ارائه روشی نوین برای توصیف ناحیه مورد علاقه در استاندارد کدگذاری ویدئو برای ماشین‌ها

بهار خدیوی بروجنی[۱] و هدی رودکی[۲]

چکیده

با گسترش کاربردهای تحلیل ویدئو در کاربردهایی مانند نظارت ویدئویی و خودروهای خودران، نیاز به روش‌های کارآمد برای کاهش پیچیدگی پردازش و فشرده‌سازی داده بیش از پیش اهمیت یافته است. در کدگذاری ویدئویی برای ماشین‌ها برخلاف استانداردهای سنتی، تمرکز بر انتقال کارآمد اطلاعات معنادار برای تحلیل ماشینی است. چالش اصلی در این حوزه، پیچیدگی بالای شبکه‌های تشخیص اشیاء و هزینه سنگین پردازش آن‌ها در دستگاه‌های بلادرنگ است. در این پژوهش، یک نسخه ساده‌سازی‌شده از YOLOv۸ ارائه می‌شود که با بهره‌گیری از چنددقتی‌سازی محاسبات، هرس وزنی و تقطیر دانش، پیچیدگی مدل را بدون افت محسوس عملکرد کاهش می‌دهد. همچنین یک ساختار کدگذار/کدگشا مبتنی بر تولید سه جریان ایجاد شده است تا نرخ بیت ورودی در مرحله کدگذاری کاهش یابد. نتایج تجربی نشان می‌دهد که روش پیشنهادی زمان تاخیر در تشخیص اشیاء را تا ۵۰ درصد کاهش می‌دهد درحالی که، دقت را تنها ۰/۰۵ درصد تحت‌تأثیر قرار می‌دهد و امکان پردازش بلادرنگ ویدئو را فراهم می‌سازد. افزون بر این، تقسیم‌بندی جریان ورودی سبب کاهش نرخ بیت بدون افت محسوس در دقت تشخیص می‌شود.نوآوری اصلی این کار، ترکیب ساده‌سازی ساختاری شبکه با معماری جدید کدگذاری سه‌جریانی است که به‌طور هم‌زمان موجب کاهش پیچیدگی محاسباتی و بهبود فشرده‌سازی در سناریوهای بلادرنگ می‌شود.

کلید واژه ها

کدگذاری ویدئویی برای ماشین‌ها، فشرده‌سازی ویدئو، تشخیص اشیاء، YOLOv۸ ، فراداده، نواحی مورد علاقه

۱- مقدمه

ویدئو به‌عنوان یکی از پرکاربردترین داده‌ها در دنیای امروز، نقشی کلیدی در حوزه‌های متنوعی نظیر نظارت ویدئویی، حمل‌ونقل هوشمند، خودروهای خودران، شهر هوشمند و سامانه‌های امنیتی ایفا می‌کند[۱]. در کاربردهای عملی مانند خودروهای خودران و سیستم‌های نظارت هوشمند، وظایفی مانند دنبال کردن اشیا، جداسازی اشیا و تشخیص اشیا نقش حیاتی دارند. در خودروهای خودران بیش از ۷۰ درصد داده‌های ادراکی از دوربین‌ها حاصل می‌شود. در سامانه‌های نظارت شهری نیز، بیش از ۸۰ درصد داده‌های تولیدشده از نوع ویدئویی هستند. در خودروهای خودران، تشخیص و جداسازی دقیق وسایل نقلیه، عابران، علائم راهنمایی و موانع به خودرو امکان می‌دهد مسیر حرکت را به‌صورت ایمن برنامه‌ریزی کرده و از برخورد جلوگیری کند، در حالی که دنبال کردن اشیا امکان پیش‌بینی حرکت آن‌ها را فراهم می‌کند. در نظارت هوشمند نیز، تشخیص و جداسازی افراد و اشیا به سیستم اجازه می‌دهد مناطق حساس را شناسایی کرده و رفتار غیرعادی را تحلیل کند. دنبال کردن اشیا نیز در این کاربرد برای

مقاله در تاریخ ۱۲ مهر ماه ۱۴۰۴ دریافت شد.

[۱] کارشناسی ارشد، گروه اینترنت اشیا، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران
رایانامه: b.khadiviborojeni@email.kntu.ac.ir

[۲] استادیار، گروه اینترنت اشیا، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی ، تهران، ایران
رایانامه: hroodaki@kntu.ac.ir

نویسنده مسئول: هدی رودکی

ردیابی سوژه‌ها در فضاهای شلوغ یا پهنه‌های وسیع ضروری است. وقتی این وظایف با کدگذاری ویدئویی برای ماشین[۱] ترکیب می‌شوند، داده‌های غیرضروری حذف شده و سیستم می‌تواند با حجم کمتر و سرعت بالاتر، عملکرد دقیقی در زمان واقعی ارائه دهد. در اغلب این کاربردها، ویدئو نه‌تنها برای مشاهده انسانی بلکه بیشتر برای تحلیل ماشینی ایجاد و مصرف می‌شود. به همین دلیل در کنار کدگذاری ویدئویی برای انسان، حوزه‌ای جدیدی تحت عنوان کدگذاری ویدئویی برای ماشین‌ها شکل گرفته است [۲]. علاوه بر این گاهی نیاز است بازسازی ویدئو برای بینایی ماشین و انسان به صورت هم‌زمان انجام شود[۳]. در برخی کاربردها، ویدئو باید هم برای تحلیل ماشین و هم برای کاربر انسانی بازسازی شود. طراحی الگوریتمی که بتواند هم‌زمان نیازهای بینایی ماشین و انسان را برآورده سازد، پیچیده است[٤] [٥]. در کدگذاری سنتی، هدف بازسازی باکیفیت برای انسان است، اما در کدگذاری ویدئویی برای ماشین‌ها، خروجی ویدئو بیشتر برای مدل‌های بینایی ماشین مصرف می‌شود و کیفیت ادراکی نقش ثانویه را دارد. برخلاف روش‌های سنتی که حفظ کیفیت بصری و معیارهایی مثل نسبت سیگنال به نویز[۲] و معیار شاخص شباهت ساختاری[۳] مهم است، در کدگذاری ویدئویی برای ماشین‌ها عملکرد وظایفی مانند تشخیص و طبقه‌بندی پس از فشرده‌سازی معیار اصلی است. بنابراین معیارهایی مثل میانگین دقت[٤] و صحت[٥] استفاده می‌شوند و فقط ویژگی‌های مهم برای مدل‌ها حفظ می‌شود. در این روش ممکن است کیفیت بصری کاهش یابد، اما نرخ بیت بسیار کم و عملکرد وظیفه‌ای حفظ می‌شود، و حتی می‌توان اطلاعات سطح بالا مانند ویژگی‌ها یا نقشه‌های توجه را ذخیره کرد. مهم‌ترین چالش در این استاندارد، بالابودن پیچیدگی محاسباتی در کدگذار و کدگشا است. در کدگذاری ویدئویی برای ماشین‌ها این پیچیدگی ناشی از نیاز به استخراج، فشرده‌سازی و بازسازی اطلاعات نواحی مورد علاقه و ویژگی‌های مورد نیاز برای تحلیل ماشین است[٦]. برخلاف کدگذاری سنتی که تنها بر بازسازی ویدئو برای انسان متمرکز است. در بسیاری از پژوهش‌های اخیر در حوزه کدگذاری ویدئویی برای ماشین‌ها، برای انجام وظیفه تشخیص اشیا از مدل‌های مبتنی بر YOLO استفاده می‌شود، زیرا این مدل‌ها در تشخیص سریع و دقیق اشیا عملکرد بسیار مناسبی دارند[۸]. با این حال، پیچیدگی بالای YOLOv۸ مانعی برای به‌کارگیری مستقیم آن در یک چارچوب کدگذاری کم‌هزینه و بلادرنگ است. نوآوری اصلی این پژوهش در بهره‌گیری از روش‌های ساده‌سازی مدل YOLOv۸ و ترکیب آن با ارسال فراداده در کنار تعریف الگوریتم کدگذار و کدگشا جدید است. این روش علاوه بر کاهش حجم داده، امکان بازسازی دقیق‌تر در سمت گیرنده و بهبود عملکرد تشخیص اشیاء را برای ماشین فراهم می‌کند. به طور کلی نوآوری‌های این مقاله عبارتند از:

۱. ارائه چارچوبی جدید برای کدگذاری ویدئویی در ماشین‌ها با سرعت و دقت بالا.
۲. اجرای روش‌های ساده‌سازی بر روی الگوریتم YOLO به منظور کاهش پیچیدگی محاسباتی جهت تشخیص اشیاء.
۳. حفظ تعادل بین سرعت و دقت در ساده سازی مدل
٤. استفاده از یک کلاس فراداده[٦] و ارسال آن همراه با جریان بیت اصلی به سمت کدگشا جهت کمک.

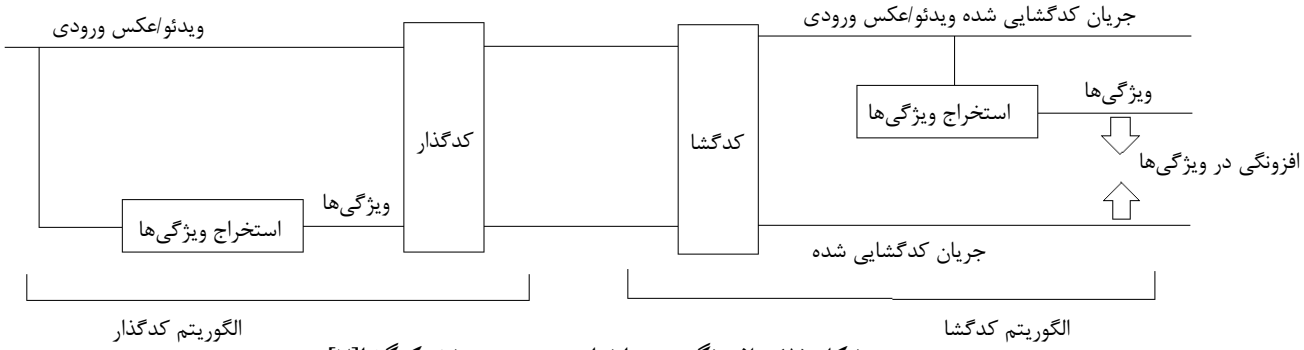
۲- مروری بر کارهای پیشین

هدف اصلی کدگذاری ویدئو آن است که به‌جای تمرکز صرف بر کیفیت بصری برای انسان، اطلاعات معنادار و ضروری برای تحلیل ماشین حفظ شود. در این راستا، روش‌های متعددی معرفی شده‌اند که در ادامه به بررسی مهم‌ترین آن‌ها پرداخته می‌شود.

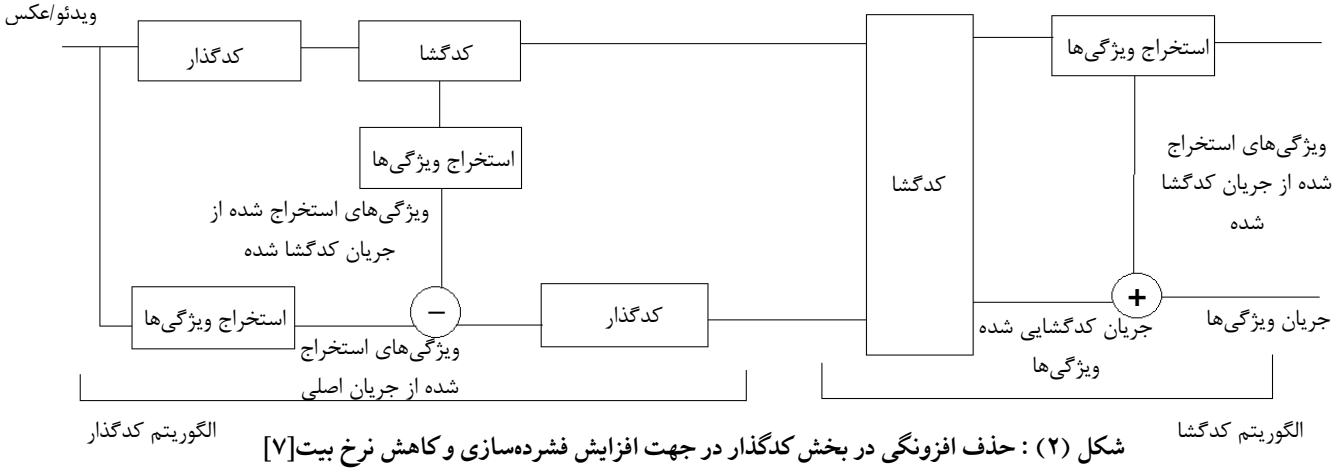
۲-۱- فشرده‌سازی جریان ارسالی مبتنی بر حذف افزونگی

هدف رویکرد پیشنهادی با در نظر گرفتن نیازهای انسان و ماشین در هنگام کدگذاری ویدئو، کاهش نرخ بیت ارسالی با حذف افزونگی است. همان‌طور که در ساختار ابتدایی کدگذاری ویدوئویی ماشین در شکل (۱) نشان داده شد، ورودی به دو بخش تقسیم می‌شود. بخش اول ویدئو به صورت عادی کد می‌شود و برای ماشین جهت تحلیل ارسال می‌شود. در بخش دوم، ویژگی‌های ویدئو که شامل اطلاعات مورد نیاز ماشین است استخراج می‌شود و داده های اضافی حذف می‌شوند. ارسال ویدئو و ویژگی‌های ویدئو به صورت دو جریان داده باعث ایجاد افزونگی داده می‌شود، زیرا ویدئو کدگشایی شده این قابلیت را دارد تا خود ویژگی‌ها را استخراج کند[۷]. برای حل این مشکل، شکل (۲) ساختاری را برای حذف افزونگی داده پیشنهاد می‌کند. در این رویکرد با استخراج ویژگی از ویدئو کدگذاری شده و سپس کدگشایی شده و سپس کسر آن‌ها، تنها داده‌های باقیمانده از جریان اصلی و ویژگی‌های استخراج شده ارسال می‌شود. در بخش کدگشایی، ویژگی‌های استخراج شده از ویدئو بازسازی شده با جریان داده ویژگی‌ها ترکیب می‌شود. در این رویکرد برای بخش بینایی ماشین، تمرکز بر انتقال نقاط کلیدی است که به طور گسترده برای کارهایی مانند تشخیص، ردیابی و بازسازی اشیاء استفاده می‌شوند. رویکرد پیشنهادی می‌تواند در طیف وسیعی از کاربردها، از جمله وسایل نقلیه خودران، روباتیک و نظارت مفید باشد[۷]، اما این رویکرد تنها بر حذف افزونگی تمرکز داشته و فشرده‌سازی

[۱] Video Coding for Machine
[۲] Peak Signal-to-Noise Ratio (PSNR)
[۳] Structural Similarity Index Measure (SSIM)
[٤] Mean-Average Precision(mAP)
[٥] Accuracy
[٦] Metadata



شکل (۱) : افزونگی در ساختار عمومی در بخش کدگشا[۷]



شکل (۲) : حذف افزونگی در بخش کدگذار در جهت افزایش فشرده‌سازی و کاهش نرخ بیت[۷]

بیشتری در جهت کاهش نرخ بیت ارسالی ارائه نمی‌دهد. به علاوه، اگر شیء در زمان کدگذاری به اشتباه تشخیص داده نشود چون به کدگشا ارسال نمی‌شود در زمان کدگشایی هم آن شیء برای ماشین قابل تشخیص نخواهد بود.

۲-۲- ساده‌سازی قاب‌ها براساس نواحی مورد علاقه[۱]

هدف اصلی این رویکرد استفاده از روش کدگذاری ویدئویی مبتنی بر نواحی مورد علاقه است تا بخش‌های مهم تصویر برای تحلیل ماشین با دقت بالاتر حفظ شوند و در عین حال نرخ بیت کاهش یابد. این رویکرد امکان فشرده‌سازی کارآمد و حفظ اطلاعات کلیدی برای وظایف بینایی ماشین را فراهم می‌کند. ایده‌ی اصلی این روش طراحی یک کدگذار ویدئویی برای ماشین‌ها است که بر پایه‌ی کدگذار/کدگشاهای موجود عمل می‌کند و شامل چند بلوک کلیدی است[۹]:

- تشخیص و طبقه‌بندی نواحی مهم
- ساده‌سازی که در آن جزئیات کاهش داده می‌شوند تا نرخ بیت پایین بیاید.
- کاهش خطوط و ستون‌های تصویر در نواحی غیرمهم برای کاهش بیشتر حجم داده
- ارسال فراداده به‌عنوان اطلاعات جانبی برای همگام‌سازی و بازسازی دقیق در سمت کدگشا

در سمت کدگشا، فرایند هدف‌گیری معکوس[۲] برای بازسازی تصویر اصلی و اشیای حذف‌شده انجام می‌شود تا بینایی ماشینی و بینایی انسانی پشتیبانی شوند[۱۱]. محدودیت‌های این رویکرد شامل موارد زیر است: تمرکز صرف روی نواحی مورد علاقه که می‌تواند منجر به از دست رفتن اطلاعاتی در سایر بخش‌های تصویر شود که ممکن است برای برخی کاربردهای مربوط به انسان مهم باشند. همچنین پیاده‌سازی و بهینه‌سازی الگوریتم برای ویدئوهای با صحنه‌های پویا ممکن است چالش‌برانگیز باشد و نیاز به منابع محاسباتی بالاتر داشته باشد.

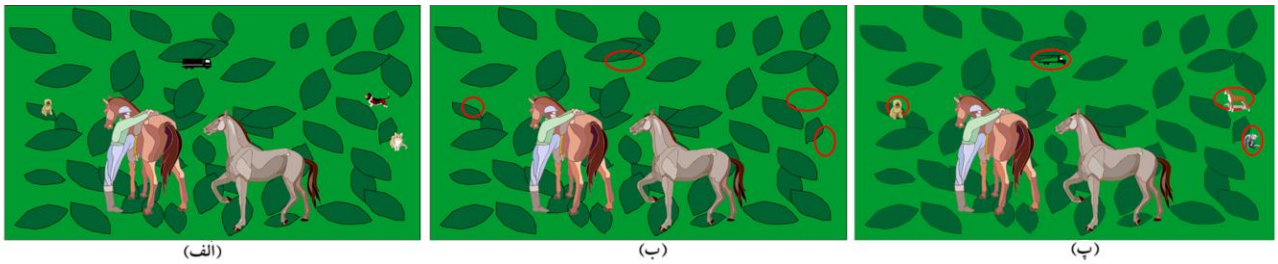
۳-۲- ساده‌سازی و رنگ‌آمیزی[۳] قاب‌ها بر اساس نواحی مورد علاقه

تمرکز رویکرد قبلی بر روی ساده‌سازی و تغییر ابعاد هدفمند بود. این رویکرد با افزودن بلوک رنگ آمیزی و معرفی سه جریان فراداده رویکردی پیشرفته‌تر ارائه داد و در حقیقت تکمیل کننده رویکرد قبلی است. در این طرح اشیای کم‌اهمیت یا تکراری در مرحله کدگذاری حذف و با محتوای ساده جایگزین می‌شوند و سپس اطلاعات آن‌ها به‌صورت فراداده به کدگشا منتقل می‌گردد تا بازسازی شوند[۱۰]. این تفاوت موجب می‌شود این رویکرد علاوه بر کاهش حجم داده، توانایی بالاتری در حفظ توازن بین کیفیت ادراکی برای انسان و کارایی الگوریتم‌های بینایی ماشین ارائه دهد.

[۱] Region of Interest (RoI)

[۲] Inverse Retargeting

[۳] Inpainting



شکل (۳) : فرآیند رنگ آمیزی مجدد برای یک تصویر ورودی نمونه. (الف): تصویر ورودی. (ب): اشیاء غیر مهم حذف و اطلاعات آن‌ها به صورت فراداده برای بازسازی ارسال می‌شود. (پ): بازسازی تصویر و اشیاء غیر مهم. [۱۰]

در شکل (۳) فرآیند رنگ‌آمیزی و بازسازی در سه مرحله نشان داده شده است[۱۰]:

۱. تصویر اصلی (تصویر ابتدایی از سمت چپ): تصویر ورودی شامل همه‌ی اشیا.
۲. فرآیند رنگ آمیزی (تصویر وسط): در این مرحله، بعضی اشیا که توسط الگوریتم تشخیص اشیاء به‌عنوان کم‌اهمیت یا تکراری تشخیص داده شده مثلاً یک سگ کوچک یا ماشین، از تصویر حذف می‌شوند و به جای آن‌ها ، پس‌زمینه‌ی ساده گذاشته می‌شود.
۳. بازسازی در کدگشا (تصویر انتهایی از سمت چپ): وقتی بازسازی ویدئو انجام شد، دستگاه با استفاده از فراداده که شامل اطلاعات اشیای حذف‌شده است، آن اشیاء را دوباره بازسازی می‌کند. البته بازسازی دقیق نیست و ممکن است فقط نمایشی تقریبی از شیء باشد.

در این رویکرد وابستگی شدید به تشخیص دقیق نواحی مورد علاقه می‌تواند باعث افت عملکرد در ویدئوهای پیچیده شود به علاوه که در جهت افزایش سرعت مدل تشخیص اشیاء بهبودی انجام نگرفته است. علاوه بر این، تقسیم جریان‌ها به نواحی مورد علاقه و پس‌زمینه نیازمند محاسبات اضافه است که ممکن است سرعت پردازش را کاهش دهد، پس محاسبه جریان فراداده نباید باعث افزایش محاسبات و پیچیدگی شود.

٤-۲- فشرده‌سازی ویدئو مبتنی بر نواحی مورد علاقه

ایده و هدف اصلی این رویکرد این است که به جای ارسال کل قاب ویدئو، تنها نتایج پیش‌بینی شبکه‌ی تشخیص اشیا به صورت یک توصیف‌گر نواحی مهم[۱] (شامل مختصات، کلاس و نمره اطمینان) استخراج، فشرده‌سازی و منتقل شوند. این توصیف‌گرها سپس می‌توانند برای وظایفی مانند ردیابی اشیاء استفاده شوند، بدون آنکه نیاز به پردازش دوباره‌ی کل تصویر باشد. نتایج آزمایش‌ها روی مجموعه داده TVD[۲] [۱۲] نشان داد که روش پیشنهادی توانسته نرخ بیت را بیش از ۵۰ درصد نسبت به VVC[۳] کاهش دهد و در عین حال معیار MOTA[٤] در ردیابی اشیاء را در سطح قابل قبولی حفظ یا حتی بهبود بخشد. همچنین تکنیک‌های تکمیلی مانند انتخاب رنگ‌های پس‌زمینه مناسب، افزایش حاشیه اطراف اشیاء، گروه‌بندی نواحی مجاور و روش‌های مختلف انباشت قاب‌ها مورد بررسی قرار گرفتند که هر یک بهبودهای جزئی در دقت یا کارایی فشرده‌سازی ایجاد کردند. در مجموع، این کار نشان می‌دهد که استفاده از توصیف‌گرها می‌تواند راهکاری مؤثر برای کاهش حجم داده و حفظ عملکرد الگوریتم‌های یادگیری ماشین در کاربردهای ویدئویی باشد[۱۳].

در کنار مزایای کاهش نرخ بیت و تمرکز بر نواحی مهم تصویر، روش‌های مبتنی بر نواحی مورد علاقه با محدودیت‌هایی نیز همراه هستند. از جمله اینکه استخراج دقیق ناحیه‌ی مورد علاقه وابسته به الگوریتم تشخیص شیء است و خطای آن مستقیماً بر کیفیت بازسازی و عملکرد بینایی ماشین اثر می‌گذارد و در رابطه با ساده‌سازی یا افزایش سرعت تشخیص و کاهش حجم مدل استفاده شده فرآیندی صورت نگرفته است. به علاوه در جهت بازسازی ویدئو برای کاربرد انسان فرآیندی پیاده‌سازی نشده است.

٥-۲- ساده‌سازی مبتنی بر هرس ساختاری و نیمه‌ساختاری

یکی از مهم‌ترین چالش‌های مدل‌های تشخیص اشیاء در کاربردهای بلادرنگ، پیچیدگی محاسباتی بالا و نیاز به توان پردازشی بالا است. این امر منجر شده تا بخش قابل‌توجهی از پژوهش‌ها بر ساده‌سازی این مدل‌ها بدون ایجاد افت محسوس در دقت تمرکز کنند. در این پژوهش، ایده اصلی این روش این است که به‌جای حذف کامل برخی وزن‌ها یا ایجاد پراکندگی تصادفی در وزن‌ها، وزن‌های شبکه در قالب بلوک‌های کوچک و ساختاریافته (بلوک‌های ۱×٤ یا ۲×٤) دسته‌بندی شوند و کوچک‌ترین وزن‌ها در این بلوک‌ها حذف می‌شود تا پراکندگی به‌صورت کنترل‌شده ایجاد شود. این نوع هرس دارای دو مزیت کلیدی است[١٤]:

[۱] RoI Descriptor
[۲] Tencent Video Dataset
[۳] Versatile Video Coding
[٤] Multi-Object Tracking Accuracy

۱. ساختار تنسورهای [۱] شبکه به‌گونه‌ای حفظ می‌شود که پردازشگرهای گرافیکی قادر به اسفاده از موازی‌سازی باشند.
۲. سربار ناشی از تنسورهای پراکنده [۲] در معماری‌های سخت‌افزاری کاهش می‌یابد.

از نظر عملکرد، نتایج نشان دادند که این روش قادر است مدل‌های YOLO را تا ۴/۴ برابر فشرده‌تر کند و زمان استنتاج را در پلتفرم‌های کم‌منبعی مانند Jetson TX۲ به مقدار ۲/۱۵ برابر بهبود بخشد، در حالی‌که افت دقت مدل کمتر از ۱ درصد باقی بماند[۱۴]. هرچند در این پژوهش و پژوهش بعدی، هدف از ساده‌سازی مدل مربوط به کدگذاری ویدئویی برای ماشین‌ها نیست و نتایج آن‌ها برروی کدگذار ویدئویی برای ماشین‌ها مشخص نیست.

۶-۲- ساده‌سازی مبتنی بر بهینه‌سازی معماری و تقطیر دانش[۳]

در محیط‌های زیرآبی به دلیل وجود نویزهای شدید، کاهش شفافیت، پراکندگی نور و محدودیت توان پردازشی، از چالش‌برانگیزترین کاربردها برای تشخیص اشیاء محسوب می‌شوند. در این پژوهش تلاش شده تا با بهره‌گیری از جستجوی معماری شبکه[۴]، ساختار بهینه انتخاب شود؛ ساختاری که ضمن کاهش پارامترها، قادر باشد ویژگی‌های مهم را در شرایط تصویری سخت استخراج نماید. جستجوی معماری شبکه عملا فضایی از معماری‌های مختلف را جستجو کرده و بهترین ترتیب و تعداد لایه‌ها را متناسب با محدودیت‌های سخت‌افزاری انتخاب می‌کند. علاوه بر این، تقطیر دانش سبب شده تا مدل کوچکتر YOLO بتواند رفتار مدل بزرگ‌تر YOLO را تقلید کرده و بخش بزرگی از دقت مدل معلم را حفظ کند[۱۵].

نتایج این پژوهش نشان داد که ترکیب جستجوی معماری شبکه و تقطیر دانش موفق شده است حافظه مصرفی مدل را کاهش داده و در عین حال دقت تشخیص را در محیط‌های زیرآبی افزایش دهد که نشان‌دهنده اثربخشی آن در کاربردهای واقعی است. این دستاورد مهم است زیرا ثابت می‌کند که حتی در شرایطی که کیفیت تصویر به‌شدت کاهش می‌یابد، می‌توان با طراحی ساختار مناسب و هدایت یادگیری مدل، نسخه‌های سبک YOLO را به‌گونه‌ای آموزش داد که عملکردی مشابه یا حتی بهتر از نسخه‌های اصلی ارائه دهند.

بر اساس تمامی موارد بررسی شده، با وجود پیشرفت‌های صورت‌گرفته در زمینه کدگذاری ویدئویی و تشخیص اشیاء، هنوز خلاهای پژوهشی مهمی وجود دارد. بسیاری از روش‌های موجود یا بر سرعت و یا بر دقت تمرکز دارند و چارچوب یکپارچه‌ای که بتواند هم‌زمان سرعت و دقت بالا را در کدگذاری ویدئو برای ماشین‌ها فراهم کند، به اندازه کافی مورد بررسی قرار نگرفته است. همچنین الگوریتم‌های تشخیص اشیاء مانند YOLO در نسخه‌های اصلی خود پیچیدگی محاسباتی بالایی دارند و روش‌های ساده‌سازی برای کاهش این پیچیدگی بدون افت قابل توجه دقت، هنوز به طور جامع ارائه نشده است. علاوه بر این، سیستم‌های کدگذاری ویدئو اغلب تنها به جریان بیت اصلی توجه دارند و استفاده عملی از فراداده برای بهبود عملکرد کدگشا کمتر مورد مطالعه قرار گرفته است. بنابراین، نیاز به پژوهش‌هایی که بتوانند با ساده‌سازی الگوریتم‌ها، تعادل سرعت و دقت را حفظ کرده و با بهره‌گیری هوشمندانه از فراداده، کارایی سیستم‌های کدگذاری ویدئویی را بهبود دهند، همچنان محسوس است.

۳- روش پیشنهادی

در این بخش، رویکرد پیشنهادی در این مقاله با جزئیات شرح داده می‌شود. شکل (٤) روند کلی انجام الگوریتم پیشنهادی را نشان می‌دهد. ابتدا به منظور کاهش پیچیدگی محاسباتی و نیز افزایش کارایی، قاب‌های ویدئویی نمونه‌برداری کاهشی می‌شوند. سپس فرآیند ساده‌سازی شبکه برای تشخیص اشیا انجام می‌شود. در ادامه نواحی مورد علاقه و نواحی پس‌زمینه جداسازی شده و با نرخ بیت‌های متفاوت کد می‌شوند. جزئیات هر یک از این بخش‌ها در ادامه توضیح داده می‌شود.

۱-۳- نمونه برداری کاهشی[۵]

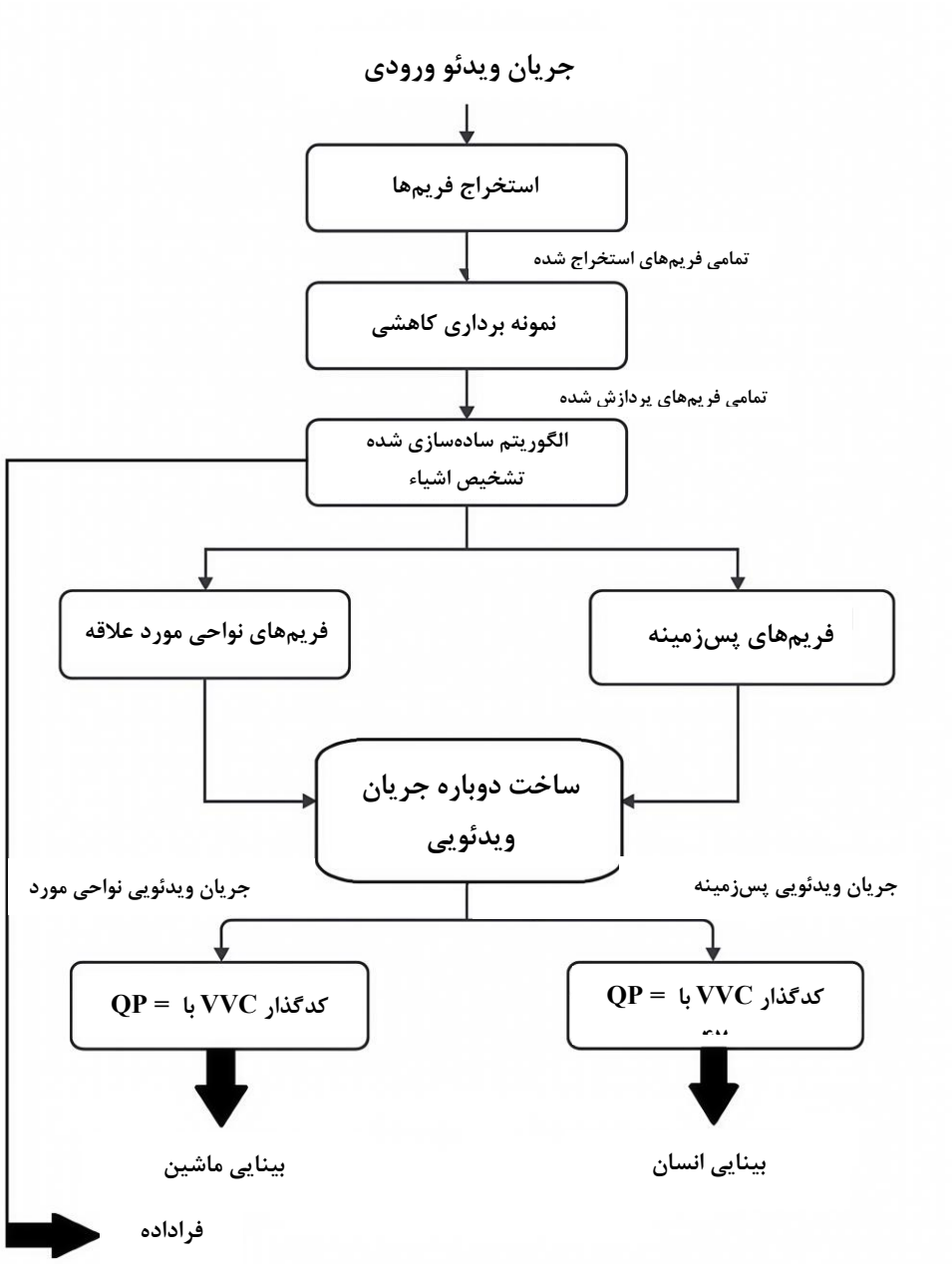
پیچیدگی محاسباتی بالا و کارایی فشرده‌سازی نیز یکی از نیازهای اساس در استاندارد کدگذار/کدگشا ویدئو برای ماشین است، بنابراین حجم داده ارسالی باید کنترل شود. به‌منظور پاسخ به این نیاز، در روش پیشنهادی از نمونه برداری کاهشی استفاده شده است. این فرآیند با کاهش ابعاد تصاویر ورودی، حجم داده‌ها را کاهش می‌دهد و در نتیجه، نیاز به پهنای باند و نرخ بیت کمتری در مرحله‌ی کدگذاری خواهد بود. علاوه بر این، کاهش ابعاد ورودی منجر به سبک‌تر شدن محاسبات شده و به بهبود سرعت اجرا و کاهش مصرف منابع سخت‌افزاری کمک خواهد کرد. این مرحله از کاهش ابعاد به‌عنوان یکی از گام‌ها در فرآیند کدگذار در نظر گرفته شده است. دلیل انتخاب این روش، توانایی آن در کاهش چشم‌گیر حجم داده و در نتیجه کاستن از نرخ بیت بدون افت محسوس در دقت تشخیص اشیاء توسط مدل است.

روش‌های مختلفی برای نمونه‌برداری کاهشی وجود دارد اما روش استفاده شده در این پژوهش، روش تغییر اندازه[۶] نسبی است. در این فرآیند، قاب‌های استخراج‌شده از جریان ویدئو به‌صورت

---

[۱] Tensor
[۲] Sparse Tensor
[۳] Knowledge Distillation
[۴] Neural Architecture Search

[۵] Down Sampling
[۶] Resizing



شکل (٤) : ساختار کدگذار پیشنهادی

فریم‌به‌فریم پردازش شده و ابعاد مکانی هر قاب بر اساس یک نسبت مشخص کاهش می‌یابد. در پیاده‌سازی حاضر، عرض و ارتفاع هر قاب به ۴۰ درصد اندازه اصلی مقیاس‌گذاری شده است. این مقدار پس از آزمایش مقیاس‌های مختلف انتخاب شده است تا ضمن کاهش چشمگیر حجم داده، افت محسوسی در دقت تشخیص اشیاء ایجاد نشود. برای انجام عملیات بازنمونه‌برداری، از فیلتر LANCZOS استفاده شده است که به‌عنوان یکی از فیلترهای بازسازی باکیفیت در پردازش تصویر شناخته می‌شود و توانایی حفظ جزئیات مهم در هنگام کاهش ابعاد را دارد [١٦]. فرآیند کلی به شرح زیر است:

١. بارگذاری قاب‌ها
٢. اعمال کاهش ابعاد مکانی
٣. به‌کارگیری فیلتر LANCZOS
٤. ذخیره‌سازی بهینه‌شده
٥. استفاده از پردازش چندریسمانی[1]

در آزمایش‌های انجام‌شده، مقیاس‌گذاری به ۴۰ درصد ابعاد اصلی بهترین توازن میان کاهش حجم و حفظ کیفیت اطلاعات محتوایی را ایجاد کرده است، به‌گونه‌ای که هم فضای ذخیره‌سازی و هم زمان پردازش کاهش یافته و در عین حال خروجی برای شناسایی اشیاء توسط YOLO همچنان قابل اتکا باقی مانده است. برای انتخاب نسبت مناسب کاهش ابعاد، چندین مقیاس مورد آزمایش قرار گرفتند که نتایج این مقایسه در جدول (١) آورده شده است.

[1] Multi Thread

جدول (۱) : مقایسه نتایج درصد کاهش ابعاد در نمونه برداری کاهشی براساس دقت مدل و زمان پردازش

مقیاس (درصد)	میانگین دقت مدل	تاخیر برای هر فریم(ثانیه)
۲۰	۰/۸۷۰۵	۰/۹۵
۳۰	۰/۸۷۰۱	۰/۹۱
٤٥	۰/۸٦۸۲	۰/۸٤
٥٥	۰/۸٦۳٦	۰/۸۱
٦۰	۰/۸٦۰۲	۰/۷٦
٦٥	۰/۸٤٦۱	۰/۷۲
۷۰	۰/۸۰۹۷	۰/٦۷

نتایج نشان داد که کاهش بیش از ۴۰ درصد باعث افت محسوس در تشخیص اشیاء کوچک می‌شود، در حالی که مقیاس ۴۰ درصد بهترین توازن میان کاهش حجم داده و حفظ دقت تشخیص را ایجاد می‌کند و نتایج نشان دهنده کاهش ۰/۰۱ دقت تشخیص اشیاء بوده است. هرچند در کاربردهایی که اشیاء در آن از ابعاد کوچکتری برخوردار هستند و نیازمند به حفظ دقت بیشتر است می‌توان ابعاد این کاهش را کمتر کرد تا تشخیص اشیاء تحت تاثیر قرار نگیرد. اما کاهش ابعاد بیش از این مقدار باعث افت محسوس دقت در تشخیص اشیای کوچک می‌شود و این موضوع با اهداف کاربردی ما سازگار نیست. به همین دلیل، مقیاس حفظ ابعاد تا ٤۰ درصد به‌عنوان بهترین توازن میان کاهش حجم داده، افزایش سرعت پردازش و حفظ دقت تشخیص، مخصوصاً در نواحی مورد علاقه انتخاب شد. علاوه بر این، برای کاهش اثرات منفی احتمالی ناشی از کوچک‌سازی تصویر، از فیلتر LANCZOS استفاده شده است که یکی از فیلترهای پیشرفته بازنمونه‌برداری به شمار می‌رود و جزئیات ساختاری و لبه‌های مهم تصویر را تا حد زیادی حفظ می‌کند. استفاده از این فیلتر نقش مهمی در جلوگیری از کاهش دقت YOLO هنگام اندازه‌گذاری مجدد ورودی داشته است. بدین ترتیب، این مرحله نه‌تنها یک عملیات پیش‌پردازش ساده، بلکه بخشی ضروری از جریان کدگذاری محسوب می‌شود که زمینه‌ساز دستیابی به کارایی بالاتر در مراحل بعدی مانند فشرده‌سازی، انتقال و تحلیل است. هدف اصلی کاهش پیچیدگی محاسباتی، افزایش سرعت پردازش بلادرنگ و کاهش حجم داده در کدگذاری ویدئویی برای ماشین‌ها است.

در ادامه روشی برای کاهش پیچیدگی شبکه YOLOv۸ جهت تشخیص اشیا و افزایش فشرده سازی جریان ورودی در فرآیند کدگذاری ویدئویی برای ماشین‌ها ارائه شد. هدف اصلی این روش، دستیابی به یک الگوریتم بهینه است که بتواند در کنار رفع نیازهای تحلیل ماشین و دید انسان، با کاهش پیچیدگی در تشخیص اشیاء و حفظ دقت، جریان ویدئویی ورودی را با هدف کاهش پهنای باند فشرده کند و بتواند به صورت بلادرنگ جریان ورودی را کدگذاری کند و برای کدگذار ارسال نماید.

۲-۳- ساده سازی YOLOv۸ به سه روش

در رویکرد پیشنهادی هدف این است که فرآیند تشخیص اشیاء با محاسبات کمتر، حافظه مصرفی کمتر و توان مصرفی کمتر، بدون افت محسوس در دقت انجام شود، چرا که نسخه‌های بزرگ YOLO مصرف حافظه و منابع محاسباتی زیادی دارند؛ پس باید مدل را کوچک‌تر و سریع‌تر کرد. در این پژوهش به‌منظور ساده‌سازی مدل YOLOv۸ و کاهش مصرف منابع محاسباتی، سه رویکرد اصلی به‌کار گرفته شده است:

چند دقتی[۱]: پیاده‌سازی این روش به‌شرح زیر است:
شناسایی نوع عملیات: هنگامی که یک عملیات ریاضی بر روی پردازنده اجرا می‌شود، به صورت خودکار نوع آن را شناسایی می‌شود و تصمیم گرفته می‌شود که آیا می‌توان آن را با دقت ممیز شناور ۱۶ بیتی انجام داد یا نیاز به اجرای آن در ممیز شناور ۳۲ بیتی وجود دارد[۱۷]. به طور کلی عملیات‌های حساس به دقت عددی مانند برخی محاسبات آماری در ممیز شناور ۳۲ بیتی اجرا می‌شوند.

پس از شناسایی نوع عملیات، اگر عملیات نیاز به دقت نداشت، ورودی‌ها به ممیز شناور ۱۶ بیتی تبدیل می‌شوند. در مقابل، اگر عملیات حساس باشد، حتی اگر داده‌ها در ممیز شناور ۱۶ بیتی باشند، قبل از اجرا به ممیز شناور ۳۲ بیتی بازگردانده می‌شوند. این فرآیند که تبدیل خودکار نوع داده نامیده می‌شود، تضمین می‌کند که سرعت محاسبات در عملیات ساده افزایش یابد و از افت دقت در عملیات حساس جلوگیری شود. از آنجا که داده‌های ممیز شناور ۱۶ بیتی تنها ۱٦ بیت فضا اشغال می‌کنند، استفاده از آن‌ها به جای ممیز شناور ۳۲ بیتی باعث کاهش مصرف حافظه می‌شود. این موضوع امکان استفاده از دسته[۲] بزرگ‌تر را بدون نیاز به سخت‌افزار قوی فراهم می‌کند. به طور کلی:

- لایه‌های کانولوشن[۳] و سایر لایه‌های سازگار در ممیز شناور ۱۶ بیتی اجرا می‌شوند.
- بیشینه‌سازی نرم[٤]، ساده‌سازی دسته[٥] و دیگر عملیات حساس در ممیز شناور ۳۲ بیتی انجام می‌گیرند.

---
۱ Mixed Precision
۲ Batch
۳ Convolution

لایه‌های کانولوشن در بخش‌های ستون فقرات شبکه[۱]، بخش میانی[۲] و بخش خروجی[۳]، به همراه توابع فعال‌سازی و نیز محاسبات عنصری مانند جمع و ضرب ساده، همگی در ممیز شناور ۱۶ بیتی اجرا شدند. این عملیات‌ها شامل ضرب و جمع‌های تکراری هستند که از نظر عددی نسبت به خطاهای ناشی از کوانتیزاسیون در ممیز شناور ۱۶ بیتی مقاوم هستند. خروجی این لایه‌ها پس از عبور از توابع فعال‌سازی و لایه‌های بعدی هموار[۴] و پایدار می‌شود، بنابراین خطاهای دقت پایین تأثیر محسوسی بر پایداری مدل YOLO ایجاد نمی‌کنند. در آزمایش‌های این پژوهش نیز مشخص شد که استفاده از ممیز شناور ۱۶ بیتی در این بخش‌ها بدون افت معنی‌دار دقت و با کاهش قابل توجه هزینه محاسباتی و مصرف حافظه همراه است. در مقابل، لایه‌های حساس نظیر ساده‌سازی دسته‌ای متکی بر محاسبه میانگین و واریانس هستند و خروجی‌های مبتنی بر تابع سیگموند و بیشنه‌سازی نرم وابسته به اختلافات بسیار کوچک بین مقادیر ورودی‌اند و در دقت پایین‌تر رفتار ناپایدار[۵] نشان می‌دهند و توسط بهینه‌ساز در ممیز شناور ۳۲ بیتی نگه داشته شدند. در فرآیند اجرای مدل، مکانیزم تشخیص خودکار عملیات این لایه‌های حساس را به‌صورت خودکار از محاسبات ممیز شناور ۱۶ بیتی جدا می‌کند تا تنها قسمت‌هایی با حساسیت کمتر (مثل لایه‌های کانولوشنی) در دقت پایین اجرا شوند و بخش‌های حساس در در ممیز شناور ۳۲ بیتی باقی بمانند. این ترکیب بهینه به مدل اجازه می‌دهد ضمن حفظ دقت، سرعت و کارایی سخت‌افزاری افزایش یابد. برای جلوگیری از ناپایداری ناشی از استفاده از ممیز شناور ۱۶ بیتی، معیار توقف بر اساس عدم بهبود مقدار میانگین دقت برای ۱۰ دوره تنظیم شد. در طول آزمایش‌ها، آستانه پایش‌شونده[۶] مقدار کاهش دقت[۷] بود و در صورت افزایش پیوسته آن طی ۵ دوره، آموزش در حالت ممیزشناور ۱۶ بیتی متوقف و محاسبات حساس در ممیزشناور ۳۲ بیتی ادامه می‌یافت.

هرس وزنی[۸]: مراحل کلی پیاده‌سازی به شرح زیر است:

آموزش اولیه مدل: مدل YOLOv۸ به طور کامل بر روی مجموعه داده دلخواه آموزش داده می‌شود.

اعمال هرس: در این پژوهش از هرس غیرساختاری[۹] برای کاهش تعداد پارامترهای شبکه استفاده شده است. در این ساختار وزن‌های جداگانه‌ای که مقدار کمی دارند حذف می‌شوند و باعث ایجاد الگوهای پراکنده در وزن‌ها می‌شود. انتخاب آستانه هرس بسیار مهم است؛ آستانه بالا می‌تواند منجر به افت محسوس در دقت شود[۱۸].

تنظیم دقیق: پس از هرس، مدل نیاز به تنظیم دقیق دارد تا وزن‌های باقی‌مانده خود را با شرایط جدید تطبیق دهد. این مرحله به حفظ دقت مدل کمک می‌کند.

در YOLOv۸ ، لایه‌های کانولوشن بیشترین وزن را دارند و بخش عمده حافظه و توان پردازشی را مصرف می‌کنند و هرس این لایه‌ها باعث کاهش اندازه مدل و افزایش سرعت می‌شود. در این مقاله، هرس وزنی بر روی لایه‌های کانولوشن مدل YOLOv۸ انجام شد، زیرا این لایه‌ها بیشترین سهم را در تعداد متغیرها و مصرف حافظه دارند. برخلاف روش‌های عمومی که صرفاً به حذف تصادفی یا یکنواخت وزن‌های کوچک می‌پردازند، ما با انتخاب آستانه بهینه بر اساس آزمایش‌های تدریجی و ارزیابی دقت، توانستیم تعادلی میان کاهش اندازه مدل و حفظ دقت ایجاد کنیم. پس از هرس، مرحله تنظیم دقیق نیز به‌طور خاص برای بازتنظیم وزن‌های باقی‌مانده در همین لایه‌ها طراحی شد. برای انتخاب آستانه، مقادیر ۱۰ درصد، ۲۰ درصد و ۳۰ درصد هرس آزمایش شدند. نتایج نشان داد که آستانه ۲۰ درصد بهترین توازن میان کاهش اندازه مدل و حفظ دقت را دارد. پس از اعمال هرس، مدل به مدت ۵۰ چرخه بازآموزی شد تا وزن‌های باقی‌مانده با معماری جدید سازگار شوند. معیار توقف در این مرحله نیز عدم بهبود میانگین دقت برای ۵ دوره بود.

تقطیر دانش: در این مقاله، برای کاهش حجم و پیچیدگی مدل YOLOv۸، از یک مدل YOLOv۸ بزرگ‌تر به عنوان معلم و نسخه کوچک‌تر به عنوان دانش‌آموز استفاده شده است. روند کار به صورت است:

آموزش مدل معلم: ابتدا نسخه بزرگ‌تر YOLOv۸-x بر روی مجموعه داده مورد نظر آموزش داده می‌شود.

تهیه خروجی نرم: در مرحله پیش‌بینی، خروجی مدل معلم با یک دمای مشخص T با فرمول زیر نرم می‌شود:

$$T^P = \left(\frac{T^z}{T}\right) \quad (۱)$$

T^z خروجی logits مدل معلم است.

آموزش مدل دانش‌آموز: نسخه کوچک‌تر YOLOv۸ با ترکیبی از دو تابع هزینه آموزش داده می‌شود:

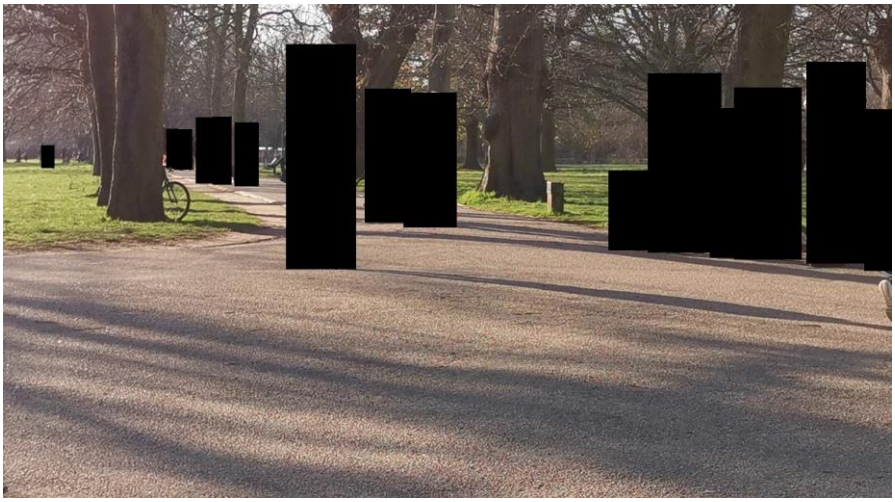
تابع هزینه عادی تشخیص اشیاء و تابع هزینه تقطیر که بر اساس خروجی نرم مدل معلم است.

تنظیم وزن ترکیب و دما: متغیرهای α و T با آزمایش‌های متعدد به گونه‌ای تنظیم می‌شوند که بهترین تعادل بین دقت و کارایی حاصل شود.

---

[۴] Softmax
[۵] Batch Normalization
[۱] Backbone
[۲] Neck
[۳] Head
[۴] Smooth
[۵] Instability
[۶] Monitor
[۷] mAP Loss
[۸] Weight Pruning
[۹] Unstructured Pruning

(الف)



(ب)

شکل (۵) : (الف) تصویر نواحی مورد علاقه مربوط به قاب اول شکل و (ب) تصویر پس‌زمینه مربوط به قاب اول

در نسخه موجود از استاندارد کدگذاری ویدئویی برای ماشین‌ها، برای بخش تشخیص اشیاء هیچ سازوکاری به منظور کاهش پیچیدگی محاسباتی در نظر گرفته نشده است و همین موضوع باعث می‌شود استفاده مستقیم از آن در کاربردهای بلادرنگ با محدودیت منابع دشوار باشد. به همین دلیل، رویکرد پیشنهادی در این مقاله رویکردی با بهره‌گیری از مدل‌های یادگیری عمیق موجود و اعمال تکنیک‌های کاهش پیچیدگی، شبکه پیشنهادی برای تشخیص اشیاء در کدگذاری ویدئویی برای ماشین‌ها را سبک خواهد کرد.

۳-۳- استخراج فریم‌ها و جداسازی نواحی مورد علاقه و ایجاد دو جریان ویدئویی

هدف اصلی این بخش پردازش ویدئوهای ورودی به منظور استخراج نواحی مورد علاقه و جداسازی از پس‌زمینه است. این عملیات به منظور تسهیل پردازش‌های بعدی مانند فشرده‌سازی انتخابی و رسیدن به هدف کدگذاری ویدئویی برای ماشین‌ها یعنی ایجاد جریان ویدئویی مناسب برای تحلیل ماشین‌ها و در صورت نیاز انسان انجام می‌شود.

۱-۳-۳- شناسایی نواحی مورد علاقه با مدلYOLO

با توجه به نتایج بدست آمده در ساده‌سازی YOLOv۸ برای شناسایی اشیاء و نواحی مورد علاقه در ساختار پیشنهادی کدگذار در این مقاله، از مدل YOLOv۸ نسخه‌ی سفارشی‌شده با چند دقتی مخلوط استفاده شده و با استفاده از پردازش دسته‌ای[۱]، که در جهت کاهش زمان اجرا استفاده شده است، فریم‌ها به دسته‌هایی با اندازه مشخص تقسیم شده و پردازش روی هر دسته انجام می‌شود. این روش باعث کاهش سربار ناشی از فراخوانی مدل برای هر فریم جداگانه می‌شود.

۲-۳-۳- استخراج موقعیت جعبه‌های محدودکننده[۲]

خروجی مدل شامل مختصات جعبه‌های محدودکننده برای هر شیء در فریم است. برای هر جعبه مختصات (x۱, y۱, x۲, y۲) استخراج و به شکل استاندارد x مرکز ، y مرکز ، عرض و ارتفاع ذخیره شد. این قالب برای هماهنگی با استانداردهای پردازش

[۱] Batch Processing
[۲] Bounding Boxes

ویدئو و فشرده‌سازی انتخابی استفاده می‌شود. پس از شناسایی جعبه‌های محدودکننده، دو تصویر از هر فریم تولید می‌شود که در جهت بازسازی مجدد نامگذاری آن‌ها بر اساس شماره فریم آن‌ها است:

- تصویر نواحی مورد علاقه: تنها نواحی درون جعبه‌های محدودکننده حفظ می‌شوند و سایر نواحی حذف می‌شوند. این فریم‌ها در جهت تحلیل ماشین‌ها استفاده می‌شوند(شکل (۵)(الف)).
- تصویر پس‌زمینه: نواحی شناسایی‌شده با جعبه‌های محدودکننده صفر (خالی[۱]) می‌شوند. این تصویر فقط شامل پس‌زمینه است و اشیاء اصلی در آن حذف می‌شوند(شکل(۵)(ب)).

همچنین، مختصات جعبه‌های محدودکننده برای هر فریم در فایل‌های متنی جداگانه ذخیره شدند تا امکان بازسازی دقیق فریم‌ها وجود داشته باشد.

۳-۳-۳- بازسازی ویدئو از قاب‌ها

پس از جداسازی نواحی، قاب‌ها به ویدئوهای مستقل بازسازی شدند و دو جریان ویدئویی مختلف داریم:

- ویدئو نواحی مورد علاقه: شامل تنها نواحی مورد علاقه شناسایی شده.
- ویدئو پس‌زمینه: شامل پس‌زمینه و بدون نواحی مورد علاقه.

در بازسازی ویدئو، ویژگی مهم این مرحله عبارت است از مرتب‌سازی قاب‌ها بر اساس شماره تا ترتیب زمانی حفظ شود. بازسازی ویدئوها امکان تجزیه و تحلیل را فراهم می‌کند و پایه‌ای برای اجرای الگوریتم‌های فشرده‌سازی انتخابی است. اهمیت این مرحله در تقسیم‌بندی محتوا با جداسازی ناحیه مورد علاقه و پس‌زمینه، آماده‌سازی داده برای تحلیل ماشین و انسان است. در مرحله بعدی به سراغ اعمال فشرده‌سازی پیشرفته بر اساس استاندارد VVC می‌رویم. پس از آماده‌سازی داده‌ها، فشرده‌سازی با استفاده از VVC انجام شد. در این مرحله، استراتژی تفکیک کیفیت بین ناحیه مورد علاقه و پس‌زمینه اعمال می‌شود؛ به گونه‌ای که قاب‌های ناحیه مورد علاقه با متغیر کوانتیزاسیون[۲] پایین‌تر و کیفیت بالاتر فشرده شدند تا جزئیات مهم برای تحلیل‌های ماشینی و تشخیص اشیاء حفظ شود، در حالی که قاب‌های پس‌زمینه با متغیر کوانتیزاسیون بالاتر فشرده شدند تا حجم داده کاهش یابد. این روش به طور مؤثر تعادلی بین کیفیت تصویر و نرخ بیت ایجاد می‌کند و از انتقال اطلاعات غیرضروری جلوگیری می‌کند. این روش نشان‌دهنده یک رویکرد کارآمد برای فشرده‌سازی ویدئو با تمرکز بر ماشین است. بر اساس شکل (٤) که ساختار کلی کدگذار روش پیشنهادی را نشان می‌دهد، ابتدا توالی ویدئویی ورودی به قاب‌های مجزا استخراج می‌شود. سپس تمام قاب‌ها تحت عملیات کاهش ابعاد[۳] قرار می‌گیرند. پس از آن، مدل ساده شده YOLOv۸ بر روی قاب‌ها اعمال می‌شود تا نواحی مورد علاقه از پس‌زمینه تفکیک گردد. در مرحله بعد، دو جریان ویدئویی بازسازی می‌شوند: یکی شامل قاب‌های نواحی مورد علاقه و دیگری شامل قاب‌های پس‌زمینه .هرکدام از این توالی‌ها به صورت جداگانه کدگذاری می‌شوند. برای قاب‌های نواحی مورد علاقه از کدگذار VVC با مقدار متغیر کوانتیزاسیون برابر با ۳۷ استفاده می‌شود تا کیفیت مورد نیاز برای بینایی ماشین حفظ شود[۹] [۱۱]. در مقابل، قاب‌های پس‌زمینه با مقدار متغیر کوانتیزاسیون بالاتر (٤٧) کدگذاری می‌شوند تا نرخ بیت برای بخش‌هایی که اهمیت کمتری دارند کاهش یابد. مقدار متغیر کوانتیزاسیون برای ناحیه پس‌زمینه به صورت تجربی انتخاب شده است به نحوی که اطلاعات غیر مفید با کیفیت پایین و نرخ بیت کم ارسال شوند.

در نهایت، علاوه بر دو توالی ویدئویی کدگذاری‌شده، فراداده شامل مختصات و اطلاعات نواحی مورد علاقه نیز استخراج و ذخیره می‌شود تا در وظایف بینایی ماشین مورد استفاده قرار گیرد. این رویکرد به‌طور هم زمان به کاهش نرخ بیت و حفظ دقت وظایف بینایی ماشین کمک می‌کند.

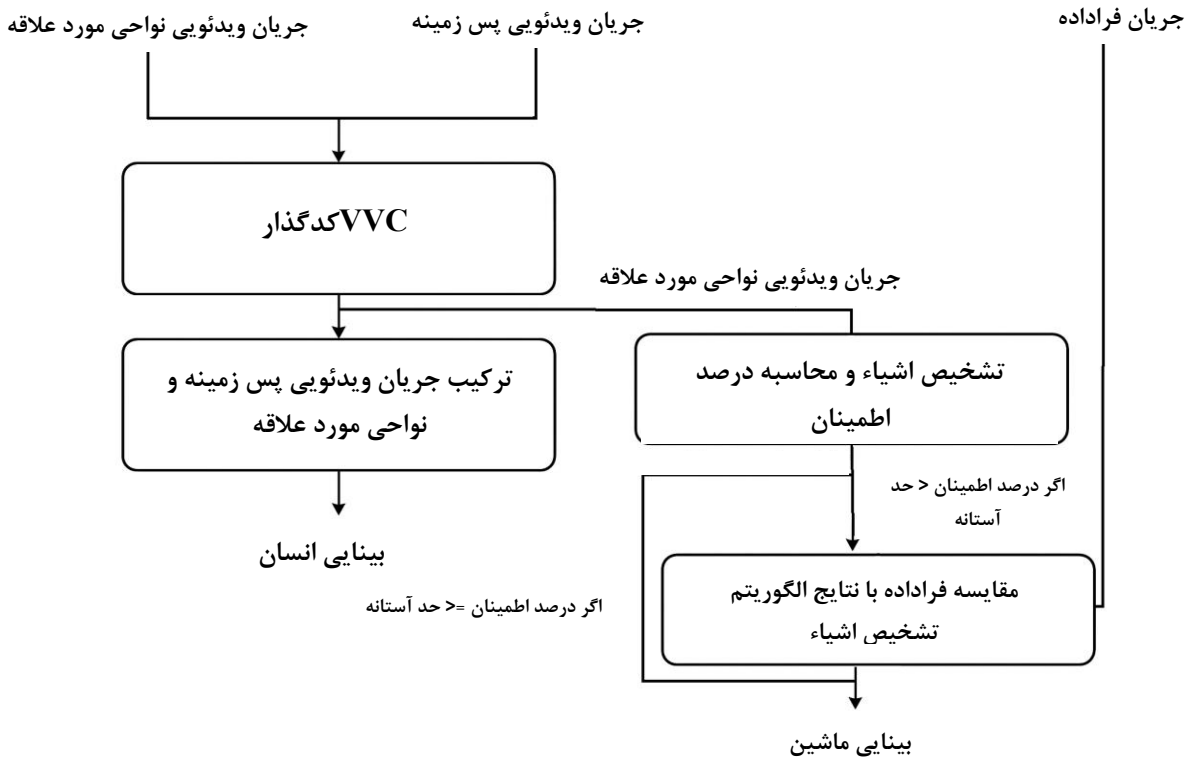
٤-۳- استفاده از فراداده

یکی از روش‌های افزایش دقت تشخیص، استفاده از فراداده مرتبط با قاب‌ها برای انتقال اطلاعات مربوط به موقعیت نواحی مورد علاقه در ویدئو است. در مرحله استخراج قاب‌ها و تشخیص اشیاء با مدل YOLO، برای هر قاب، مختصات جعبه‌های محدودکننده اشیاء شناسایی شده و ذخیره می‌شود. این مختصات شامل مرکز جعبه و ابعاد آن است و در قالب یک فایل متنی برای هر قاب نگهداری می‌شود. به این ترتیب، برای هر قاب یک فایل فراداده تولید می‌شود که اطلاعات موقعیت و اندازه اشیاء مهم را در اختیار دستگاه قرار می‌دهد. از آنجایی که داده‌های فراداده در جریان کدگذار محاسبه و ذخیره می‌شوند، استفاده از آن‌ها هیچگونه پیچیدگی عملیاتی اضافه بر کدگذار اعمال نمی‌کند و همچنین بخاطر حجم کم این داده‌ها ارسال آن‌ها در کنار جریان ویدئویی خروجی تاثیر به خصوصی در حجم داده ارسالی ندارد.

در مرحله کدگشایی، این فراداده می‌تواند موقعیت دقیق اشیاء مهم را در هر قاب مشخص کند. این قابلیت امکان تحلیل ویدئویی برای ماشین‌ها را، با ارائه مکان دقیق نواحی مورد علاقه در صورت عدم تشخیص صحیح شیء فراهم می‌کند. همچنین، این فراداده می‌تواند برای هماهنگی بین دو جریان ویدئویی مورد استفاده قرار گیرد.

---

[۱] Blank
[۲] Quantization Parameter (QP)
[۳] Down-sampling



شکل (٦) : ساختار کدگشای پیشنهادی

به طوری که بازسازی بخش‌های مهم و غیرمهم به صورت دقیق انجام شود و درصورت نیاز بتوان با بررسی مکان نواحی مهم و بازسازی باکیفیت‌تر آن نقاط، دقت تحلیل را برای بینایی ماشین افزایش داد. مزایای استفاده از فراداده عبارتند از:

- حفظ جزئیات حیاتی
- افزاش دقت تحلیل ویدئوهای فشرده برای ماشین

۵-۳-کدگشایی و بازسازی

پس از انجام فشرده‌سازی ویدئو، فایل‌های تولیدشده با پسوند VVC. شامل جریان ویدئویی نواحی مورد علاقه و پس‌زمینه، توسط کدگشا VVC بازسازی شده و سپس در صورت نیاز برای تحلیل توسط کاربر انسانی ترکیب مجدد آن‌ها انجام می‌شود. برای بازسازی ویدئوی برای کاربر انسانی، ویدئوی نواحی مورد علاقه به‌عنوان یک ماسک عمل می‌کند تا فقط بخش‌های دارای اطلاعات شیء از آن استخراج و با ویدئوی پس‌زمینه ترکیب شوند. الگوریتم به صورت زیر عمل می‌کند. قاب‌های متناظر از دو ویدئو خوانده می‌شوند، یک ماسک باینری از قاب ویدئوی نواحی مورد علاقه تولید می‌شود؛ به گونه‌ای که پیکسل‌های غیرمشکی (مربوط به اشیاء) مقدار ۱ گرفته و بقیه صفر می‌شوند، با استفاده از این ماسک، اشیاء در قاب نواحی مورد علاقه متناظر جدا شده و سپس بر روی پس‌زمینه متناظر قرار می‌گیرند.این فرآیند برای تمام قاب‌ها تکرار شده و در نهایت ویدئوی بازسازی‌شده توسط کاربر انسانی قابل تحلیل است. این فرآیند باعث می‌شود که در عین کاهش نرخ بیت در پس‌زمینه، جزئیات کامل اشیاء مهم حفظ شوند. برای تحلیل ماشین، جریان ویدئویی نواحی مورد علاقه مستقیما به مدل‌های تشخیص اشیاء داده می‌شود. برای هر شیء تشخیص داده شده یک آستانه اطمینان[1] وجود دارد، که اگر شیء تشخیص داده شده که از این آستانه اطمینان مقدار کمتری داشته باشد در حالت معمول به عوان ناحیه مورد علاقه در نظر گرفته نمی‌شود. برای افزایش دقت تشخیص جریان فراداده ارسال شده به کدگشا استفاده می‌شود. بعد از اینکه در قابی یک شیء مقدار اطمینان کمتر از آستانه داشته باشد داده‌های مکانی آن در فایل فراداده متناظر مورد جستجو قرار می‌گردد؛ اگر این داده در فایل فراداده هم وجود داشته باشد نشان دهنده یک شیء مهم است پس برخلاف این که مقدار آن از آستانه اطمینان کمتر است همچنان به عنوان شیء مهم به ماشین معرفی می‌شود، در غیر این صورت تشخیص مدل درست بوده است و شیء با اطمینان کمتر از آستانه در تحلیل در نظر گرفته نمی‌شود. شکل (٦) روند پردازش در کدگشا را نشان می‌دهد. در این بخش، توالی ویدئویی ناحیه‌ی مورد علاقه و توالی پس‌زمینه پس از دریافت، توسط کدگذار معکوس بازسازی می‌شوند. برای بینایی انسانی، دو توالی نواحی مورد علاقه و پس‌زمینه با یکدیگر ترکیب شده و ویدئوی نهایی بازسازی می‌شود تا کیفیت نمایش برای کاربر حفظ گردد. برای بینایی ماشین، ابتدا مجددا فرآیند آشکارسازی اشیاء روی توالی نواحی مورد علاقه انجام شده و

[1] Confidence Threshold

میزان اطمینان[۱] برای هر شیء محاسبه می‌شود. اگر میزان اطمینان بالاتر یا مساوی آستانه تعریف‌شده باشد، نتایج آشکارسازی پذیرفته می‌شوند. در غیر این صورت، موقعیت اشیاء شناسایی‌شده با فراداده ارسال‌شده در مرحله کدگذاری مقایسه می‌شود تا دقت وظایف بینایی ماشین تضمین گردد .این رویکرد باعث می‌شود که حتی در شرایطی که کیفیت بازسازی پایین‌تر است یا آشکارساز دچار خطا می‌شود، وظایف بینایی ماشین همچنان با کمک فراداده از دقت مناسبی برخوردار باشند.

٤- ارزیابی

٤-١- ارزیابی روش‌های ساده‌سازی مدل

روش‌های مطرح شده همگی برای ساده‌سازی مدل بدون افت قابل توجه دقت تشخیص مطرح شده‌اند، با توجه به نتایج مربوط به سرعت، دقت و حافظه مصرفی درنهایت مدل چند دقتی برای ادامه کار در کدگذار انتخاب شد. نتایج مقایسه روش‌های ساده‌سازی در جدول (۲) نشان داده شده است. متغیرهای مقایسه عبارتند از:

- زمان آموزش: مدت زمانی است که مدل برای طی‌کردن فرآیند آموزش روی مجموعه داده نیاز دارد. به طور تقریبی زمان آموزش مدل اصلی ۲۹ دقیقه و تعداد دفعات آموزش برابر ۱۰۰ است و زمان‌های نوشته شده برای مدل‌ها مجموع زمان ساده‌سازی آن‌ها به اضافه آموزش اولیه است. برای مثال، در چند دقتی ۲۹ دقیقه زمان آموزش و ١٤ ثانیه زمان اجرای ساده‌سازی آن است.
- اندازه یا اندازه مدل: به حجم فایل مدل نهایی (بر حسب مگابایت) اشاره دارد.
- دقت میانگین: مقدار آن بین ۰ و ۱ (یا درصدی بین ۰ تا ۱۰۰) است؛ هرچه بالاتر باشد، عملکرد مدل بهتر است. دقت برابر است با مجموع دقت تشخیص اشیاء موجود در همه کلاس‌ها تقسیم بر تعداد کل کلاس‌ها.
- تاخیر: مدت زمانی است که طول می‌کشد تا یک تصویر ورودی به خروجی تبدیل شود.
- حافظه مصرفی: مقدار حافظه‌ی RAM یا VRAM که مدل هنگام بارگذاری و تحلیل نیاز دارد.
- عملیات ممیز شناور در هر ثانیه[۲]: نشان‌دهنده پیچیدگی محاسباتی مدل است. هرچه FLOPs کمتر باشد، سرعت اجرا بالاتر و مصرف انرژی کمتر خواهد بود.
- تعداد قاب در ثانیه[۳]: تعداد قاب‌هایی که مدل می‌تواند در یک ثانیه پردازش کند.

[۱] Confidence
[۲] Floating Point Operations Per Second (FLOPs)

نتایج برروی ویدئوهای ورودی با طول متفاوت اندازه گرفته شده است و به صورت میانگین برای هر قاب اندازه‌گیری شده است و در جدول ذکر شده است. برای ارزیابی عملکرد روش‌های مختلف ساده‌سازی، از معیارهای زمان آموزش، اندازه مدل، دقت، تاخیر، حافظه مصرفی، تعداد عملیات محاسباتی و نرخ پردازش قاب استفاده شده است. با بررسی زمان آموزش، چند دقتی و تقطیر دانش به ترتیب ١٤ ثانیه و ۱ دقیقه به زمان کل اضافه کردند. اما هرس وزنی چون نیازمند یک مرحله تنظیم دقیق کامل بود، زمان آموزش آن عملاً دو برابر شد. ترکیب سه روش نیز بیشترین زمان آموزش را به خود اختصاص داد. حافظه مصرفی و اندازه فایل مدل در تمامی روش‌ها کاهش محدودی داشته است. به عنوان نمونه، اندازه مدل تغییراتی جزئی (کمتر از ۱ درصد) داشته است.

نتایج نشان دادند که استفاده از مدل‌های ساده‌سازی شده تنها باعث کاهش بسیار ناچیز (کمتر از ۰/۰۵ درصد) در دقت شدند. در حالی که استفاده از تقطیر دانش افت بیشتری نسبت به دیگر مدل‌ها دارد. به طور کلی می‌توان گفت تمامی روش‌های موجود توانسته‌اند دقت مدل اصلی را تا حدودی حفظ کنند. یکی از مهم‌ترین معیارها برای کاربردهای بلادرنگ، زمان تاخیر مدل است. استفاده از چند دقتی موجب کاهش چشمگیر تاخیر تا ٦/٣٣ میلی‌ثانیه، حدود ٥٧ درصد شد. هرس وزنی نیز زمان را به ٨/٩٥ میلی‌ثانیه و تقطیر دانش به ١١/٥٧ میلی‌ثانیه کاهش دادند. تعداد عملیات محاسباتی در تمامی مدل‌ها تقریباً ثابت باقی ماند. به طور کلی بر اساس نتایج، چند دقتی به تنهایی بهترین تعادل میان سرعت و دقت را فراهم کرده و زمان تاخیر را بیش از ۵۰ درصد کاهش داده است، بدون اینکه دقت مدل کاهش قابل توجهی داشته باشد. همان‌گونه که مشاهده می‌شود، تعداد عملیات محاسباتی در تمام نسخه‌ها تقریباً ثابت باقی مانده است(حدود ٤/٠٩—٤/٠٨)، زیرا معماری اصلی شبکه تغییر نکرده و تنها نوع نمایش عددی یا وزن‌های غیرفعال حذف شده‌اند.

از نظر تعداد پارامترها و اندازه مدل، بیشترین کاهش در روش هرس وزنی مشاهده می‌شود (حدود ۳ درصد کاهش). در مقابل، روش چنددقتی مصرف حافظه را به‌طور قابل توجهی کاهش داده و زمان تاخیر را تا ٥٧ درصد کاهش داده است. روش تقطیر دانش نیز کوچک‌ترین افت دقت را ایجاد کرده است. نتیجه کلی نشان می‌دهد که چنددقتی بهترین تعادل میان سرعت و دقت را ارائه می‌دهد و برای کاربردهای بلادرنگ توصیه می‌شود، در حالی‌که هرس وزنی و تقطیر دانش بیشتر برای کاهش اندازه مدل و بهبود قابلیت استقرار مناسب هستند. هرس وزنی و تقطیر دانش هر کدام به تنهایی مزایای محدودی ایجاد کرده‌اند، در نهایت، ترکیب سه روش اگرچه کاهش خوبی در تاخیر ایجاد کرده است، اما به دلیل افزایش زمان آموزش، نسبت به چند دقتی مزیت چندانی نشان

[۳] Frames Per Second (FPS)

نمی‌دهد. این نتایج نشان می‌دهند که برای کاربردهای بلادرنگ، استفاده از چند دقتی می‌تواند بهترین انتخاب باشد.

در شکل (۷) نتایج مقایسه میان نسخه‌های مختلف شبکه YOLOv۸ تحت روش‌های گوناگون ساده‌سازی مدل شامل مدل پایه[۱] ، چند دقتی، هرس وزنی، تقطیر دانش و روش ترکیبی (Mixed + Prune + KD) نمایش داده شده است. هدف اصلی از این مقایسه، بررسی اثر هر یک از روش‌های فشرده‌سازی و بهینه‌سازی مدل بر معیارهای کلیدی نظیر زمان تاخیر (ستون زرد)، تعداد قاب در ثانیه (ستون آبی)، دقت (ستون سبز)، اندازه مدل (نمودار بنفش) و مصرف حافظه گرافیکی (نمودار قرمز) است. بر اساس نتایج مشاهده‌شده، روش چند دقتی موجب کاهش چشم‌گیر زمان تاخیر و مصرف حافظه گرافیکی شده و همچنان دقت مدل را در سطح بالایی حفظ کرده است. روش هرس وزنی با حذف بخشی از وزن‌های غیرضروری توانسته اندازه مدل را کاهش دهد و سرعت اجرا را افزایش دهد، اما به‌طور نسبی مصرف حافظه بیشتری نسبت به چند دقتی دارد. در مقابل، روش تقطیر دانش باعث پایداری بهتر در دقت مدل و کاهش بار محاسباتی گرافیکی شده، هرچند زمان تاخیر آن نسبت به چند دقتی بیشتر است. در کاربردهای بلادرنگ که کاهش مصرف منابع و افزایش سرعت اهمیت بیشتری دارد، استفاده از چند دقتی مناسب‌تر است، در حالی‌که در کاربردهایی که حفظ دقت در کنار کاهش اندازه مدل اهمیت دارد، روش‌های مبتنی بر هرس وزنی و تقطیر دانش یا ترکیب آن‌ها گزینه بهتری خواهند بود.

۲-٤- ارزیابی الگوریتم پیشنهادی کدگذاری ویدئویی برای ماشین‌ها

نتایج مربوط به الگوریتم پیشنهادی و الگوریتم‌های مرجع در جدول (۳) نشان داده شده است. متغیرهای مقایسه عبارتنداند از:

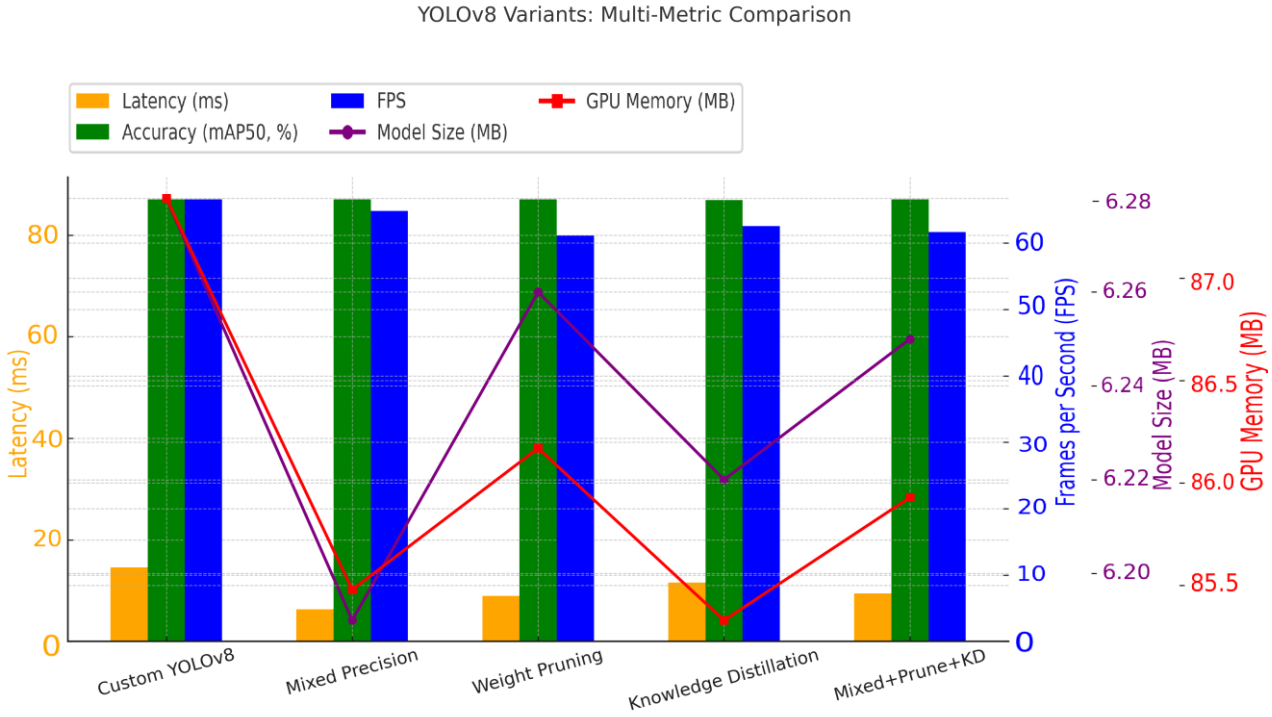
- بیت بر پیکسل[۲]: نشان‌دهنده‌ی میزان فشرده‌سازی است و بیان می‌کند برای نمایش هر پیکسل چه تعداد بیت مصرف شده است.
- نرخ بیت[۳]: تعداد بیت‌هایی که در هر ثانیه برای نمایش ویدئو تولید می‌شود. به صورت مستقیم روی حجم فایل و پهنای باند مورد نیاز تأثیر دارد. در کدگذاری، نرخ بیت پایین به معنای صرفه‌جویی در ذخیره‌سازی و انتقال است، اما ممکن است باعث کاهش کیفیت شود. در ردیف اول نرخ بیت گزارش شده برای یک قاب است درحالی که برای بقیه ستون‌ها میانگین نرخ بیت برای جریان ویدئویی اندازه‌گیری شده است.
- میانگین دقت: نشان می‌دهد الگوریتم فشرده‌سازی تا چه حد توانسته ویژگی‌های ضروری برای تحلیل ماشینی را حفظ کند. معمولاً با mAP یا معیارهای مشابه گزارش می‌شود که مقدار عددی آن بین ۰ و ۱ است یا به صورت درصدی گزارش می‌شود. در ردیف اول، الگوریتم برای وظیفه ردیابی شیء بهینه شده است و دقت آن براساس MOTA گزارش شده است. بقیه ردیف‌ها با mAP گزارش شده است. دقت برروی ویدئوی کدگشایی شده در مراجع و روش پیشنهادی انجام می‌شود. دقت برابر است با مجموع دقت تشخیص اشیاء موجود در همه کلاس‌ها تقسیم بر تعداد کل کلاس‌ها در ویدئو کدگشایی شده.
- تاخیر[٤]: زمان مورد نیاز برای پردازش یک قاب یا یک ویدئو.
- میزان فشرده‌سازی نسبت به مرجع[٥]: بیانگر میزان کاهش حجم داده در مقایسه با یک کدگذار/کدگشا یا روش مرجع است. در جدول ۲مرجع استاندارد VVC است که نسبت فشرده‌سازی براساس ویدئوی ورودی با کیفیت ۱۹۲۰×۱۰۸۰ است.

نتایج برروی ویدئوهای ورودی با طول متفاوت اندازه گرفته شده است و به صورت میانگین برای هر قاب اندازه‌گیری شده است و پیشنهادی توانسته است بهینه‌سازی فشرده‌سازی را بدون افت دقت بالایی حفظ کند. در صورت استفاده از فراداده، نرخ بیت کمی افزایش یافته، در حالی که دقت از ۹۱/٤۷ درصد به ۹۵/۸۲ درصد افزایش می‌یابد. این نشان‌دهنده نقش کلیدی فراداده در بهبود عملکرد شناسایی اشیاء بدون افزایش قابل توجه حجم داده‌هاست. در مقیاس دقت، الگوریتم دوم دقت حدود ۹۱/۱ درصد و الگوریتم سوم دقت ۸۲/٤ درصد ارائه می‌دهند، در حالی که دقت الگوریتم پیشنهادی حتی بدون استفاده از فراداده به حدود ۹۱/٤۷ درصد می‌رسد و با استفاده از فراداده به ۹۵/۸۲ درصد می‌رسد. که نشان‌دهنده توانایی الگوریتم پیشنهادی در حفظ یا افزایش دقت شناسایی اشیاء همزمان با کاهش نرخ بیت، است. از لحاظ زمان پردازش یا تاخیر، الگوریتم پیشنهادی با زمان حدود ۱/۵٦ ثانیه برای هر قاب کمی کندتر از باقی الگوریتم‌هاست. این افزایش زمان پردازش، نشان‌دهنده مبادله بین افزایش دقت و زمان محاسباتی است.

[۱] Custom Model
[۲] Bits Per Pixel (BPP)
[۳] Bitrate
[٤] Delay
[٥] Anchor

جدول (۲) : مقایسه نتایج روش‌های موجود برای ساده‌سازی مدل سفارشی YOLOv۸ برروی مجموعه‌داده TVD و Open Images

نام مدل	زمان آموزش	اندازه مدل	میانگین دقت	تاخیر	حافظه مصرفی	FLOPs	FPS
استاندارد YOLOv۸	۰۰:۲۹:۰۰	٦/۲۸ MB	۰/۸۷۰۸	۱٤/٥۸ ms	۸۷/۳۹ MB	٤/۰۹ Gmac	٦٦/٦
چنددقتی	۰۰:۲۹:۱٤	٦/۱۹ MB	۰/۸۷۰۷	٦/۳۳ ms	۸٥/٤۸ MB	٤/۰۹ Gmac	٦٤/۸٥
هرس وزنی	۰۰:٥۸:۰۰	٦/۲٦ MB	۰/۸۷۰٦	۸/۹٥ ms	۸٦/۱۷ MB	٤/۰۸ Gmac	٦۱/۱٤
تقطیر دانش	۰۰:۳۰:۰۰	٦/۲۲ MB	۰/۸٦۹۱	۱۱/٥۷ ms	۸٥/۳۳ MB	٤/۰۸ Gmac	٦۲/٥٥
ترکیب سه روش	۰۱:۱۹:۰۰	٦/۲٥ MB	۰/۸۷۰٤	۹/٤٦ ms	۸٥/۹۳ MB	٤/۰۸ Gmac	٦۱/٦۷



شکل (۷) : مقایسه مدل‌های ساده‌سازی YOLOv۸

با افزودن فراداده، زمان پردازش به ۱/٦۹ ثانیه افزایش می‌یابد، که هنوز در محدوده قابل قبول برای کاربردهای با تحلیل دقیق قاب به قاب است. نسبت فشرده‌سازی نیز نشان می‌دهد که الگوریتم پیشنهادی با مقدار ۲/۹۲ بهترین عملکرد را در میان روش‌های بررسی‌شده دارد، که به معنای توانایی الگوریتم در کاهش حجم داده‌ها تقریباً تا سه برابر در مقایسه با استاندارد VVC است. بر اساس این اعداد درصد بهبود در الگوریتم پیشنهادی (بدون استفاده از جریان فراداده) نسبت به روش‌های قبلی [۹]، [۱۰]، و [۱۳] به ترتیب ۱۸/۷٪، ٦۰٪ و ۱٦/۸٪ است. درصد بهبود در الگوریتم پیشنهادی (با استفاده از جریان فراداده) نسبت به روش‌های قبلی [۹]، [۱۰]، و [۱۳] نیز به ترتیب ۱۸/۳٪ ، ٥۹٪ و ۱٦/٤٪ است. بنابراین، الگوریتم پیشنهادی ترکیبی از نرخ بیت پایین، دقت بالا و نسبت فشرده‌سازی مناسب ارائه می‌دهد. استفاده از فراداده باعث افزایش دقت بدون افزایش چشم‌گیر حجم داده‌ها می‌شود و نشان می‌دهد که انتقال اطلاعات اضافی به صورت هوشمند می‌تواند بهبود عملکرد شناسایی اشیاء را بدون افت کارایی فشرده‌سازی فراهم کند. این نتایج نشان می‌دهند که الگوریتم پیشنهادی برای کاربردهایی که همزمان به دقت بالا و کاهش حجم داده نیاز دارند، گزینه‌ای مناسب و بهینه محسوب می‌شود.

در شکل (۸) نتایج مقایسه میان سه روش موجود شامل روش مبتنی بر ناحیه مورد علاقه[۹]،ساده‌سازی قاب‌ها بر اساس نواحی مورد علاقه [۱۳]، رنگ آمیزی مجدد قاب‌ها[۱۰] با روش پیشنهادی این پژوهش و نسخه بهبودیافته آن است. معیارهای ارزیابی شامل:

- نرخ بیت: که بیانگر میزان فشرده‌سازی داده‌هاست (ستون آبی).
- دقت: به درصد بیان شده و نشان‌دهنده کارایی مدل در وظیفه تشخیص شیء است (ستون سبز).

- زمان تاخیر: به‌صورت ثانیه بر قاب، که نشان‌دهنده سرعت پردازش است (ستون نارنجی).
- نسبت فشرده‌سازی: که براساس مقایسه با یک کدگذار/کدگشا استاندار محاسبه می‌شود (نمودار قرمز)

بر اساس نتایج، روش [۹] نرخ بیت بسیار پایینی دارد اما دقت آن کمتر از سایر روش‌هاست. روش [۱۳] دقت بالاتری ارائه می‌دهد ولی از نظر نرخ بیت و تاخیر ضعیف‌تر عمل می‌کند. روش [۱۰] نسبت به [۱۳] بهبود در نرخ بیت و تعادل میان دقت و تاخیر نشان می‌دهد. در مقابل، روش پیشنهادی توانسته است هم‌زمان دقت بالا و تاخیر کمتر را به دست آورد و همچنین نسبت فشرده‌سازی بهتری نسبت به روش‌های موجود ارائه کند. افزون بر این، نسخه Proposed+Meta با استفاده از فراداده باعث افزایش بیشتر دقت و کاهش جزئی در نرخ بیت شده و به‌عنوان بهترین موازنه بین دقت، سرعت و فشرده‌سازی مطرح می‌شود.

شکل (۹) قسمت (الف) و (ب) به ترتیب نشان‌دهنده یک فریم از ویدئوی ورودی و سپس همان فریم پس از بازسازی است.

هدف این مقایسه آن است که نشان دهد روش پیشنهادی نه‌تنها از نظر پیچیدگی محاسباتی برای کاربردهای کدگذاری ویدئویی برای ماشین‌ها مناسب است، بلکه از نظر حفظ دقت تشخیص اشیاء و کاهش منابع موردنیاز نیز برتری نسبت به روش‌های موجود دارد.

۵- بررسی و تحلیل نتایج

نتایج این پژوهش نشان می‌دهد که ساده‌سازی مدل YOLOv۸ و طراحی معماری سه‌جریانی برای کدگذاری ویدئویی، در کاهش پیچیدگی و افزایش فشرده‌سازی تاثیر داشته است، در حالی که دقت تشخیص اشیاء تقریبا ثابت مانده است. عملکرد بهتر نسخه چنددقتی مدل را می‌توان به کاهش عملیات نقطه‌شناور در لایه‌های کانولوشنی نسبت داد؛ لایه‌هایی که بخش عمده بار محاسباتی شبکه را تشکیل می‌دهند و نسبت به کاهش دقت عددی حساسیت کمتری دارند. این ویژگی سبب شده تا ساده‌سازی چند دقتی بیشترین کاهش زمان تاخیر و بیشترین بهبود سرعت را ایجاد کند، بدون آنکه دقت خروجی مدل افت محسوسی داشته باشد.

در بخش کدگذاری ویدئویی، ساختار سه‌جریانی پیشنهادی شامل جریان نواحی مورد علاقه، جریان پس‌زمینه و جریان فراداده باعث شده است که بخش عمده اطلاعات غیرضروری حذف شده و اطلاعات مرتبط با تحلیل ماشین با کیفیت بالاتر حفظ شود. همچنین مشخص شد که این رویکرد نسبت به روش‌های مرجع مانند VVC یا مدل‌های مبتنی بر استخراج توصیف‌گر، عملکرد بهتری از نظر تعادل میان دقت، نرخ بیت و زمان پردازش ارائه می‌دهد. این نتایج نشان می‌دهند که ترکیب مدل ساده‌سازی‌شده YOLOv۸ با معماری کدگذاری پیشنهادی می‌تواند یک راهکار مؤثر برای سناریوهای بلادرنگ باشد.

با وجود مزایای به‌دست‌آمده، روش پیشنهادی محدودیت‌هایی نیز دارد. نمونه‌برداری کاهشی و کاهش دقت محاسباتی ممکن است در شناسایی اشیای بسیار کوچک منجر به افت دقت شود. همچنین اجرای چند دقتی و ساختار سه‌جریانی نیاز به سخت‌افزار سازگار دارد. از سوی دیگر، پیاده‌سازی سه جریان مجزا پیچیدگی بیشتری در ساختار کدگذار و کدگشا ایجاد می‌کند که ممکن است در سیستم‌های توکار محدودیت‌هایی به همراه داشته باشد. با این حال، نتایج نشان می‌دهد که در کاربردهایی مانند نظارت هوشمند، پردازش لبه‌ای و خودروهای خودران، روش ارائه‌شده می‌تواند با حفظ دقت و کاهش مصرف منابع محاسباتی عملکرد قابل توجهی ارائه دهد. در مجموع، این پژوهش نشان می‌دهد که ترکیب ساده‌سازی هدفمند مدل تشخیص اشیاء با طراحی هوشمندانه معماری کدگذاری، می‌تواند گامی مؤثر در جهت توسعه استانداردهای جدید کدگذاری ویدئویی برای ماشین‌ها باشد.

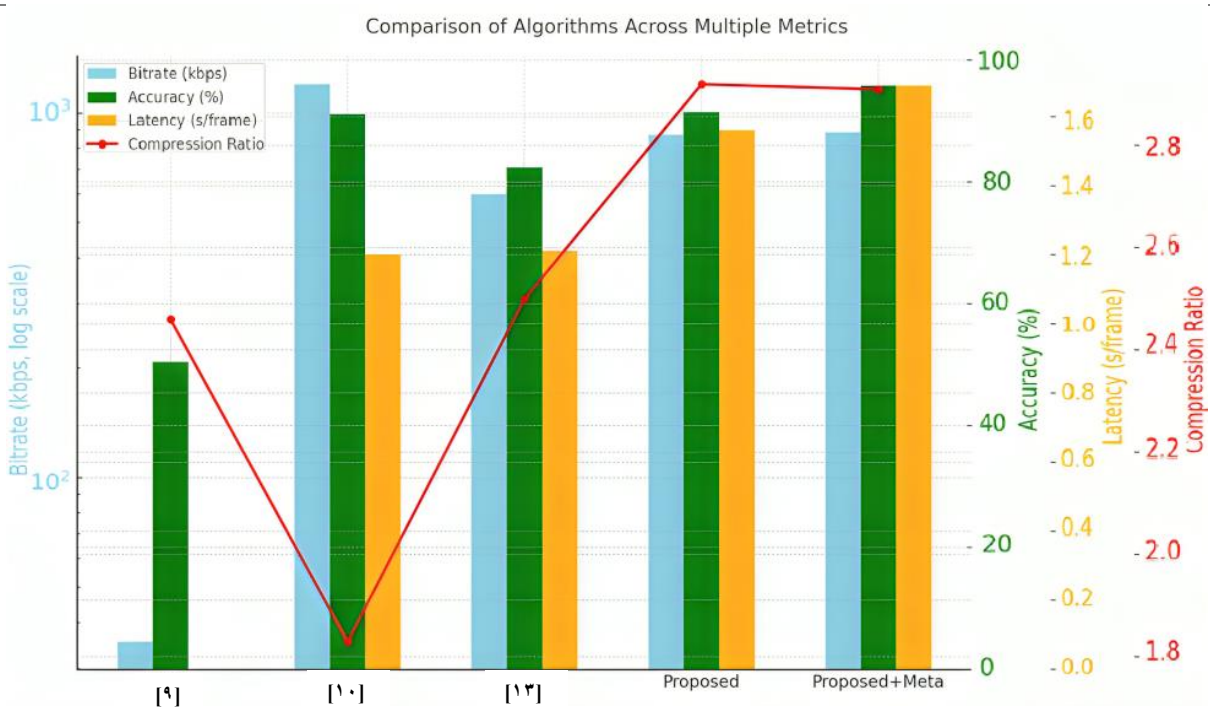
۶- نتیجه گیری و کارهای آتی

در این پژوهش با هدف کاهش پیچیدگی و افزایش کارایی در فرآیند تشخیص اشیاء، یک نسخه ساده‌سازی‌شده از YOLOv۸ ارائه شد. رویکرد پیشنهادی با تمرکز بر نواحی مورد علاقه و جداسازی آن‌ها از پس‌زمینه، علاوه بر کاهش حجم پردازش، امکان استفاده بهینه‌تر از منابع محاسباتی را فراهم ساخت. همچنین با تعریف معماری جدید کدگذار—کدگشا و تولید سه جریان مجزا شامل نواحی مورد علاقه، پس‌زمینه و فراداده، توانستیم هم‌زمان فشرده‌سازی مؤثر و بازسازی دقیق اطلاعات را تضمین کنیم. نتایج به‌دست‌آمده نشان می‌دهد که روش ارائه‌شده می‌تواند بدون افت محسوس در دقت، پیچیدگی محاسباتی را کاهش داده و زمینه را برای توسعه سامانه‌های تشخیص اشیاء سبک‌تر و کارآمدتر فراهم سازد.

اگرچه استفاده از مدل‌های جدیدتر مانند YOLOv۱۰ ، RT-DET یا SAM می‌تواند دقت تشخیص را افزایش دهد، اما لازم است چالش‌های عملی این انتقال نیز مورد توجه قرار گیرد. مدل‌های نسل جدید اغلب نیازمند منابع سخت‌افزاری قدرتمندتر، حافظه گرافیکی بیشتر و پهنای باند بیشتر هستند؛ بنابراین بهینه‌سازی آن‌ها برای کاربردهای بلادرنگ یک چالش اساسی محسوب می‌شود. علاوه بر این، استفاده از معماری‌های پیچیده‌تر ممکن است تاخیر را افزایش دهد و با محدودیت‌های پردازنده ناسازگار باشد. در تحقیقات آینده می‌توان مسیرهای عملی‌تری را دنبال کرد؛ از جمله طراحی نسخه‌های سبک‌شده از مدل‌های جدید، استفاده از تکنیک‌های موازی‌سازی، بهینه‌سازی اختصاصی برای سخت‌افزارهای کم‌مصرف، و توسعه روش‌های سازگار با معماری‌های واقعی برای ماشین‌ها که محدودیت نرخ بیت و تاخیر در آن‌ها اهمیت بالایی دارد. چنین رویکردهایی می‌توانند موجب شوند که مدل‌های پیشرفته‌تر نیز در سناریوهای بلادرنگ به‌صورت پایدار و قابل اتکا مورد استفاده قرار گیرند.

جدول (۳) : مقایسه نتایج الگوریتم پیشنهادی با روش‌های مرجع، با جریان ورودی با کیفیت ۱۹۲۰×۱۰۸۰ و مدل YOLO

الگوریتم	BPP	نرخ بیت	دقت	تاخیر	نسبت فشرده‌سازی
[۹]	~۰/۰٤	Kbps۳۵/٤۸	۵۰/۵	گزارش نشده	x ۲/٤٦
[۱۰]	~۰/۰٦	۱/۲ Mbit	۹۱/۱	~۱/ ۲ ms	x ۱/۸۳
[۱۳]	~۰/۰۵	٦۰۰ Kbps	۸۲/٤	~۱/۲۱ ms	x ۲/۵
الگوریتم پیشنهادی (بدون استفاده از جریان فراداده)	~۰/۰۳۷٤	Kbps۸۷۲/۸۱	۹۱/٤۷	~۱/ ۵٦ ms	x ۲/۹۲
الگوریتم پیشنهادی (با استفاده از جریان فراداده)	~۰/۰۳۷٤	Kbps۸۸۵/۰۸	۹۵/۸۲	~۱/ ٦۹ ms	x ۲/۹۱



شکل (۸) : مقایسه الگوریتم‌ها براساس متغیرهای مطرح شده

(الف)

(ب)

شکل (۹) : (الف) یک فریم از ویدئوی ورودی - (ب) فریم ورودی بازسازی شده

مراجع

[۱] C. Chen, S. Liu, and F. Wu, "Towards Intelligent Video Coding for Machines," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. ۳۰, no. ۷, pp. ۲۱۱٤–۲۱۲۸, ۲۰۲۰.

[۲] ISO/IEC JTC۱/SC۲۹/WG۱۱ (MPEG), "Draft Call for Proposals on Video Coding for Machines (VCM)," ۲۰۱۹.

[۳] Ultralytics, "YOLOv۸ Documentation," ۲۰۲۳.

[٤] H. Ye, et al., "Edge Intelligence for Video Analytics: Challenges and Opportunities," *IEEE Network*, vol. ۳۵, no. ۳, pp. ۳۸–٤۵, ۲۰۲۱.

[۵] Z. Chen, J. Li, and F. Wu, "Machine-Centric Video Coding: A New Paradigm," *IEEE Transactions on Image Processing,* vol. ۲۹, pp. ۳۳۳٦–۳۳۵۱, ۲۰۲۰.

[٦] Z. Ma, J. Liu, and F. Wu, "A Framework for Video Coding for Machines," *IEEE Communications Magazine*, vol. ۵۸, no. ۷, pp. ۵٤–۵۹, ۲۰۲۰.

[۷] S. Maćkowiak, et al., "Video coding for machines: Partial transmission of SIFT features," arXiv preprint arXiv:۲۲۰۱/۰۲٦۸۹, ۲۰۲۲..

[۸] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Proc. *IEEE Computer Society* Conference *on Computer Vision and Pattern Recognition,* ۲۰۱٦.

[۹] L. Zhang, et al., "ROI-Based Video Coding for Object Detection," Proc. *IEEE International Conference on Image Processing*, pp. ۳۳۱٤–۳۳۱۸, ۲۰۱۹.

[۱۰] M. Domański, O. Stankiewicz, S. Maćkowiak, S. Różek, T. Grajek, J. Szekiełda, D. Cywiński, "[VCM] Poznań University of Technology Proposal C in response to CfP on Video Coding for Machines," ISO/IEC JTC ۱/SC ۲۹/WG ٤ m٦۱۵۲۰ , Oct. ۲۰۲۰.

[۱۱] Video Coding for Machines," ISO/IEC JTC ۱/SC ۲۹/WG ٤ m٦۱۵۲۰, Oct. ۲۰۲۰.

[۱۲] X. Xu, S. Liu and Z. Li, "A Video Dataset for Learning-based Visual Data Compression and Analysis," ۲۰۲۱ *International Conference on Visual Communications and Image Processing* (*VCIP*), Munich, Germany, ۲۰۲۱.

[۱۳] M. Domański, O. Stankiewicz, S. Maćkowiak, J. Stankowski, S. Różek, M. Wawrzyniak, M. Lorkiewicz, T. Grajek, "Region-of-Interest-Based Video Coding for Machines," *IEEE International Conference on Visual*

Communications and Image Processing (*VCIP*), ۲۰۲۳, pp. ۱-٦.

[١٤] A. Balasubramaniam, F. Sunny and S. Pasricha, "R-TOSS: A Framework for Real-Time Object Detection using Semi-Structured Pruning," ۲۰۲۳ ٦۰*th ACM/IEEE Design Automation Conference* (*DAC*), San Francisco, CA, USA, ۲۰۲۳.

[١٥] Y. Zhang, X. Wu, and L. Li, "A lightweight model of underwater object detection based on YOLOv٨n for an edge computing platform," *Journal of Marine Science and Engineering*, vol. ١٢, no. ٥, p. ٦٩٧, ۲۰۲٤.

[١٦] P. Wang and J. Li, "Evaluation of Modern Interpolation and Resampling Filters for Image Scaling," *Journal of Mathematical Imaging and Vision*, ۲۰۲۱.

[١٧] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, S. Venkatesh, and H. Wu, "Mixed precision training," *International Conference on Learning Representations* (*ICLR*), ۲۰۱۸.

[١٨] T. Gale, E. Elsen, and S. Hooker, "The state of sparsity in deep neural networks," arXiv preprint arXiv:١٩٠٢/٠٩٥٧٤, ۲۰۱۹.

[١٩] M. Domański, O. Stankiewicz, S. Maćkowiak, S. Różek, T. Grajek, J. Szekiełda, and D. Cywiński, "[VCM] Poznań University of Technology Proposals A and B in response to CfP on Video Coding for Machines," ISO/IEC JTC ١/SC ٢٩/WG ٤, Document m٦١٥١٩ , Oct. ۲۰۲۰.

[٢٠] CfP on Video Coding for Machines," ISO/IEC JTC ١/SC ٢٩/WG ٤, Document m٦١٥١٩, Oct. ۲۰۲۰.

[٢١] H. Yang, et al., "Scalable Video Coding for Human and Machine Vision," *IEEE Transactions on Multimedia*, vol. ٢٢, no. ١٠, pp. ٢٥٧٣–٢٥٨٥, ۲۰۲۰.

[٢٢] N. Carion, et al., "End-to-End Object Detection with Transformers," *The European Conference on Computer Vision* (*ECCV*), ۲۰۲۰.

[٢٣] 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. ٣٩, no. ٦, pp. ١١٣٧–١١٤٩, ۲۰۱۷.

[٢٤] ITU-T Rec. H.٢٦٥ and ISO/IEC ٢٣٠٠٨-٢ HEVC, "High Efficiency Video Coding," ۲۰۱۳.

[٢٥] X. Xu, et al., "Efficient Video Analytics with Deep Learning on Edge Devices," *ACM Computing Surveys*, vol. ٥٣, no. ٣, pp. ١–٣٦, ۲۰۲۱.

[٢٦] ITU-T Rec. H.٢٦٤ and ISO/IEC ١٤٤٩٦-١٠ AVC, "Advanced Video Coding for Generic Audiovisual Services," ۲۰۱۰.

[٢٧] K. Iida, T. Moriyoshi, and K. Chono, "[VCM] Improvement of luma enhancement process in decoder complexity," ISO/IEC JTC ١/SC ٢٩/WG ٤, m٧٢١٨٨, Apr. ۲۰۲٥.

[٢٨] S. Raschka, "Noteworthy AI Research Papers of ۲۰۲٤ (Part Two)," *Amyris Switzerland*, Jan. ۲۰۲٥.

[٢٩] M. Jani, J. Fayyad, Y. Al-Younes, and H. Najjaran, "Model Compression Methods for YOLOv٥: A Review," arXiv, Jul. ۲۰۲۳.

[٣٠] A. Chan, A. Gujarati, K. Pattabiraman, and S. Gopalakrishnan, "The fault in our data stars: Studying mitigation techniques against faulty training data in machine learning applications," ٥٢*nd Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.* (*DSN*), Jun. ۲۰۲۲, pp. ١٦٣–١٧١.

[٣١] M. A. Santana, R. Calinescu, and C. Paterson, "Risk-aware real-time object detection," ١٨*th Eur. Dependable Comput. Conf.* (*EDCC*), Sep. ۲۰۲۲, pp. ١٠٥–١٠٨.

[٣٢] W. Gao, X. Xu, M. Qin, and S. Liu, "An open dataset for video coding for machines standardization," *IEEE Int. Conf. Image Process.* (*ICIP*), Oct. ۲۰۲۲, pp. ٤٠٠٨ ٤٠١٢.

بهار خدیوی بروجنی دانش‌آموخته مهندسی کامپیوتر در مقطع کارشناسی از دانشگاه شهرکرد (فارغ‌التحصیل ۱۳۹۶) و فارغ التحصیل کارشناسی ارشد مهندسی کامپیوتر گرایش معماری دستگاه‌های کامپیوتری در دانشگاه صنعتی خواجه نصیرالدین طوسی، که زمینه‌های پژوهشی و علاقه‌مندی‌های او شامل کدگذاری ویدئو برای ماشین‌ها، فشرده‌سازی ویدئو، بهینه‌سازی مدل‌های یادگیری عمیق و بینایی ماشین برای کاربردهای بلادرنگ می‌باشد.

هدی رودکی استادیار دانشکده مهندسی کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی است. ایشان کارشناسی خود را در رشته مهندسی کامپیوتر از دانشگاه تهران در سال ۱۳۸۳، کارشناسی ارشد را در معماری کامپیوتر از دانشگاه صنعتی شریف در سال ۱۳۸۶ و دکتری را نیز در معماری کامپیوتر از دانشگاه تهران در سال ۱۳۹۳ دریافت کرده‌اند. زمینه‌های پژوهشی اصلی او شامل فشرده‌سازی ویدئو با استفاده از روش‌های مبتنی بر هوش مصنوعی، ارزیابی کیفیت ویدئو و بینایی ماشین است.