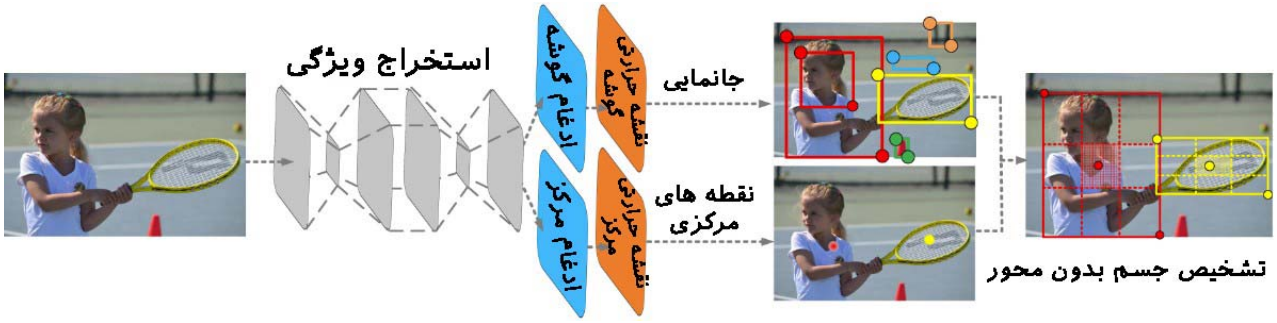
این مدل از یک معماری شبکه کاملاً کانولوشن برای پیش‌بینی نقطه‌ی مرکزی جعبه، اندازه و دسته‌بندی اجسام در یک تصویر استفاده می‌کند.به جای استفاده از جعبه‌های محور از پیش تعریف شده، CenterNet به‌طور مستقیم نقطه مرکزی هر جسم را با استفاده از رویکرد مبتنی بر نقشه حرارتی[۵] پیش‌بینی می‌کند. معماری شبکه شامل یک شبکه استخراج ویژگی و به دنبال آن سه بخش موازی است که نقشه حرارتی، اندازه و موقعیت هر جسم را پیش‌بینی می‌کند. بخش‌های اندازه و موقعیت به ترتیب اندازه و مکان جعبه مشخص‌کننده هر جسم را پیش‌بینی می‌کنند، در حالی که بخش نقشه حرارتی نقطه مرکزی هر جسم را پیش‌بینی می‌کند. سپس نقاط مرکزی پیش‌بینی‌شده برای به‌دست آوردن مکان‌های نهایی اجسام مورد استفاده قرار می‌گیرند.

همچنینCenterNet مانند RetinaNet از یک تابع خطای کانونی برای تمرکز بر نمونه‌های سخت در طول آموزش استفاده می‌کند. به‌طور کلی، رویکرد بدون محور CenterNet امکان تشخیص ساده‌تر و کارآمدتر اجسام را فراهم می‌کند و در عین حال به عملکرد بهتر در معیارهای مختلف دست می‌یابد. شکل (۱۴) ساختار کلی این مدل را نشان می‌دهد.

[۱] Attention Module
[۲] Deconvolution
[۳] Focal Loss
[۴] Anchor Boxes
[۵] Heatmap



شکل (۱۳): ساختار کلی SSD.



شکل (۱۴): ساختار کلی مدل CenterNet [۳۸].

۴-۲-۵-YOLO

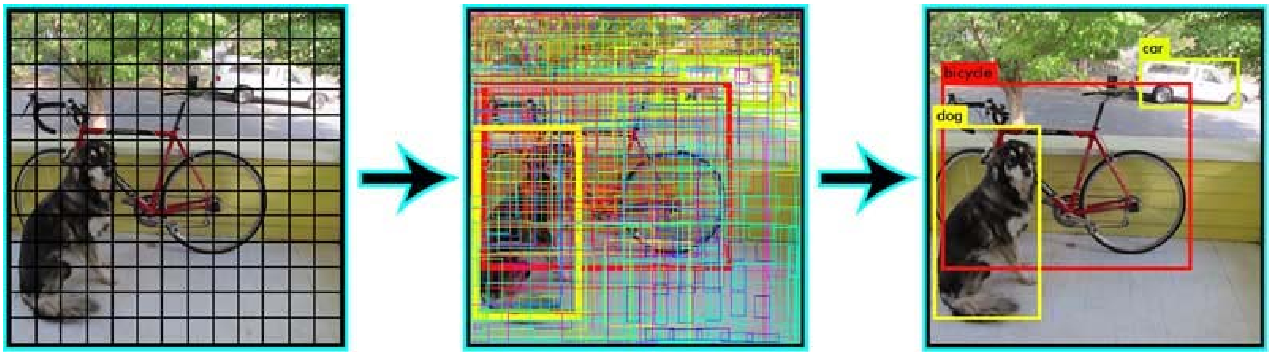
مدل [You Only Look Once]۹یکی از مدل‌های شناسایی و دسته‌بندی اجسام است که به عنوان یک مدل یک مرحله‌ای شناخته می‌شود. این مدل به مسئله تشخیص جسم به عنوان یک مسئله رگرسیون ساده نگاه می‌کند و می‌تواند به طور همزمان چندین جسم را در یک تصویر شناسایی و احتمال تعلق آن‌ها به هر دسته را پیش‌بینی کند.این مدل عمل تعیین موقعیت و دسته‌بندی را در یک مرحله انجام می‌دهد و به مسئله تشخیص جسم به شیوه‌ای جدید نگاه می‌کند که با مدل‌های قبلی متفاوت است و عملکرد بهتری در زمان و سرعت اجرا دارد.

در مدل YOLO ، ابتدا یک شبکه عصبی پیچشی روی تصویر کامل اعمال می‌شود و سپس تصویر به S سلول تقسیم می‌شود. برای هر سلول، جعبه‌هایی که احتمال حضور جسم در آنجا وجود داشته باشد و همچنین احتمال تعلق آن‌ها به هر دسته را به صورت همزمان پیش‌بینی می‌کند. در نهایت، با استفاده از عملیاتNMS جعبه‌ها ترکیب می‌شوند. NMS جعبه‌های اضافی و نامربوط را فیلتر می‌کند و فقط دقیق‌ترین آن‌ها را برای هر جسم حفظ می‌کند.شکل (۱۵) نحوه عملکرد الگوریتمNMSو شکل (۱۶) ساختار کلی اولین نسخه YOLO را نشان می‌دهد.

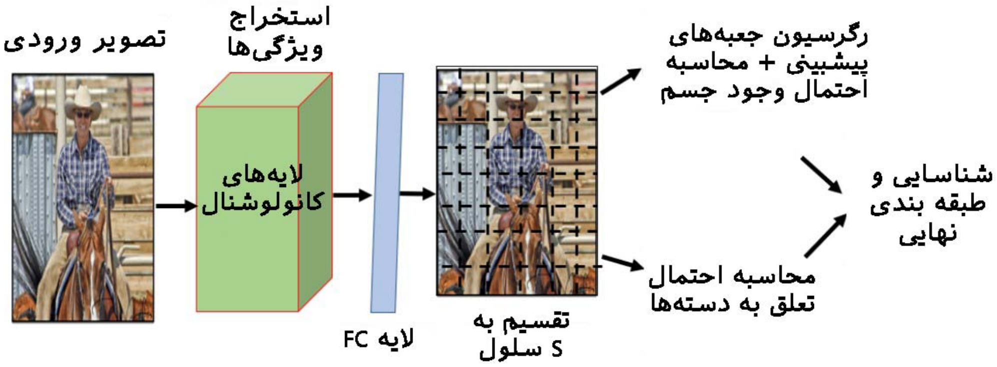
از مزایای این مدل نسبت به مدل‌های پیشین مانند RCNN می‌توان به سرعت تشخیص بالای آن برای مسائل بلادرنگ و توانایی بهتر در تشخیص اجسام دارای همپوشانی اشاره کرد.

با وجود بهبود قابل توجه سرعت شناسایی، اولین نسخه‌هایYOLO در مقایسه با مدل‌های دو مرحله‌ای، در جانمایی جعبه‌هاپیش‌بینی ضعیف‌تر عمل می‌کردند. نسخه‌های بعدیYOLO و برخی از مدل‌های بر پایهYOLO بهبود یافته، توجه بیشتری به این مشکل داشتند. برای مثال، مدل YOLOv2 که در سال ۲۰۱۷ معرفی شد [۳۹] از نقشه‌های ویژگی چند مقیاسی برای شناسایی اجسام استفاده می‌کند و برای دستیابی به شناسایی چند مقیاسی، نقشه‌های ویژگی با وضوح بالا را با نقشه‌های ویژگی با وضوح پایین ترکیب می‌کند. با تنظیم جعبه‌های پیش‌بینی در مقیاس‌های مختلف، مدل بر روی تشخیص اجسامی متمرکز می‌شود که از نظر شکل شبیه جعبه‌ی پیش‌بینی هستند.

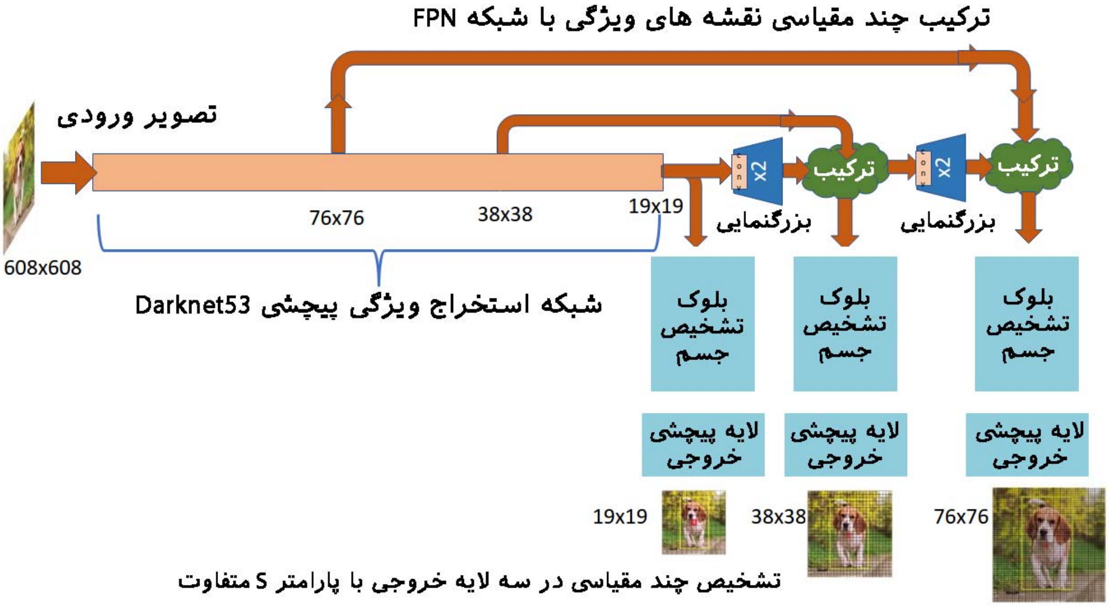
آخرین نسخه رسمیYOLO که توسط نویسندگان اصلی آن یعنیRedmon و همکاران معرفی شد، نسخه‌یYOLOv3 بود [۴۰]. اینمدل دو پیشرفت عمده را نسبت به YOLOv2 معرفی کرد. اولین پیشرفت استفاده از مدل Residual برای تعمیق بیشتر ساختار شبکه است. مورد دیگر استفاده از معماریFPN برای دستیابی به شناسایی چندمقیاسی بهتر نسبت به مدل‌های قبل است.مدل YOLOv3 از ایده شبکه ResNet [۴۱] برای طراحی شبکه جدیدDarkent53 به عنوان شبکه استخراج ویژگی استفاده می‌کند. همچنین از ایده چندمقیاسیFPN استفاده می‌کند و ۳ مقیاس مختلف را در هر لایه‌ی خروجی پیش‌بینی می‌کند. این بهبود باعث می‌شودکه این مدل بتواند اجسام با مقیاس‌های مختلف را با دقت بهتری تشخیص دهد. از طرف دیگر، مدل YOLOv3 موقعیت جسم را از طریق روش پیش‌بینی رگرسیون فریم پیش‌بینی می‌کند که از رگرسیون خطی پایدارتر است. شکل (۱۷) ساختار این مدل را نشان می‌دهد.



شکل (۱۵): نحوه عملکرد **NMS** [۹].



شکل (۱۶): ساختار کلی اولین نسخه **YOLO**.



شکل (۱۷): ساختار کلی مدل **YOLOv3** [۴۰].

از مدل YOLOv3 به بعد مدل‌های YOLO از معماری سه قسمتی استخراج ویژگی (Backbone)، ترکیب ویژگی (Neck) و تشخیص جسم (Head) استفاده می‌کنند. این ساختار انعطاف پذیری این مدل‌ها را افزایش می‌دهد، به طوری که به محققان اجازه می‌دهد که با توجه به اهداف و کاربرد مورد نظر، شبکه‌های دیگری را به جای شبکه‌های پیش‌فرض Darknet53 و FPN برای استخراج و ترکیب ویژگی‌های اجسام استفاده کنند.

با وجود عدم ارائه نسخه‌های رسمی جدید توسط نویسندگان YOLO در سال‌های بعد، راهکارهای بهبود یافته بر پایه YOLO توسط محققان دیگر نیز ارائه شده‌اند. در سال ۲۰۱۹ مدل YOLOv4 توسط Bochkovskiy و همکاران معرفی شد [۴۲]. هسته‌ی اصلی معرفی شده در این مدل، شبکه‌ی CSPDarknet53 است که برای استخراج ویژگی‌های هدف استفاده می‌شود. YOLOv4 با درس‌هایی از تجربه‌ی CSPNet[۲۷] در حفظ دقت، کاهش محاسبات مورد نیاز و هزینه‌های حافظه، CSP را به هر بلوک باقی‌مانده در شبکه اضافه می‌کند، نقشه ویژگی لایه پایه را به دو قسمت تقسیم می‌کند و آن‌ها را از طریقیک سلسله مراتب چند مرحله‌ای ادغام می‌کند.

این کار میزان محاسبه را کاهش داده،ولی حفظ دقت را تضمین می‌کند. CSPDarknet53 از تابع فعال سازی Mish [۴۳] استفاده می‌کند تا شناسایی دقیق‌تر باشد. برخلاف مدل YOLOv3 که از FPN برای تشخیص چند مقیاسی استفاده می‌کند، YOLOv4 از ایده گردش اطلاعات در شبکه تجمیع مسیر[1] [۴۴] استفاده می‌کند که باعث افزایش دقت در شناسایی چند مقیاسی می‌شود. همچنین در YOLOv4 روش‌های جدیدی برای بهبود و آماده سازی تصاویر در حین آموزش شبکه معرفی شده‌است.

در ادامه‌ی تلاش محققان برای بهبود مدل‌هایYOLO ، مدل YOLOv5 توسط Jocher و همکاران در سال ۲۰۲۰ معرفی شد [۴۵] که ساختار کلی این مدل مشابه با YOLOv4 است. همانطور که پیش‌تر گفته شد، در مدل‌هایYOLO از جعبه‌های محور برای رگرسیون جعبه‌های پیش‌بینی استفاده می‌شود. در نسخه‌های قبل این جعبه‌ها یا ثابت بوده یا مانند YOLOv4 با استفاده از الگوریتم k-means محاسبه می‌شوند.ا توجه به اینکه مجموعه‌داده‌های مختلف دارای طیف گسترده‌ای از اندازه‌ها و نسبت‌های ابعادی از اجسام هستند که ممکن است به تعداد و اندازه‌های بسیار متفاوتی از جعبه‌های محور نیاز داشته باشند، در YOLOv5 پس از مرحله k-means، یک مرحله الگوریتم ژنتیک نیز اضافه شده که نتایجk-means را به عنوان ورودی دریافت می‌کند و جعبه‌های محور بسیار بهتری را پیدا می‌کند. این جعبه‌ها برای مجموعه‌داده‌ای که در حال استفاده است و برای نوع اجسام هدف تنظیم شده‌اند که در عمل دقت نهایی مدل را برای شناسایی اجسام مورد نظر بالا می‌برد. برخلاف نسخه‌های قبلیYOLO که با زبان C نوشته شده و از کتابخانه‌ی DarkNet استفاده می‌کنند، پیاده‌سازیYOLOv5 با پایتون و با استفاده از کتابخانه‌ی Pytorch انجام شده است که یک تغییر اساسی است. بر اثر این تغییر، آموزش مدل سریع‌تر و استقرار آن بر روی دستگاه‌های مختلف بهتر شده و حجم فایل مدل نهایی نسبت به YOLOv4 کاهش یافته است. همچنین مدل YOLOv5 برای اولین بار در یک مدل مبتنی بر YOLO نسخه‌های متعدد با عمق و تعداد نورون متفاوت معرفی کرد.شکل (۱۸) ساختار کلی آخرین نسخه YOLOv5 را نشان می‌دهد.

در سال ۲۰۲۱ مدل YOLOR[2]توسط تیم تحقیقاتی که پیش‌تر رویYOLOv4 کار کرده بودند منتشر شد [۴۷]. در این مقاله، نویسندگان یک رویکرد یادگیری چند وظیفه‌ای را توسعه دادند که هدف آن ایجاد یک مدل واحد برای وظایف مختلف (به عنوان مثال، دسته‌بندی، تشخیص، تخمین موقعیت) با یادگیری ویژگی‌های کلی و استفاده از شبکه‌های فرعی برای ایجاد نمایش‌های خاص است. با این بینش که روش یادگیری مشترک سنتی اغلب منجر به تولید ویژگی‌های غیربهینه می‌شود،YOLOR سعی می‌کند با رمزگذاری دانش ضمنی شبکه‌های عصبی برای اعمال چندین کار (مشابه نحوه استفاده انسان از تجربیات گذشته برای نزدیک شدن به مشکلات جدید)، بر این مسئله غلبه کند.

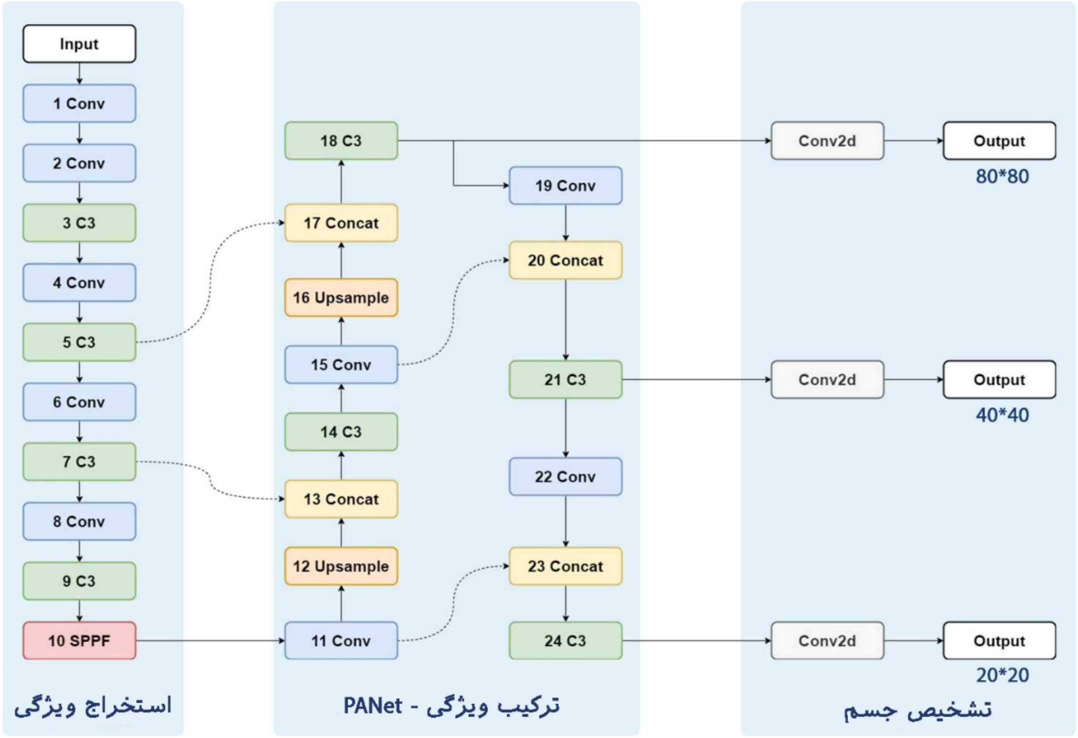
چند ماه بعد، مدل تشخیص جسم YOLOX توسط تیم تحقیقاتیMegviiTechnology منتشر شد [۴۸]. این مدل که با استفاده از Pytorch توسعه یافته و از YOLOv3 به عنوان شبکه‌ی پایه استفاده می‌کند، دارای سه تغییر اصلی نسبت به آن است: شناسایی بدون جعبه‌ی محور مشابه با CenterNet، بخش شناسایی جسم (Head) مجزا و تخصیص برچسب پیشرفته. شناسایی بدون جعبه‌ی محور به بهبود روند آموزش کمک می‌کند و جداسازی بخش پیش‌بینی موقعیت اجسام و تشخیص دسته باعث افزایش سرعت همگرایی و دقت مدل می‌شود. به علاوه، معرفی و استفاده از راهکار تخصیص برچسب پیشرفته simOTA باعث شده که این مدل در هنگامی که جعبه‌های مشخص کننده چند جسم باهم تداخل دارند عملکرد دقیق‌تری داشته باشد. مدل تشخیص جسم YOLOX با برقراری تعادل بهتر بین سرعت و دقت به نتایج بهتری نسبت به دیگر مدل‌های مبتنی بر YOLO دست یافته‌است. شکل (۱۹) تفاوت بخش شناسایی جسم در معماریYOLOX نسبت به YOLOv3 را نشان می‌دهد.

در سال‌های۲۰۲۲ و ۲۰۲۳ پیشرفت مدل‌هایYOLO با معرفیYOLOv6 [۴۹]، YOLOv7 [۵۰]وYOLOv8 [۵۱] سرعت بیشتری یافت به طوری که مدل‌هایYOLO به دردسترس‌ترین و برترین مدل‌های تشخیص جسم تبدیل شدند.

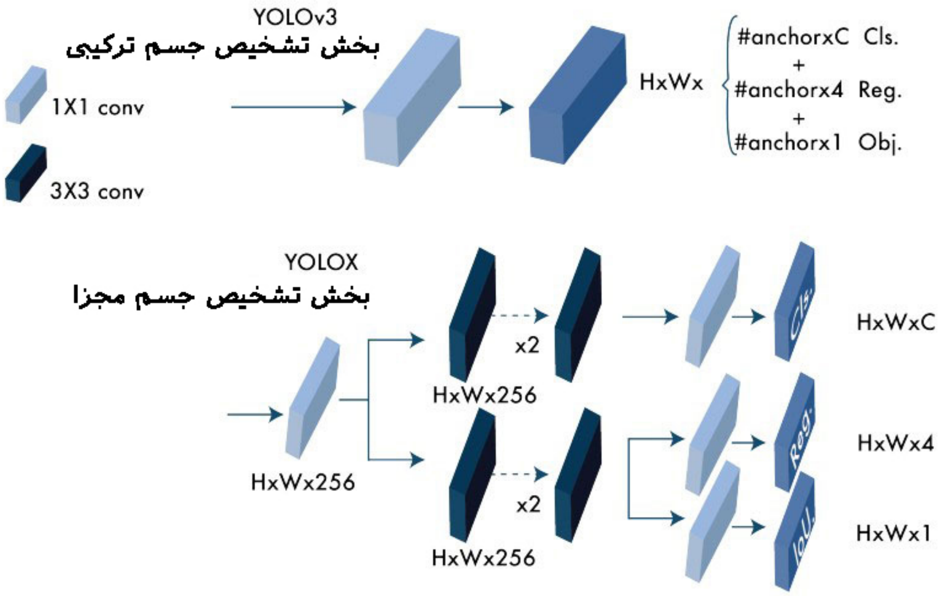
مدل تشخیص جسم YOLOv6در سال ۲۰۲۲ معرفی شد و مانند مدل YOLOv5 نسخه‌هایی با پیچیدگی محاسباتی متفاوت برایتشخیص بدون جعبه‌های محور بهره می‌برد. تفاوت اصلی این مدل معرفی و استفاده از یک شبکه استخراج ویژگی ابتکاری به نام EfficientRep است که از موازی سازی بالاتری نسبت به شبکه استخراج ویژگی قبلی در YOLO استفاده می‌کند که از شبکه RepVGG [۵۲] الهام گرفته شده است. همچنین این مدل برای قسمت ترکیب ویژگی، از PANet [۴۴] بهبود یافته با بلوک‌های RepBlock [۵۲] استفاده می‌کند و با الهام از YOLOX قسمت تشخیص جسم جدا شده دارد. شکل (۲۰) ساختار کلی مدل YOLOv6 را نشان می‌دهد.

در اوایل سال ۲۰۲۳ اولین نسخه YOLOv8 توسط شرکت Ultralytics که پیشترYOLOv5 را معرفی کرده بود منتشر شد. این مدل که هنوز در زمان نگارش این مقاله فاقد مقاله علمی رسمی است، از یک شبکه استخراج ویژگی بهبود یافته بر اساس شبکه Darknet-53 با استفاده از بلوک‌های جدید C2F و شناسایی سریع بدون محور بهره می‌برد. همچنین این مدل به غیر از تشخیص جسم،

[1] Path Aggregation Network
[2] You Only Learn One Representation



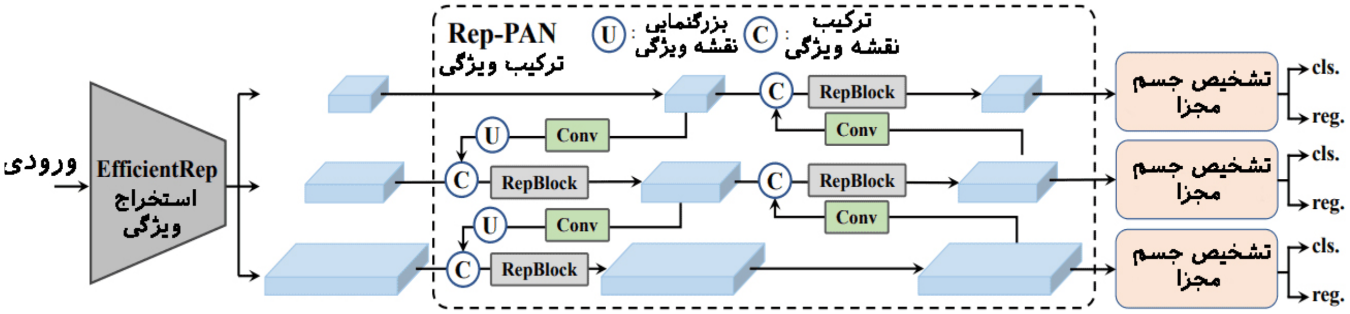
شکل (۱۸): معماریYOLOv5 . این مدل از شبکه استخراج ویژگی بهبود یافتهCSPDarkNet و خروجی مشابه YOLOv3 بهره می‌برد و برای تشخیص جسم چند مقیاسی از PaNet استفاده می‌کند.



شکل (۱۹): تفاوت بین بخش تشخیص جسم در YOLOv3 و YOLOX . مدل YOLOX برای هر سطح، ابتدا از یک لایه پیچشی ۱ × ۱ برای کاهش ابعاد نقشه‌های ویژگی استفاده می‌کند و سپس از دو بخش تشخیص مجزا، به ترتیب برای وظایف طبقه بندی و پیش‌بینی موقعیت جسم(رگرسیون) استفاده می‌کند و در نهایت یک بخش مجزای دیگر برای محاسبه IoU اضافه می‌شود که به تعیین جعبه‌های مشخص کننده‌ی جسم کمک می‌کند. در حالی که در YOLOv3 از یک بخش یکسان برای طبقه‌بندی، رگرسیون و محاسبه IoU استفاده می‌شود [۴۸].

قابلیت دسته بندی و تقسیم بندی تصویر را نیز دارد که آن را به یک مدل کاربردی برای مسائل مختلف تبدیل می‌کند. این مدل نسبت به دیگر مدل‌های مبتنی بر YOLO دارای سرعت و دقت بیشتری است و در مقایسه با YOLOv6 و YOLOv7 زمان آموزش کمتری نیاز دارد.

در این بخش، انواع مدل‌های یک مرحله‌ای و دومرحله‌ای توسعه یافته برای کاربرد تشخیص جسم مورد بررسی قرار گرفت. بخش بعد به راهکارهایی برای افزایش سرعت تشخیص در مدل‌های مختلف می‌پردازد.



شکل (۲۰): ساختار کلی YOLOv6 [۴۹].

۶-راه‌کارهای افزایش سرعت پردازش

سرعت یک مدل تشخیص جسم برای مدت طولانی یک مشکل چالش برانگیز بوده است و در سال‌های اخیر راهکارهای متعددی برای کاهش پیچیدگی محاسباتی و افزایش سرعت پردازش در این زمینه ارائه شده است.

۶-۱- پردازش مشترک نقشه ویژگی

در میان مراحل مختلف محاسباتی برای تشخیص جسم، استخراج ویژگی معمولاً بیشترین زمان را به خود اختصاص می‌دهد. رایج‌ترین ایده برای کاهش افزونگی محاسبات ویژگی‌ها، محاسبه نقشه ویژگی کل تصویر تنها برای یکبار است [۳۰] که دهها بار سرعت تشخیص را افزایش می‌دهد.

۶-۲- تشخیص آبشاری

تشخیص آبشاری یک تکنیک رایج است که یک فلسفه آن در تشخیص اجسام از درشت به ریز است: فیلتر کردن بیشتر پنجره‌های پس‌زمینه ساده با استفاده از محاسبات ساده و سپس پردازش پنجره‌های دشوارتر با پنجره‌های پیچیده. در سال‌های اخیر، تشخیص آبشاری به‌ویژه برای کاربردهای تشخیص "اجسام کوچک در صحنه‌های بزرگ" مانند تشخیص چهره [۵۴] و تشخیص عابر پیاده[۵۵] استفاده شده است.

۶-۳- هرس و کمی‌سازی شبکه

هرس شبکه" و "کمی سازی شبکه" دو روش متداول برای افزایش سرعت مدل‌هایCNN هستند که به ترتیب به هرس کردن ساختار یا وزن های شبکه و کاهش طول کدها در آن اشاره دارند. روش‌های هرس شبکه معمولاً یک فرآیند آموزش و هرس تکراری را انجام می‌دهند،یعنی تنها گروه کوچکی از وزن‌های کم‌اهمیت را پس از هر مرحله از آموزش حذف می‌کنند و آن عملیات را تکرار می‌کنند [۵۶]. کارهای اخیر در مورد کمی سازی شبکه عمدتاً بر دودویی سازی شبکه متمرکز است، که هدف آن فشرده‌سازی یک شبکه با کاهش توابع فعالسازی و وزن‌های آن به متغیرهای دودویی است تا عملیات پیچیده به عملیات منطقی تبدیل شود.

۶-۴- طراحی شبکه‌ی سبک

مهمترین راهکار برای افزایش سرعت در مدل‌های مبتنی بر CNN، طراحی مستقیم شبکه‌های سبک وزن است. این دسته از مدل‌ها شبکه‌های سبک و کارآمدی دارند که عملکرد مناسب در سیستم‌های با توان پردازشی محدود مانند اینترنت اشیا و دستگاه‌های قابل حمل را هدف قرار می‌دهند.

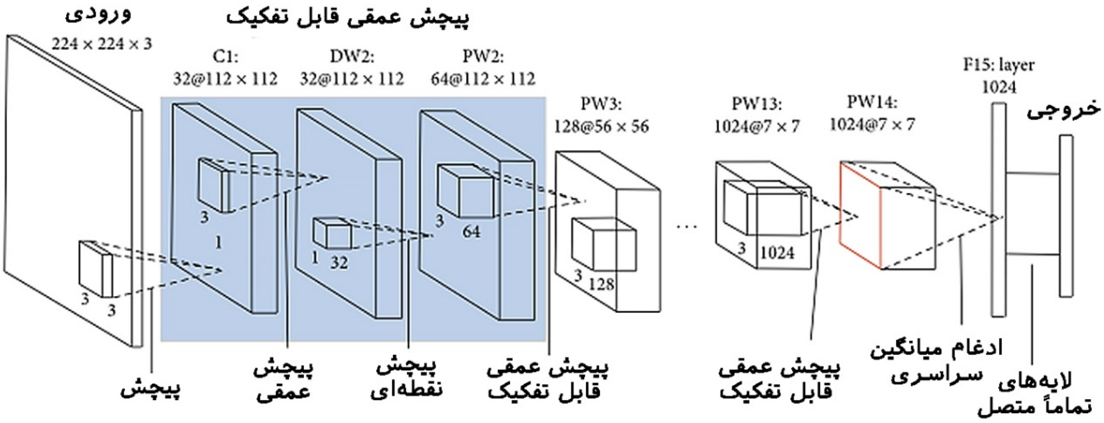
۶-۴-۱-SqueezeNet

یکی از نخستین شبکه‌های سبک مطرح شده، شبکه‌ی SqueezeNet[۵۷] است که در آن اندازه‌ی شبکه بدون کاهش شدید در دقت آن کاهش می‌یابد. پیشنهاد دهندگان این شبکه از سه استراتژی طراحی برای دستیابی به این هدف استفاده کرده‌اند: استفاده از فیلترهای کوچکتر، کاهش کانال‌های ورودی به فیلترهای ۳×۳و قرار دادن لایه‌های کوچک‌سازی[۱] در شبکه. بلوک تشکیل دهنده‌ی اصلی در SqueezeNet ماژول Fire نامیده می‌شود که از یک لایه فشرده‌سازی و یک لایه گسترش با تابع فعال‌سازیReLU تشکیل شده است. نویسندگان با استفاده از تکنیک فشرده سازی عمیق توانستند به کاهش قابل توجهی در اندازه‌ی مدل دست یابند.

۶-۴-۲-MobileNet

مدل MobileNet [۵۸] یک شبکه عصبی است که از یک معماری کارآمد استفاده می‌کند و با روش‌های مرسوم مدل‌های کوچک متفاوت است. همانطور که در شکل (۲۱) مشخص است، این شبکه از راهکار متفاوتی به نام پیچش قابل تفکیک عمقی برای افزایش سرعت استفاده می‌کند که یک پیچش استاندارد را به یک پیچش عمقی و یک پیچش نقطه‌ای تبدیل می‌کند. پیچش عمقی فیلترهای مختلفی را بر روی هر کانال ورودی اعمال کرده و از یک پیچش نقطه‌ای برای ترکیب ورودی‌ها استفاده می‌کند و هزینه محاسبات و اندازه‌ی مدل را کاهش می‌دهد. این شبکه دارای ۲۸ لایه پیچشی است که هر کدام با تابع فعال‌سازی ReLU دنبال

[۱] Downsampling



شکل (۲۱): ساختار کلی شبکه‌ی سبک MobileNet [۵۸].

می‌شوند. همچنین برای بهبود سرعت و کاهش اندازه از دو پارامتر کوچک‌کننده مدل یعنی ضریب عرض و عمق استفاده می‌کند. MobileNet دقت قابل توجهی را در مقایسه با مدل‌های بزرگتر به دست آورده و می‌تواند به مسائل مختلف مانند تشخیص اجسام تعمیم یابد. مدل‌های MobileNet2 [۵۹] و MobileNet3 [۶۰] دارای سرعت و دقت بیشتری از نسخه‌ی اصلی هستند وتکنیک‌هاییمانند تنگناهای خطی و توابع فعال‌سازی پیشرفته استفاده می کنند.

۳-۴-۶-ShuffleNet

شبکه ShuffleNet [۶۱] یک معماری شبکه عصبی محاسباتی کارآمد است که برای دستگاه‌های قابل حمل طراحی شده است. این شبکه از عملیات کانولوشن جمعی و جابجایی کانال به جای عملیات‌های کانولوشن عادی استفاده می‌کند.

این شبکه عمدتاً از یک لایه پیچشی و به دنبال آن از واحدهای ShuffleNet تشکیل شده است. واحدهای ShuffleNet دارای سه مرحله هستند، این واحدها در مرحله اول عملیات جابجایی کانال را بر ورودی اعمال می‌کنند، سپس از کانولوشن عمقی با ابعاد ۳×۳ استفاده می‌کنند، و در نهایت یک لایه کانولوشن جمعی نقطه‌ای ۱×۱ را اعمال می‌کنند. با وجود عملکرد بهتر از مدل‌های دیگر در زمان معرفی، این شبکه در مقایسه با شبکه‌های دیگر بهبود قابل توجهی در سرعت استنتاج ندارد.

مدتی پس از معرفی ShuffleNet شبکه‌ی ShuffleNetv2 [۶۲] معرفی شد. عملکرد این شبکه با معرفی یک تکنیک جدید تقسیم کانال که به شبکه اجازه می‌دهد کانال‌ها را به دو گروه تقسیم کند و نمایش‌های جداگانه‌ای برای آن‌ها بیاموزد، نسبت به ShuffleNet بهبود یافت. همچنین یک عملیات جدید المان محور را معرفی کرده است که به تبادل اطلاعات بین دو گروه کمک می‌کند. علاوه بر این، نسخه‌ی جدید آن طراحی بلوک کارآمدتری را در خود جای داده است که از کانال‌های کمتری استفاده می‌کند و یک اتصال باقی‌مانده‌ی جدید برای بهبود بیشتر دقت اضافه می‌کند.

این پیشرفت‌ها منجر به آموزش سریعتر و دقت بهتر در مقایسه با نسخه اول ShuffleNet شد.

پس از بررسی راهکارهای افزایش سرعت آموزش و استنتاج در شبکه‌های تشخیص جسم که استفاده از این شبکه‌ها را در بسیاری از کاربردهای واقعی و بی‌درنگ امکان‌پذیر می‌سازد، در قسمت بعد به ذکر پیشرفت‌های اخیر در روش‌های تشخیص جسم خواهیم پرداخت و این بررسی در راستای روشن کردن چشم‌انداز تحقیق در این زمینه خواهد بود.

۷-راهکارهای افزایش دقت تشخیص و افزایش کارآیی

راهکارهای تشخیص جسم همواره در حال تکامل بوده‌اند و نوآوری‌های متعددی در پیشرفت آن‌ها نقش داشته‌اند. در این بخش، هدف بررسی برخی از مهمترین پیشرفت‌هایی است که در بهبود دقت تشخیص، سرعت، کارایی و همچنین توسعه روش‌ها و راهکارهای جدید نقش موثری ایفا کرده‌اند.

۷-۱- تشخیص بدون پنجره‌ی متحرک

از آنجایی که یک جسم در یک تصویر را می‌توان با نقطه گوشه سمت چپ بالای آن و نقطه گوشه پایین سمت راست آن تعیین کرد، بنابراین عمل تشخیص و مکان‌یابی جسم را می‌توان با عنوان پیدا کردن این دو گوشه یا دو نقطه‌ی کلیدی تعریف کرد.

یکی از پیاده‌سازی‌های اخیر برای این ایده، پیش‌بینیک نقشه حرارتی برای گوشه‌ها است [۶۳]. برخی از روش‌های دیگر نیز با پیروی از این ایده سعی در استخراج نقاط کلیدی بیشتری برای دستیابی به عملکرد بهتر دارند [۳۸].

راهکار دیگری که برای این مورد پیشنهاد شده‌است، اجسام را به عنوان مجموعه‌ای از نقاط در نظر گرفته و مستقیماً ویژگی‌های جسم (مثل ارتفاع و عرض) را بدون گروه‌بندی پیش‌بینی می‌کند [۶۴]. مزیت این رویکرد این است که می‌توان آن را تحت یک

چارچوب تقسیم‌بندی معنایی پیاده سازی کرد و نیازی به طراحی جعبه‌های پیش‌بینی چند مقیاسی نیست.

۲-۷- بهبود تشخیص چرخش اجسام و تغییر مقیاس

در سال‌های اخیر تلاش‌هایی برای بهبود مقاومت روش‌های تشخیص جسم در برابر تغییراتی مانند چرخش و تغییرات مقیاس اجسام انجام شده است.

۱-۲-۷- بهبود تشخیص چرخش

چرخش اجسام در سناریوهای مختلف تشخیص مانند تشخیص چهره، تشخیص متن و سنجش از راه دور متداول است. یک رویکرد رایج برای کمک به رفع این مشکل، بهبود یادگیری ویژگی‌های اجسام در جهت‌های مختلف با تکرار و تغییر زاویه تصاویر در طول آموزش است [۶۵].

همچنین، می‌توان مدل‌های مستقل را برای هر زاویه به صورت جداگانه آموزش داد [۶۶]. برخی از مطالعات اخیر استفاده از توابع هزینه ثابت چرخش را پیشنهاد کرده‌اند که در آن محدودیتی به هزینه‌ی تشخیص اضافه می‌شود تا اطمینان حاصل شود که ویژگی اشیایی که چرخیده‌اند بدون تغییر باقی می‌ماند [۶۷]. راه حل دیگر یادگیری تبدیل‌های هندسی اجسام کاندیدا در تصویر است [۶۸]. در مدل‌های تشخیص جسم دو مرحله‌ای، با ادغام ناحیه‌ی مورد نظر یاROI[۱] با ناحیه‌ی هدف، استخراج نقشه ویژگی اجسام در هر ناحیه از تصویر و با هر ابعادی انجام می‌شود. با توجه به اینکه این ادغام ویژگی معمولاً در مختصات دکارتی انجام می‌شود، که معمولاً در زمان چرخش جسم ثابت نمی‌ماند، اخیراً روشی برای انجام ادغام ROI پیشنهاد شده که براساس مختصات قطبی است و ویژگی‌ها را در برابر تغییرات چرخشی مقاوم می‌کند [۶۹].

۲-۲-۷- بهبود تشخیص تغییر مقیاس و ابعاد

برای کاربرد تشخیص جسم، تاکنون مطالعات متعددی بر روی بهبود تشخیص در تغییر مقیاس در هر دو مرحله‌ی آموزش و آزمایش انجام شده‌است. مدل‌های تشخیص جسم مدرن معمولاً تصاویر ورودی را به اندازه ثابت تغییر می‌دهند و هزینه‌ی تشخیص اجسام را در همه مقیاس‌ها به عهده‌ی روند انتشار به عقب[۲] می‌اندازند. با این حال، این رویکرد منجر به بروز مشکل "عدم تعادل مقیاس"[۳] می‌شود. یکی از راه‌حل‌ها استفاده از هرم تصویر در حین تشخیص است، نرمال سازی مقیاس برای هرم تصویری (SNIP)[۴] [۷۰] یک راهکار نوین است که هرم تصویری را با استفاده از تصویر در مقیاس‌های مختلف در هر دو مرحله‌ی آموزش و تشخیص ایجاد می‌کند و تنها هزینه‌ی برخی از مقیاس‌های انتخاب شده را به عقب انتشار می‌دهد. بعدها، گروهی از محققان استراتژی آموزشی کارآمدتری را به جایSNIP پیشنهاد دادند که با عنوان نمونه‌برداری مجدد کارآمد (SNIPER) [۷۱] شناخته می‌شود. این استراتژی تصویر را به مجموعه‌ای از مناطق برش داده و دوباره مقیاس‌بندی می‌کند تا در آموزش از مجموعه‌داده‌ی غنی‌تری بهره‌مند شود.

باید توجه داشت که در مدل‌های تشخیص جسم مبتنی بر CNN، اندازه و نسبت ابعاد اجسام در محدوده‌ی مشخصی در نظر گرفته می‌شود که در صورت مواجهه شبکه با تغییرات مقیاس غیرمنتظره در اجسام، کارآیی آن در یادگیری و تشخیص این اجسام کاهش می‌یابد.مطالعات اخیر تکنیک‌هایی را برای بهبود تشخیص اجسام کوچک پیشنهاد کرده‌اند که از جمله‌ی آنها می‌توان به "بزرگنمایی تطبیقی" [۷۲] اشاره کرد که اجسام کوچک را به صورت تطبیقی بزرگ می‌کند. یکی دیگر از پیشرفت‌های انجام شده در این زمینه، پیش‌بینی توزیع مقیاس اجسام در یک تصویر و تغییر مقیاس تصویر مطابق با آن است [۷۳]. با این وجود به نظر می‌رسد که علیرغم تحقیقات انجام‌شده،یکی ازچالش‌های موجود در این زمینه وجود اجسام با اندازه‌ها و مقیاس‌های متفاوت در تصویر است.

۳-۷- یادگیری با تابع هزینه‌ی تقسیم بندی معنایی

مطالعات اخیر نشان می‌دهد که تشخیص جسم را می‌توان با تغییر تابع هزینه در شبکه و استفاده از تابع هزینه‌ی تقسیم‌بندی معنایی بهبود بخشید.

برای بهبود تشخیص با تقسیم بندی معنایی، ساده ترین راه این است که شبکه تقسیم‌بندی را به عنوان یک استخراج‌کننده ویژگی ثابت در نظر بگیریم و آن را به عنوان ویژگی‌های کمکی در یک مدل تشخیص جسم ادغام کنیم [۷۴]. اگرچه پیاده‌سازی این روش آسان است، اما ممکن است که استفاده از شبکه تقسیم‌بندی به صورت مجزا بار محاسبات اضافی را به همراه داشته باشد. راه دیگر معرفی یک شاخه‌ی تقسیم بندی اضافی در بالای مدل تشخیص جسم اصلی و آموزش این مدل با توابع هزینه ترکیبی است (ترکیب تابع هزینه تشخیص و تابع هزینه‌ی تقسیم بندی) [۷۵]. مزیت این راهکار این است که سرعت تشخیص تحت تأثیر آموزش یک شبکه‌ی اضافی قرار نمی‌گیرد. اما نکته‌ی مفید در آموزش این روش نیاز به برچسب گذاری تصاویر به صورت بسیار دقیق و در سطح پیکسل است.

---

[۱] Region of Interest
[۲] Back Propagation
[۳] Scale Imbalance
[۴] Scale Normalization for Image Pyramids

۴-۷- آموزش تخاصمی

شبکه تخاصمی مولد (GAN)[1]، معرفی شده توسط گودفیلو و همکاران [۷۶]، در بسیاری از کاربردها در پردازش تصویر مانند تولید تصویر، انتقال سبک تصویر، و بهبود وضوح تصویر مورد توجه قرار گرفته است. اخیراً از این شبکه برای کاربرد تشخیص جسم به‌ویژه برای بهبود تشخیص اجسام کوچک و مسدود شده[2]نیز استفاده شده است. برای تشخیص اجسام کوچک، GANمی‌تواند به عنوان یک شبکه‌ی افزایش‌دهنده‌ی رزولوشن استفاده شود تا با بزرگ کردن اجسام کوچک، استخراج ویژگی از آن‌ها به گونه‌ی مناسب‌تری انجام شود [۷۷]. همچنین برای بهبود تشخیص اجسام مسدود شده، یکی از ایده‌ها تولید تصاویری با این نوع از اجسام با استفاده از آموزش تخاصمی است [۷۸]. شبکه تخاصمی به جای آموزش برای تولید نمونه‌های اصلی از اجسام، مستقیماً برای تولید تصاویر دارای انسداد آموزش می‌بینند تا با افزایش این نمونه از تصاویر،یادگیری در روش تشخیص جسم را بهبود بخشد.

۵-۷- تشخیص جسم با نظارت ضعیف

آموزش یک مدل تشخیص جسم مبتنی بر یادگیری عمیق معمولاً به مقدار زیادی از داده‌های برچسب گذاری شده به صورت دستی نیاز دارد. هدف تشخیص جسم با نظارت ضعیف (WSOD)[3] کاهش اتکا به برچسب گذاری داده‌ها با آموزش مدل با برچسب گذاری در سطح تصویر به جای جعبه‌های مشخص کننده‌ی اجسام است [۷۹].

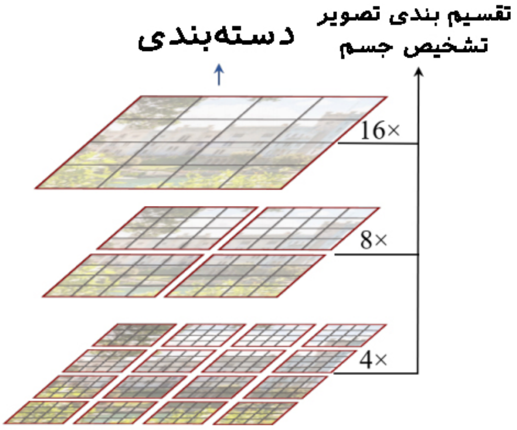
یکی از راهکارهایWSOD الگوریتم یادگیری چند نمونه‌ای است که کاربرد گسترده‌ای در WSOD داشته است [۸۰]. این الگوریتم‌ها به جای یادگیری با مجموعه‌ای از نمونه‌ها که به صورت خاص برچسب‌گذاری شده‌اند، مجموعه‌ای از داده‌های کلی مرتبط با هم را دریافت می‌کند که حاوی نمونه‌های زیادی است. اگر اجسام کاندیدا در یک تصویر را به عنوان مجموعه و برچسب‌ها در سطح تصویر را به عنوان برچسب اصلی در نظر بگیریم،WSOD را می‌توان به عنوان یک فرآیند یادگیری چند نمونه‌ای فرموله کرد.

نگاشت فعالسازی دسته، گروه دیگری از روش‌های اخیر برایWSOD است [۸۱]. تحقیقات رویCNN ها نشان داده است که لایه پیچشی یکCNN به عنوان یک آشکارساز جسم عمل می‌کند، علیرغم اینکه هیچ نظارتی بر محل جسم وجود ندارد. نگاشت فعال‌سازی دسته نشان می‌دهد که چگونه می‌توان با استفاده از برچسب گذاری سطح تصویر به قابلیت محاسبه مختصات اجسام بدون نظارت بر محل جسم دست یافت [۸۲].

۶-۷- تشخیص جسم با تطبیق دامنه

فرآیند آموزش بیشتر مدل‌های تشخیص جسم را می‌توان اساساً به عنوان یک فرآیند تخمین احتمال با فرض داده‌های مستقل با توزیع یکسان (i.i.d)[4]تعریف کرد. تشخیص جسم با داده‌های غیرi.i.d به ویژه برای برخی از مسائل کاربردی دنیای واقعی، هنوز یک چالش باقی مانده است.

از آنجایی که در کاربردهای تشخیص جسم ممکن است توزیع و دامنه‌ی اجسام در زمان آموزش و استفاده از مدل آموزش دیده کاملاً بر هم منطبق نباشد، از روش‌های تطبیق دامنه برای کاهش تفاوت بین دامنه‌ها استفاده می‌شود. برای به‌دست آوردن نمایش ویژگی بدون وابستگی به دامنه، روش‌های منظم‌سازی ویژگی و



شکل (۲۲): ترانسفورمر Swin [۸۹].

آموزش تخاصمی درسطوح تصویر، دسته یا جسم مورد بررسی قرارگرفته‌اند [۸۳ و ۸۴]. تبدیل با چرخه‌ی ثابت[5] [۸۵] نیز برای کاهش تفاوت بین دامنه منبع و هدف در تشخیص جسم استفاده شده است [۸۶].

۷-۷- ترانسفورمرهای بینایی برای تشخیص جسم

شبکه‌های عصبی ترانسفورمر [۸۷] از زمان معرفی تأثیر عمیقی در حوزه پردازش زبان طبیعی (NLP) داشته‌اند و کاربرد آنها در مدل‌های زبانی موفقی مانند GPT (ترانسفورمر از پیش آموزش‌دیده مولد) [۸۸]، علاقه به استفاده از آنها در بینایی ماشین را نیز برانگیخته است. در حالی که CNN ها همواره به عنوان شبکه‌های استخراج ویژگی یکی از بخش‌های اصلی مدل‌های تشخیص جسم بوده‌اند، اما دارای کاستی‌های ذاتی مانند عدم اهمیت به زمینه‌یادگیری در کل مدل، وزن‌های ثابت پس از آموزش و... هستند.

ترانسفورمرهای بینایی مانند Swin [۸۹] به دنبال ارائه یک شبکه استخراج ویژگی بر پایه ترانسفورمرها برای وظایف بینایی ماشین مانند تشخیص جسم هستند.همان‌طور که در شکل (۲۲)

[1] Generative Adversarial Network
[2] Occluded
[3] Weak Supervised Object Detection
[4] Independent and Identically Distributed
[5] Cycle Consistent Transformation

مشخص است، این شبکه تصاویر ورودی را به تکه[1]های متعدد و بدون همپوشانی تقسیم می‌کند.سپس تعداد زیادی بلوک‌های ترانسفورمر Swin در ۴ مرحله بر روی تکه‌های تصاویر اعمال می‌شود و در مرحله‌های متوالی تکه‌ها به هم متصل شده و تعداد آن‌ها کاهش می‌دهد تا نمایش سلسله مراتبی حفظ شود. بلوک ترانسفورمر Swin از ماژول‌های خودتوجهی استفاده کرده که فاصله‌ی بین بلوک‌های متوالی را مدیریت می‌کند [۸۹]. همچنین نشان می‌دهد که چگونه پنجره‌های جابجا شده دقت تشخیص را با سربار کمی افزایش می‌دهند. اگرچه که ترانسفورمرها یک تغییر قابل توجه در شبکه‌های عصبی مبتنی بر CNN را ارائه می‌دهند،کاربرد آن‌ها در مسئله‌ی تشخیص جسم هنوز در مراحل اولیه است و باید توجه داشت که ترنسفورمرها پارامترهای نسبتاً بیشتری نسبت به مدل‌های پیچشی عادی دارند و در نتیجه دارای سرعت محدودتری هستند.

مقالات متعددی از ترانسفورمرهای بینایی در مدل‌های تشخیص جسم استفاده کرده‌اند. در [۹۰] یک مدل ترکیبی متقابل به نام SwinNet(شکل ۲۲) برای تشخیص اجسام برجسته در تصاویرRGB-D و RGB-T پیشنهاد شد که از ترانسفورمر Swin برای استخراج ویژگی‌های سلسله مراتبی استفاده می‌کند. این مدل با مکانیسم توجه نیز تقویت می‌شود.و در مقایسه با دیگر مدل‌های پیشرفته در مجموعه داده‌هایRGB-D و RGB-T بهتر عمل می‌کند.در سال‌های اخیر برخی از محققان موفق شدند با استفاده از ترانسفورمرهای بینایی مدل‌های تشخیص جسم بر پایهYOLO را بهبود دهند.

در [۹۱] یک مدل YOLO بهبود یافته بر اساس ترانسفورمرهای بینایی معرفی شد که دارای یک شبکه استخراج ویژگی بهبود یافته به نام MHSA-Darknet است که برای حفظ و استخراج ویژگی‌های متمایزتر از طریق مکانیزم خود توجهی چند سر طراحی شده است. همچنین در [۹۲] مدل YOLOv5 با استفاده از ترانسفورمرهایSwin در قسمت تشخیص جسم بهبود یافته‌است. در این راهکار سرهای پیش‌بینی کانولوشن اصلی برای اولین بار با سرهای پیش‌بینی ترانسفورمر SPH جایگزین شده‌اند. SPH نشان دهنده یک مکانیسم پیشرفته خود توجهی است که طراحی پنجره‌ی تغییر یافته آن می‌تواند پیچیدگی محاسباتی را به طور خطی کاهش دهد.

۸-مروری بر تحقیقات جدید در زمینه‌ی تشخیص جسم

همان‌طور که پیش‌تر اشاره شد، تحقیقات برای ارائه راهکارهای تشخیص جسم به طور روز افزون ادامه دارد. در بخش‌های قبل به شبکه‌ها و مدل‌های پایه در تشخیص جسم از ابتدا تاکنون پرداخته شد. محققان بسیاری در سال‌های اخیر بر اساس مدل‌های پایه تشخیص جسم اقدام به توسعه و معرفی راهکارهای بهبود یافته و تکامل یافته کرده‌اند، به طوری که برخی از آن‌ها قابلیت‌های تشخیص جسم و مسائل دیگر بینایی ماشین از جمله بخش‌بندی تصویر را ترکیب کرده‌اند. در این بخش به مرور چند کار شاخص جدید در این زمینه می‌پردازیم.

در [۹۳] یک مدل جدید بر پایهYOLO با عملکرد بهتر و توسعه یافته با قابلیت تشخیص جسم و قطعه‌بندی آن به نام Poly-YOLO ارائه شده است. Poly-YOLO بر اساس ایده های اصلیYOLOv3 به‌وجود آمده‌است، اما دو نقطه ضعف آن را برطرف می کند: تعداد زیاد برچسب های بازنویسی شده و توزیع ناموازن جعبه‌های محور. Poly-YOLO با ترکیب نقشه‌های ویژگی با معرفی یک شبکه استخراج ویژگی سبک به نام SE-Darknet-53 و با استفاده از تکنیک افزایش وضوح نقشه‌های ویژگی، نواقص YOLOv3 را کاهش می‌دهد و خروجی با مقیاس پایدار و وضوح بالا تولید می‌کند. در مقایسه با YOLOv3 این مدل تنها ۶۰ درصد از پارامترهای قابل آموزش را دارد، اما میانگین دقت متوسط را تا ۴۰ درصد بهبود می‌بخشد. به علاوه، Poly-YOLO قابلیت قطعه‌بندی تصویر را با چندضلعی‌های مشخص‌کننده‌ی قطعه‌ها انجام می‌دهد. رئوس هر چندضلعی با اطمینان پیش‌بینی می‌شوند و بنابراین،Poly-YOLO چندضلعی‌هایی دقیق با تعداد رئوس متفاوت تولید می‌کند که قطعه‌های مختلف را مشخص می‌کنند.

برای دستیابی به راهکاری که امکان تشخیص و قطعه‌بندی جسم را به طور توام امکان پذیر می‌کند،Li و همکاران [۹۶] یک چارچوب یکپارچه تشخیص و قطعه‌بندی جسم به نام MaskDINO معرفی کردند که مدل، DINO[2][۹۷] را با افزودن یک شاخه پیش‌بینی ماسک گسترش می‌دهد و از یک نقشه‌ی پیکسلی با وضوح بالا برای پیش‌بینی مجموعه‌ای از ماسک‌های باینری استفاده می‌کند. در این روش، برخی از مؤلفه‌های کلیدی در DINO برای قطعه‌بندی از طریق یک معماری و فرآیند آموزشی مشترک توسعه یافته‌اند. MaskDINO ساده، کارآمد، و مقیاس‌پذیر است و از مجموعه داده‌های تشخیص و قطعه‌بندی در مقیاس بزرگ نیز استفاده کرده‌است. آزمایش‌ها نشان می‌دهد که MaskDINO به طور قابل‌توجهی از همه روش‌های قطعه‌بندی تخصصی موجود، هم با استفاده از شبکه استخراج ویژگیResNet-50 و هم در یک مدل از پیش آموزش‌دیدهبا شبکه SwinL بهتر عمل می‌کند. MaskDINO نتایج خوبی را در تشخیص جسم و انواع قطعه‌بندی (مانند قطعه بندی مبتنی بر جسم یا معنایی) در بین مدل‌های باکمتر از یک میلیارد پارامتر به دست آورده است.

[1] Patch

[2] Detr with improved denoising anchor boxes for end-to-end object detection

در [۹۸] مدل Cut-and-LEaRn(CutLER) پیشنهاد شده است که یک رویکرد ساده برای آموزش مدل‌های تشخیص و قطعه‌بندی جسم بدون نظارت است. در این تحقیق از ویژگی‌های مدل‌های خود نظارت برای "کشف" بدون نظارت اجسام و تقویت این دسته از مدل‌ها برای آموزش یک مدل موقعیت‌یابی جسم پیشرفته بدون هیچ‌گونه برچسب گذاری انسانی استفاده شده است. CutLER ابتدا از رویکرد جدیدی به نام MaskCut برای تولید ماسک‌هایی کلی برای چندین جسم در یک تصویر استفاده می‌کند و سپس با استفاده از یک تابع هزینه‌ی خلاقانه، یک آشکارساز را بر روی این ماسک‌ها می‌آموزد. در این راهکار عملکرد با خودآموزی مدل بر روی پیش‌بینی‌های آن بهبود می‌یابد. در مقایسه با کارهای قبلی،CutLER ساده‌تر است، با معماری‌های تشخیص مختلف سازگار است و چندین جسم را در تصویر تشخیص می‌دهد.

یکی دیگر از راهکارهای جدید و متفاوت بر پایهYOLO مدل Dist-YOLO [۹۴] است. این راهکار نشان می‌دهد که چگونه می‌توانYOLO را به منظور پیش‌بینی فاصله‌ی مطلق اجسام با استفاده از اطلاعات یک دوربین تک چشمی بهبود بخشید. با گسترش بردارهای پیش‌بینی، به اشتراک گذاشتن وزن های شبکه استخراج ویژگی با رگرسیور جعبه‌های مشخص‌کننده جسم، و به‌روزرسانی تابع هزینه اصلی، بخشی که مسئول تخمین فاصله است، به طور کامل در معماری اصلی ادغام شده است. در این راهکار دو روش برای به دست آوردن فاصله طراحی شده است، بدون توجه به دسته و با آگاهی از دسته‌ی جسم. روش بدون توجه به دسته بردارهای پیش‌بینی کوچکتری نسبت به روش آگاه از دسته ایجاد می‌کند و به نتایج بهتری می‌رسد. در نتیجه این مدل توانایی تشخیص جسم و اندازه‌گیری فاصله را به صورت توام دارد که منجر به افزایش دقت جعبه‌های مشخص‌کننده جسم نیز می‌شود. این روش بر روی مجموعه‌داده‌یKITTI [۹۵] توسعه داده و آزمایش شده است.

در سال‌های اخیر همچنین ارائه راهکارهای خلاقانه بر پایه‌ی مدل‌های دومرحله‌ایRCNN نیز انجام شده است . در [۹۹] یک مدل تشخیص اجسام زیر آب دو مرحله ای به نام R-CNN تقویت شده پیشنهاد شده که شامل سه جزء کلیدی است. ابتدا، یک شبکه برای پیشنهاد ناحیه‌ی جدید به نام RetinaRPN معرفی شده‌است که پیشنهادهای با کیفیت بالا ارائه می‌کند و دو عامل جسم بودن و پیش‌بینیIoU را برای مدل‌سازی موقعیت جسم در نظرمی‌گیرد. سپس، سلسله مراتب استنتاج احتمالی برای ترکیب احتمال موقعیت جسم از مرحله اول و امتیاز طبقه‌بندی جسم از مرحله دوم برای مدل سازی امتیاز تشخیص نهایی معرفی شده است. در نهایت،یک روش استخراج نمونه پیچیده و جدید به نام وزن‌دهیمجدد بهبودیافته پیشنهاد شده‌است که قابلیت اصلاح خطای مرحله‌ی اول را در طول مرحله استنتاج دارا است. آزمایش‌های جامع بر روی دو مجموعه داده زیر آب و دو مجموعه داده عمومی تشخیص جسم اثربخشی و پایدار بودن این راهکار را نشان می‌دهد.

در نهایت، در [۱۰۰] یک شبکه عصبی کانولوشن مبتنی بر ناحیه [۱۰۱] برای تشخیص اجسام سه بعدی در ابر نقاط[1]پیشنهاد شده‌است. این مدل پیشنهادی، عملکرد تشخیص سه‌بعدی را با ادغام یادگیری ویژگی انتزاعی در سطح پیکسل و کانولوشن مبتنی بر وکسل[2]افزایش می دهد و در آن یک چارچوب کارآمد و دقیق برای تشخیص اجسام سه بعدی پیشنهاد شده‌است. این چارچوب مبتنی بر دو ایده‌ی اصلی است: نمونه‌برداری پیشنهاد-محور برای تولید مؤثر نقاط کلیدی، و مکانیسمی برای تجمیع بهتر ویژگی‌های محلی با مصرف منابع بسیار کمتر. آزمایش‌ها نشان می‌دهند که چارچوب پیشنهادیPV-RCNN++ به عملکرد تشخیص سه بعدی پیشرفته در مجموعه داده باز Waymo در مقیاس بزرگ و بسیار رقابتی دست یافته‌است.

در این بخش به چند راهکار تشخیص جسم جدید و خلاقانه پرداخته شد. مرور این راهکارهای شاخص نشان می‌دهد حل مسائل متفاوت تشخیص جسم و حل توام آنها با مسائل دیگر با معرفی مدل‌های پیشرفته در حال حاضر مورد توجه محققان است و نویدبخش راهکارهای کامل‌تر در آینده است.

۹-نتایج و بحث

۱-۹- معیارهای ارزیابی

مدل‌های تشخیص جسم معمولاً با معیارهای مختلفی مانند دقت تشخیص، سرعت استنتاج[3]، بهینگی، و پیچیدگی محاسباتی مورد ارزیابی قرار می‌گیرند که در این بخش به معرفی آنها می‌پردازیم.

۱-۱-۹-معیارهای ارزیابی دقت تشخیص

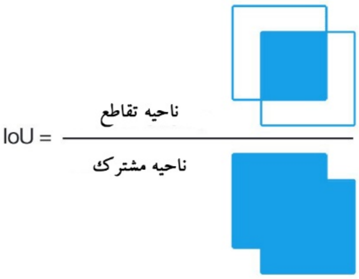
خروجی مدل‌های تشخیص جسم معمولاً شامل مجموعه‌ای از جعبه‌های مشخص کننده‌ی جسم پیش‌بینی‌شده، سطوح اطمینان و دسته‌های پیش‌بینی‌شده است. در چالش‌ها، رقابت‌ها و اغلب مجموعه‌داده‌های تشخیص جسم، از تصاویر واقعی که در آنها مشخصات اجسام از قبل مشخص شده است، استفاده می‌شود. دقت تشخیص مدل‌ها با مقایسه موقعیت جعبه‌های پیش‌بینی شده برای اجسام و دسته‌بندی آنها نسبت به مشخصات حقیقی اجسام در تصاویر مجموعه‌ی ارزیابی به دست می‌آید.

یکی از مهم‌ترین معیارهای موثر در ارزیابی دقت تشخیص مدل‌های تشخیص جسم IoU[4]است. IoU میزان هم‌پوشانی بین

---

[1] Point Cloud
[2] Voxel
[3] Inference
[4] Intersection Over Union

دو جسم است که با یکدیگر تقاطع یا اشتراک دارند. این معیار به صورت نسبت اشتراک یا تقاطع بین مساحت جعبه‌ی پیش‌بینی شده و مساحت جعبه‌ی اصلی جسم به مساحت اجتماع بین این دو جعبه‌تعریف می‌شود (شکل (۲۳)).



شکل (۲۳): فرمول کلی IoU برای تشخیص میزان مطابقت جعبه‌ی پیش‌بینی شده برای جسم با جعبه‌ی واقعی.

بیشترین مقدار IoUیک است که مشابهت کامل بین جعبه‌ی پیش‌بینی شده و جعبه‌ی واقعی را نشان می‌دهد وIoU صفر به معنی عدم تطابق بین دو جعبه است.

با وجود اینکه معرفی تابع IoU پیشرفتی قابل توجه در زمینه‌ی تشخیص جسم مبتنی بر یادگیری عمیق بود، این تابع دارای محدودیت‌هایی است. مشکل اصلی تابع IoU این است که اگر جعبه‌ی پیش‌بینی شده برای جسم با جعبه‌ی واقعی هیچ اشتراکی نداشته باشد، حاصل این تابع صفر می‌شود که روند یادگیری را با اختلال مواجه می‌کند.در نتیجه به مرور نسخه‌های بهبود یافته‌ی این تابع مهم معرفی شدند. به منظور رفع مشکل ذکر شده، تابع GIoU [۱۰۲] معرفی شد که با افزودن معیار کوچکترین جعبه‌ای که هر دو جعبه پیش‌بینی شده و واقعی را در بر می‌گیرد، تا حدودی رگرسیون جعبه را بهبود داد، اما مشکل آن این است که همچنان ابعاد اجسام را در نظر نمی‌گیرد. در نتیجه در سال‌های اخیر تابع CIoU معرفی شدکه اشتراک جعبه پیش‌بینی و واقعی، فاصله مرکز دو جعبه و نسبت ابعاد دو جعبه را در نظر می‌گیرد و بهبود قابل توجهی نسبت به IoU عادی و GIoU ایجاد می‌کند. فرمول این تابع به صورت (۱) است:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b_{gt})}{c^2} + \beta v \quad (۱)$$

به‌طوری‌که b و b_{gt}نقاط مرکز جعبه‌های پیش‌بینی شده و جعبه واقعی جسم هستند،ρفاصله اقلیدسی این دو مرکز است، c طول مورب کوچکترین جعبه محصور که دو جعبه را در بر می‌گیرد است،

βیک ضریب مثبت است و v ثابت بودن نسبت طول و عرض در دو جعبه را اندازه می‌گیرد.

به طور کلی دقت مدل به درست بودن یا اشتباه بودن یک پیش‌بینی بستگی دارد. در تشخیص جسم این درستی با توجه به IoU و یک آستانه که می‌تواند عددی بین۰ تا ۱ باشد و بر اساس معیارهایTP و FP و FN محاسبه می‌شود:

- TP[۱]: یک پیش‌بینی درست، در صورتی که IOU بزرگتر مساوی سطح آستانه‌ی تعیین‌شده باشد.
- FP[۲]: یک پیش‌بینی نادرست از جسمی که وجود ندارد یا تشخیص یک جسم با IoU کمتر از آستانه‌ی تعیین‌شده.
- FN[۳]: یک جسم در تصویر واقعی که هیچ جعبه‌ای برای آن پیش‌بینی نشده است.

نکته قابل توجه در این زمینه این است که در تشخیص جسم معیارTN[۴]استفاده نمی‌شود، زیرا ممکن است در تصویر تعداد زیادی جعبه‌ی پیش‌بینی وجود داشته باشد که جسمی در آن‌ها وجود ندارد و نباید توسط مدل پیش‌بینی شوند.براساس این تعاریف، مدل‌هایتشخیص جسم توسط معیارهای دقت[۵] و یادآوری[۶]ارزیابی می‌شوند که به ترتیب با P و R (رابطه‌ی(۲)) نمایش داده می‌شوند. معیارPمشخص می‌کند که از بین اجسامی که مدل آن‌ها را پیش‌بینی کرده است، چه تعدادی به درستی پیش‌بینی شده‌اند (یعنی نسبت تعداد پیش‌بینی‌های مثبت و درست به کل پیش‌بینی‌های مثبت انجام شده). معیارR برابر با نسبت اجسامی است که توسط مدل به درستی پیش‌بینی شده‌اند، به تعداد کل اجسامی که در تصویر وجود دارد.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \quad (۲)$$

یکی دیگر از معیارهای مهم در ارزیابی مدل‌های تشخیص جسم، معیار mAP[۷] (۳) است:

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \quad (۳)$$

که در آن N تعداد دسته‌های اجسام است. به طور کلیAP میانگینP در مقادیر مختلف R بین ۰ و ۱ است و mAP یعنی متوسط AP همه‌ی دسته‌ها. برای تشخیص درست بودن پیش‌بینی یک جسم، یک سطح آستانه در IoU درنظر گرفته می‌شود که معمولاً بین ۰/۵ و ۰/۹۵ است و با قدم‌های ۰/۰۵ تایی افزایش می‌یابد.

۲-۱-۹-معیارهای ارزیابی پیچیدگی محاسباتی و سرعت مدل

بسیاری از مسائل تشخیص جسم در دنیای واقعی نیاز به سرعت استنتاج مناسب دارند. هرچند معمولاً مدل‌های پیچیده‌تر دارای دقت بالاتری هستند، اما محاسبات و حافظه بیشتری نیاز دارند.

---

[۱] True Positive
[۲] False Positive
[۳] False Negative
[۴] True Negative
[۵] Precision
[۶] Recall

یکی از معیارهای مهم برای ارزیابی سرعت مدل‌های تشخیص جسم، معیار فریم بر ثانیه (FPS) است که بر اساس زمان متوسط استنتاج (۴) به دست میاید. زمان متوسط استنتاج، میانگین زمان پردازش هر تصویر از N تصویر یا فریم است که معمولاً با واحد میلی‌ثانیه اندازه‌گیری می‌شود.

$$\text{زمان متوسط استنتاج} = \frac{\sum_{i=1}^{N} T_i}{N} \quad (۴)$$

به طوری که N تعداد تصاویر یا فریم‌ها است و T_i زمان پردازش تصویریا فریم i ام است. معیار فریم بر ثانیه تعداد فریم‌هایی است که در یک ثانیه یعنی ۱۰۰۰ میلی ثانیه پردازش می‌شوند.

۹-۲-مقایسه‌ی کارآیی روش‌های موجود

پس از بیان معیارهای ارزیابی برای سنجش کارآیی مدل‌های تشخیص جسم، در این قسمت، عملکرد تعدادی از مدل‌های ذکر شده یک مرحله‌ای و دو مرحله‌ای را بر روی مجموعه داده‌های[۱۲PASCAL VOC] و [۱۳MS-COCO]ارزیابی و مقایسه می‌کنیم.این دو مجموعه‌داده معمولاً به عنوان مجموعه معیار برای بررسی عملکرد کلی مدل‌ها استفاده می شوند. عملکرد مدل‌های تشخیص جسم تحت تأثیر تعدادی از عوامل مانند مقیاس ورودی، شبکه استخراج ویژگی، روش آموزش، تابع هزینه و... است.

جدول‌های(۳) و (۴) نتایج گردآوری شده چند مدل تشخیص جسم ذکر شده در این مقاله را به ترتیب دومرحله‌ایایک مرحله‌ای بودن و بر اساس نتایج حاصل از مقالات آن‌ها نشان می‌دهند. طبق نتایج گردآوری شده در جدول (۳)، مدل‌های دومرحله‌ای مثل خانواده RCNN و SPPNet به معیار mAP بالاتری نسبت به نخستین مدل YOLO دست یافته‌اند، درحالی که معیارFPS برای مدل YOLO بسیار بیشتر است. همچنین طبق نتایج گردآوری شده در جدول (۴) مدل دومرحله‌ایFPN دارای معیارهایmAP بالاتر نسبت به مدل‌هاییک مرحله‌ای ابتدایی مانند SSD و YOLOv2است، اما مدل‌های دو مرحله‌ای یا یک مرحله‌ای خلاقانه و پیچیده مانند RetinaNet و CenterNet برخلاف مدل‌های با پیچیدگی کمتر نمی‌توانند به سرعت بی‌درنگ که برای برخی از کاربردها ضروری است، دست یابند.همچنین علی‌رغم ضعف و دقت پایین مدل‌هاییک مرحله‌ای ابتدایی، پیشرفت روزافزون موجب شده است که این مدل‌ها به سرعت تشخیص بالاتر از مدل‌های دیگر و دقت مناسب دست پیدا کنند، اما باید توجه داشت که انتخاب مدل مناسب به کاربرد و مسئله هدف بستگی دارد.

به طور کلی، نتایج نشان می‌دهد که مدل‌های تشخیص جسم در طول سال‌ها پیشرفت قابل‌توجهی داشته‌اند، به طوری که آخرین مدل‌ها به معیارهایmAP بالاتر و FPS سریع‌تر که از عوامل ضروری برای کاربردهای بی‌درنگ هستند دست یافته‌اند که این نویدبخش آینده‌ی امیدوارکننده در زمینه‌ی تشخیص جسم است.

۱۰- آینده‌ی تشخیص جسم

در سال‌های اخیر دستاوردهای قابل توجهی در زمینه تشخیص جسم حاصل شده است. برخی از نوآوری‌های امیدوارکننده برای آینده‌ی تحقیق در این زمینه می‌توانند شامل موارد زیر باشد، هرچند که گستره‌ی تحقیق به این موارد محدود نمی‌شود. مرور اینموارد به ما کمک می‌کند تا بینش بیشتری درباره آینده تشخیص جسم کسب کنیم:

پیشرفت مدل‌های سبک:هدف اصلی مدل‌های سبک اجرا در دستگاه‌های قابل حمل یا کم مصرف است. برخی از برنامه‌های کاربردی مهم برای مدل‌های سبک عبارتند از: اینترنت اشیا، واقعیت افزوده موبایل، رانندگی خودکار، شهر هوشمند، دوربین‌های هوشمند و... . اگرچه تلاش‌های زیادی در سال‌های اخیر در این زمینه انجام شده است، شکاف سرعت بین یک ماشین و چشم انسان همچنان زیاد است، به ویژه برای تشخیص برخی از اجسام کوچک یا شناسایی با اطلاعات چند منبعی [۱۰۳ و ۱۰۴].

بهبود تشخیص اجسام کوچک: تشخیص اجسام کوچک در صحنه‌های وسیع مدت‌ها است که یک چالش بوده است. برخی از کاربردهای بالقوه در این زمینه تحقیقاتی شامل شمارش جمعیت افراد یا تعداد حیوانات در مناطق حفاظت شده و شناسایی اهداف از تصاویر ماهواره‌ای است. برخی از کاربردها هم ممکن است شامل ادغام مکانیسم‌های توجه بصری و طراحی شبکه‌های سبک وزن با وضوح بالا باشند [۱۰۵ و ۱۰۶].

تشخیص اجسام سه بعدی: علیرغم پیشرفت‌های اخیر در تشخیص اجسام دو بعدی، کاربردهایی مانند رانندگی خودکار به دسترسی به مکان اجسام و حالات آن‌ها در دنیای سه‌بعدی متکی هستند. تشخیص جسم در دنیای سه بعدی و استفاده از داده‌های چندمنبعی و چند نمایی (به عنوان مثال، تصاویرRGB و نقاط LiDAR سه بعدی از چندین سنسور) یکی از چالش‌های تحقیقاتی است که می‌تواند مورد توجه بیشتری قرار بگیرد [۱۰۷ و ۱۰۸].

تشخیص جسم در ویدیو:تشخیص و ردیابی جسم به طور بی‌درنگ در ویدئوهای با وضوح بالا برای نظارت تصویری و رانندگی خودکار اهمیت زیادی دارد. مدل‌های تشخیص جسم مرسوم معمولاً برای تشخیص در تصاویر طراحی شده‌اند و همبستگی بین فریم‌های ویدیو را نادیده می‌گیرند. بهبود تشخیص و ردیابی جسم در ویدیو با بررسی همبستگی مکانی و زمانی تحت محدودیت محاسباتی یک چالش تحقیقاتی مهم است که می‌تواند در آینده بیشتر مورد توجه قرار بگیرد [۱۰۹ و ۱۱۰].

تشخیص جهان باز:تعمیم خارج از دامنه، تشخیص صفر شات، و تشخیص تدریجی موضوعات در حال ظهور در زمینه تشخیص

جسم هستند. انسان‌ها غریزه‌ی کشف اجسام ناشناخته در محیط را دارند. وقتی برچسب هر جسم مشخص می‌شود، انسان‌ها دانش جدیدی از آن می‌آموزند و الگوها را حفظ می‌کنند. با این حال، درک و تشخیص دسته‌های ناشناخته اجسام برای الگوریتم‌های تشخیص جسم فعلی دشوار است. هدف تشخیص جسم جهان باز کشف دسته‌های ناشناخته از اجسام است که در برنامه‌هایی مانند روباتیک و رانندگی خودران کاربردی است [۱۱۳ و ۱۱۴].

۱۱-جمع‌بندی و نتیجه‌گیری

در این مقاله اهمیت تشخیص جسم بر پایه شبکه‌های عصبی پیچشی و چالش‌های پیش روی آن بررسی شد و مجموعه داده‌های مهم، شبکه‌ها و مدل‌های تشخیص جسم بر پایهCNN و نوآوری‌های اخیر برای بهبود در این زمینه مرور شد و مورد بحث قرار گرفت. به علاوه، نتایج ارزیابی مدل‌های تشخیص جسم مطرح با معیارهای ارزیابی متفاوت مورد بحث و بررسی قرار گرفت. در نهایت زمینه‌های تحقیقاتی و نوآوری‌هایی محتمل در این زمینه معرفی شدند. همه‌ی مواردی که در این مقاله به آنها پرداخته شد می‌توانند به تحقیق و پیشرفت روز افزون در زمینه تشخیص جسم و بینایی ماشین کمک کنند.

با وجود پیشرفت قابل توجه مدل‌های تشخیص جسم بر پایه CNN در سال‌های اخیر، حتی برترین مدل‌ها هنوز به دقت و عملکرد بدون نقص در مسائل تشخیص جسم مختلف دست نیافته‌اند. با افزایش کاربردهای تشخیص جسم در مسائل دنیای واقعی نیاز به مدل‌های بهتر که بتوانند با سرعت مناسب اجسام متفاوت را تشخیص دهند در حال افزایش است. در این مقاله، نشان داده شد که چگونه مدل‌های تشخیص جسم دو مرحله‌ای و یک مرحله‌ای به مرور توسعه یافتند. در حالی که عموماً مدل‌های دو مرحله‌ای در ابتدا دقیق‌تر بودند، اما برای کاربردهای بی‌درنگ مانند خودروهای خودران کند هستند. با این حال، این مورد در چند سال گذشته تغییر کرده است، به طوری که مدل‌های یک مرحله‌ای همچون جدیدترین مدل‌های مبتنی بر YOLO در کنار داشتن دقت بالا، بسیار سریع هستند. با روند مثبت فعلی در عملکرد مدل‌های تشخیص جسم و نوآوری‌های محتمل ذکر شده، امید زیادی برای توسعه مدل‌های تشخیص جسم دقیق‌تر و سریع‌تر برای کاربردهای مختلف وجود دارد.

جدول (۳): جدول مقایسه مدل‌های تشخیص جسم آموزش داده شده با مجموعه داده PASCAL VOC

مدل	سال	شبکه استخراج ویژگی	$mAP_{0.5}$	$mAP_{[0.5:0.95]}$	FPS
RCNN [۸]	۲۰۱۴	AlexNet	۵۸/۵٪	-	۰/۰۲
SPPNet [۲۸]	۲۰۱۵	ZF-5	۵۹/۲٪	-	۰/۲۳
Fast-RCNN [۲۹]	۲۰۱۵	VGG-16	۶۵/۷٪	-	۰/۴۳
Faster-RCNN [۳۰]	۲۰۱۵	VGG-16	۶۷٪	-	۵
A-Fast-RCNN [۷۸]	۲۰۱۷	VGG-16	۶۹٪	-	-
YOLO [۹]	۲۰۱۶	GoogleNet	۵۷/۹٪	-	۴۵

جدول (۴): جدول مقایسه مدل‌های تشخیص جسم آموزش داده شده با مجموعه داده COCO

مدل	سال	شبکه استخراج ویژگی	$mAP_{0.5}$	$mAP_{[0.5:0.95]}$	FPS
FPN [۳۳]	۲۰۱۷	ResNet-101	۵۹/۱٪	۳۶/۲٪	۵
SSD [۳۴]	۲۰۱۶	VGG-16	۴۱/۲٪	۲۳/۲٪	۴۶
RetinaNet [۳۷]	۲۰۱۷	ResNet-101-FPN	۵۸/۵٪	۳۱/۹٪	۱۲
CenterNet [۳۸]	۲۰۱۹	Hourglass-104	۶۱/۱٪	۴۲/۱٪	۷/۸
YOLOv2 [۳۹]	۲۰۱۷	DarkNet-19	۴۴٪	۲۱/۶٪	۸۱
YOLOv3 [۴۰]	۲۰۱۸	DarkNet-53	۵۱/۵٪	۲۸/۲٪	۴۵
YOLOv4 [۴۲]	۲۰۲۰	CSPDarkNet-53	۶۴/۹٪	۴۳٪	۳۱
YOLOv5 [۴۵]	۲۰۲۰	CSPv7	۶۸/۸٪	۵۰/۴٪	۶۲
YOLOR [۴۷]	۲۰۲۱	CSPDarknet	۷۳/۳٪	۵۵/۴٪	۳۰
YOLOX [۴۸]	۲۰۲۱	CSPv5	۶۹/۶٪	۵۱/۲٪	۵۷
YOLOv6 [۴۹]	۲۰۲۲	EfficientRep	۷۰٪	۵۲/۵٪	۴۹
YOLOv7 [۵۰]	۲۰۲۲	RepConvN	۷۳/۵٪	۵۵/۹٪	۵۰

مراجع

[1] R. Ezhilarasi and P. Varalakshmi, "Tumor detection in the brain using faster r-cnn," in *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on*, pp. 388–392, IEEE, 2018.

[2] S. H. Naghavi, C. Avaznia, and H. Talebi, "Integrated real-time object detection for self-driving vehicles," in *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)*, pp. 154–158, IEEE, 2017.

[3] K. J. Liang, J. B. Sigman, G. P. Spell, D. Strellis, W. Chang, F. Liu, T. Mehta, and L. Carin, "Toward automatic threat recognition for airport x-ray baggage screening with deep convolutional object detection," *arXiv preprint arXiv:1912.06329*, 2019.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, Ieee, 2001.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, Ieee, 2005.

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, Ieee, 2008.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[8] R. Girshick, J. Donahue, T. Darrell, and J. 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.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[10] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. PietikÃ¤inen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020.

[11] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, p. 103514, 2022.

[12] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[14] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[15] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 590–597, 2019.

[16] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.

[17] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

[18] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018.

[19] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

[20] IBM, "What are convolutional neural networks?," 2020. https://www.ibm.com/topics/convolutional-neural-networks, Accessed on September 5th, 2023.

[21] M. Hemmer, H. Van Khang, K. G. Robbersmyr, T. I. Waag, and T. J. Meyer, "Fault classification of axial and radial roller bearings using transfer learning through a pretrained convolutional neural network," *Designs*, vol. 2, no. 4, p. 56, 2018.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645, Springer, 2016.

[26] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[27] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 390–391, 2020.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[29] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[31] G. Yin, M. Yu, M. Wang, Y. Hu, and Y. Zhang, "Research on highway vehicle detection based on faster r-cnn and domain adaptation," *Applied Intelligence*, vol. 52, no. 4, pp. 3483–3498, 2022.

[32] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14454–14463, 2021.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[34] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[35] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and feature fusion ssd for remote sensing object detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.

[36] Z. Chen, H. Guo, J. Yang, H. Jiao, Z. Feng, L. Chen, and T. Gao, "Fast vehicle detection algorithm in traffic scene based on improved ssd," *Measurement*, vol. 201, p. 111655, 2022.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[38] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569–6578, 2019.

[39] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

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

[41] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[42] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[43] D. Misra, "Mish: A self regularized non-monotonic neural activation function," *arXiv preprint arXiv:1908.08681*, vol. 4, no. 2, pp. 10–48550, 2019.

[44] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, 2018.

[45] A. H. L. Y. c. P. R. Glenn Jocher, Liu Changyu, "ultralytics/yolov5: Initial release," June 2020.

[46] M. H. Hamzenejadi and H. Mohseni, "Real-time vehicle detection and classification in uav imagery using improved yolov5," in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 231–236, IEEE, 2022.

[47] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint arXiv:2105.04206*, 2021.

[48] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[49] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[50] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

conference on computer vision (ECCV), pp. 734–750, 2018.

[64] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229, Springer, 2020.

[65] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 3735–3739, IEEE, 2015.

[66] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.

[67] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 265–278, 2018.

[68] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2849–2858, 2019.

[69] B. Cai, Z. Jiang, H. Zhang, Y. Yao, and S. Nie, "Online exemplar-based fully convolutional network for aircraft detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 7, pp. 1095–1099, 2018.

[70] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3578–3587, 2018.

[71] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," *Advances in neural information processing systems*, vol. 31, 2018.

[72] Y. Lu, T. Javidi, and S. Lazebnik, "Adaptive object detection using adjacency and zoom prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2351–2359, 2016.

[73] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scale-aware face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6186–6195, 2017.

[74] A. Shrivastava and A. Gupta, "Contextual priming and feedback for faster r-cnn," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 330–348, Springer, 2016.

[51] G. Jocher and A. Chaurasia, "Ultralytics/yolov8 in pytorch." https://github.com/ultralytics/ultralytics, 2023. Accessed: March 24, 2023.

[52] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13733–13742, 2021.

[53] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," *arXiv preprint arXiv:2211.04800*, 2022.

[54] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5325–5334, 2015.

[55] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 443–457, Springer, 2016.

[56] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[57] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[58] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[59] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[60] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[61] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

[62] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

[63] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European*

[75] S. Brahmbhatt, H. I. Christensen, and J. Hays, “Stuffnet: Using â€˜stuffâ€™to improve object detection,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 934–943, IEEE, 2017.

[76] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[77] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, “Sod-mtgan: Small object detection via multi-task generative adversarial network,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 206–221, 2018.

[78] X. Wang, A. Shrivastava, and A. Gupta, “A-fast-rcnn: Hard positive generation via adversary for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2606–2615, 2017.

[79] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, “Weakly supervised object localization and detection: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5866–5885, 2021.

[80] D. Zhang, J. Han, L. Zhao, and D. Meng, “Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework,” *International Journal of Computer Vision*, vol. 127, pp. 363–380, 2019.

[81] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, “Weakly supervised cascaded convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 914–922, 2017.

[82] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[83] Y. Wang, R. Zhang, S. Zhang, M. Li, Y. Xia, X. Zhang, and S. Liu, “Domain-specific suppression for adaptive object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9603–9612, 2021.

[84] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, “Exploring categorical regularization for domain adaptive object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11724–11733, 2020.

[85] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[86] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, “Progressive domain adaptation for object detection,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 749–757, 2020.

[87] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[88] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.

[89] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[90] Z. Liu, Y. Tan, Q. He, and Y. Xiao, “Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4486–4497, 2021.

[91] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, “Vit-yolo: Transformer-based yolo for object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2799–2808, 2021.

[92] H. Gong, T. Mu, Q. Li, H. Dai, C. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, H. Li, *et al.*, “Swin-transformer-enabled yolov5 with attention mechanism for small object detection on satellite images,” *Remote Sensing*, vol. 14, no. 12, p. 2861, 2022.

[93] P. Hurtik, V. Molek, J. Hula, M. Vajgl, P. Vlasanek, and T. Nejezchleba, “Poly-yolo: higher speed, more precise detection and instance segmentation for yolov3,” *Neural Computing and Applications*, vol. 34, no. 10, pp. 8275–8290, 2022.

[94] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, “Mask dino: Towards a unified transformer-based framework for object detection and segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023.

[95] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, “Dino: Detr with improved denoising anchor boxes for end-to-end object detection,” *arXiv preprint arXiv:2203.03605*, 2022.

[96] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, “Cut and learn for unsupervised object detection and instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3124–3134, 2023.

[97] M. Vajgl, P. Hurtik, and T. Nejezchleba, “Dist-yolo: fast object detection with distance estimation,” *Applied sciences*, vol. 12, no. 3, p. 1354, 2022.

[98] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *International Journal of Robotics Research (IJRR)*, 2013.

[99] P. Song, P. Li, L. Dai, T. Wang, and Z. Chen, “Boosting r-cnn: Reweighting r-cnn samples by rpnâ€™s error for underwater object detection,” *Neurocomputing*, vol. 530, pp. 150–164, 2023.

[100] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, “Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection,” *International Journal of Computer Vision*, vol. 131, no. 2, pp. 531–551, 2023.

[101] Z. Liu, H. Tang, Y. Lin, and S. Han, “Point-voxel cnn for efficient 3d deep learning,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[102] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

[103] B. Bosquet, M. Mucientes, and V. M. Brea, “Stdnet-st: Spatio-temporal convnet for small object detection,” *Pattern Recognition*, vol. 116, p. 107929, 2021.

[104] C. Yang, Z. Huang, and N. Wang, “Querydet: Cascaded sparse query for accelerating high-resolution small object detection,” in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 13668–13677, 2022.

[105] X. Zhou, X. Xu, W. Liang, Z. Zeng, S. Shimizu, L. T. Yang, and Q. Jin, “Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1377–1386, 2021.

[106] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, and J. Han, “Towards large-scale small object detection: Survey and benchmarks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[107] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, “Detr3d: 3d object detection from multi-view images via 3d-to-2d queries,” in *Conference on Robot Learning*, pp. 180–191, PMLR, 2022.

[108] Y. Wang, T. Ye, L. Cao, W. Huang, F. Sun, F. He, and D. Tao, “Bridged transformer for vision and point cloud 3d object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12114–12123, 2022.

[109] X. Cheng, H. Xiong, D.-P. Fan, Y. Zhong, M. Harandi, T. Drummond, and Z. Ge, “Implicit motion handling for video camouflaged object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13864–13873, 2022.

[110] L. He, Q. Zhou, X. Li, L. Niu, G. Cheng, X. Li, W. Liu, Y. Tong, L. Ma, and L. Zhang, “End-to-end video object detection with spatial-temporal transformers,” in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1507–1516, 2021.

[111] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, “Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection,” *IEEE Transactions on Image Processing*, vol. 31, pp. 6800–6815, 2022.

[112] F. Qingyun and W. Zhaokui, “Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery,” *Pattern Recognition*, vol. 130, p. 108786, 2022.

[113] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, “Promptdet: Towards open-vocabulary detection using uncurated images,” in *European Conference on Computer Vision*, pp. 701–717, Springer, 2022.

[114] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, “Ow-detr: Open-world detection transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9235–9244, 2022.

محمدحسین حمزه‌نژادی مدرک کارشناسی خود را در سال ۱۳۹۹ از موسسه آموزش عالی کرمان و مدرک کارشناسی ارشد خود را در سال ۱۴۰۲ در رشته مهندسی کامپیوتر گرایش نرم افزار از دانشگاه شهید باهنر دریافت کرده است. زمینه پژوهشی مورد علاقه وی بینایی ماشین، هوش مصنوعی و توسعه نرم افزار می باشد.

حدیث محسنی مدرک کارشناسی خود را در سال ۱۳۸۳ از دانشگاه صنعتی شریف در رشته مهندسی کامپیوتر گرایش سخت افزار و مدرک کارشناسی ارشد و دکترای خود را به ترتیب در سال های ۱۳۸۶ و ۱۳۹۲ از دانشگاه صنعتی شریف در رشته هوش مصنوعی اخذ نمود. ایشان هم اکنون دانشیار گروه هوش مصنوعی بخش مهندسی کامپیوتر دانشگاه شهید باهنر کرمان هستند و زمینه پژوهشی ایشان بازشناسی الگو آماری، یادگیری و شبکه های عمیق، پردازش تصاویر پزشکی و پردازش تصویر و ویدیو می‌باشد.