مقاله پژوهشی

مجله ماشین‌بینایی و پردازش تصویر

دوره ۱۳، شماره ۲، پاییز ۱۴۰۵

شناسایی رفتارهای خشونت‌آمیز در ویدئوهای نظارتی با YOLOv11 بهبودیافته با STN و بهینه‌سازی ماژول‌های استخراج ویژگی

افسانه شمس الدینی-فرسنگی[۱]، مصطفی قاضی‌زاده-احسائی[۲]

چکیده

با افزایش چالش‌های امنیتی در محیط‌های نظارتی و فضاهای پرتراکم، نیاز به سامانه‌های هوشمند برای تشخیص سریع و دقیق رفتارهای خشونت‌آمیز بیش از پیش احساس می‌شود. یکی از کاربردهای بالقوه چنین سامانه‌هایی، محیط‌های کنترل‌شده‌ای مانند زندان‌ها است که تشخیص به‌موقع درگیری‌ها اهمیت بالایی دارد. در پژوهش حاضر، یک نسخه بهبودیافته از مدل تشخیص شیء YOLOv11 برای شناسایی رفتارهای خشونت‌آمیز ارائه شده است. در معماری پیشنهادی، از یک شبکه تبدیل مکانی برای افزایش تمرکز مدل بر نواحی مهم تصویر استفاده شده و همچنین با بهینه‌سازی بخش استخراج ویژگی‌ها و تقویت نمایش اطلاعات زمینه‌ای، توانایی مدل در تشخیص صحنه‌های پیچیده و پرتراکم بهبود یافته است. با توجه به محرمانه بودن داده‌های واقعی زندان، ارزیابی روش پیشنهادی بر روی مجموعه‌داده‌های عمومی انجام شد تا امکان مقایسه منصفانه با پژوهش‌های پیشین فراهم گردد. نتایج تجربی نشان داد مدل پیشنهادی به دقت متوسط ۹۲/۸ درصد در آستانه همپوشانی ۰/۵ و ۷۶/۱ درصد در بازه آستانه‌های همپوشانی ۰/۵ تا ۰/۹۵ دست یافته است که نسبت به مدل پایه بهبود قابل توجهی را نشان می‌دهد. همچنین ارزیابی بر روی مجموعه‌داده‌های نظارتی عمومی بیانگر پایداری مناسب روش پیشنهادی در شرایط مختلف تصویربرداری است. نتایج به‌دست‌آمده نشان می‌دهد که روش پیشنهادی ظرفیت مناسبی برای استفاده به‌عنوان یک راهکار پایه و قابل انتقال به سناریوهای نظارتی واقعی، از جمله محیط‌های زندان، پس از تنظیم و ارزیابی تکمیلی بر داده‌های همان دامنه دارد.

کلیدواژه‌ها

تشخیص خشونت، تشخیص شیء، YOLOv11، یادگیری عمیق، سامانه‌های نظارتی

۱- مقدمه

با گسترش زیرساخت‌های شهری و افزایش تراکم جمعیت در فضاهای عمومی، رخدادهای خشونت‌آمیز و درگیری‌های فیزیکی به یکی از چالش‌های جدی در حوزه امنیت عمومی تبدیل شده‌اند [۱]. در محیط‌های کنترل‌شده و پرتنش مانند زندان‌ها، این مسئله اهمیت دوچندان دارد؛ زیرا تأخیر در تشخیص یک درگیری می‌تواند منجر به گسترش بحران و آسیب‌های جدی شود. ازاین‌رو، توسعه سامانه‌های نظارتی هوشمند که بتوانند رفتارهای خشونت‌آمیز را به‌صورت سریع و قابل اعتماد شناسایی کنند، به‌عنوان یک نیاز کاربردی مطرح است.

تشخیص خشونت در ویدئوهای نظارتی با چالش‌های متعددی همراه است؛ از جمله تغییرات نور و شرایط محیطی، تفاوت زاویه و ارتفاع دوربین، پوشیدگی و هم‌پوشانی افراد، تراکم جمعیت، و کیفیت پایین تصاویر دوربین‌های نظارتی .علاوه بر این، وابستگی صرف به نیروی انسانی برای پایش مداوم تصاویر، به دلیل خستگی و خطای انسانی، در مقیاس عملیاتی قابل اتکا نیست. بنابراین،

این مقاله در ۲۰ بهمن ۱۴۰۴ دریافت و در ۲۳ شهریور ۱۴۰۵ پذیرفته شد.
این پژوهش با پشتیبانی مالی . اداره کل زندان‌های استان کرمان بر اساس قرارداد شماره ۱۷۹۳۲/۱۰/۴۳۷/۰۲ انجام شده است.

[۱] دانشکده مهندسی کامپیوتر، دانشگاه شهید باهنر کرمان.
رایانامه: ashamsadini29@gmail.com
[۲] استادیار، دانشکده مهندسی کامپیوتر، دانشگاه شهید باهنر کرمان.
رایانامه: mghazizadeh@uk.ac.ir
نویسنده مسئول: مصطفی قاضی‌زاده-احسائی

رویکردهای خودکار مبتنی بر بینایی ماشین می‌توانند نقش مهمی در افزایش کارایی سامانه‌های نظارت ایفا کنند.

یکی از موانع اصلی در توسعه روش‌های دقیق تشخیص خشونت، کمبود داده‌های برچسب‌خورده و قابل اعتماد در سناریوهای واقعی است. فرآیند برچسب‌گذاری ویدئوهای نظارتی به‌صورت دستی، زمان‌بر و پرهزینه بوده و در بسیاری از کاربردهای واقعی—از جمله زندان—دسترسی به داده‌ها با محدودیت‌های امنیتی و محرمانگی همراه است [۲]. این مسئله باعث می‌شود بخش قابل توجهی از پژوهش‌ها برای آموزش و ارزیابی مدل‌ها از مجموعه‌داده‌های عمومی استفاده کنند تا امکان بازتولید نتایج و مقایسه منصفانه با پژوهش‌های پیشین فراهم شود.

در سال‌های اخیر، روش‌های مبتنی بر یادگیری عمیق به دلیل توانایی بالا در استخراج ویژگی‌های پیچیده، در تشخیص رخدادهای خشونت‌آمیز مورد توجه قرار گرفته‌اند [۳]. برخی پژوهش‌ها از مدل‌های مکانی-زمانی)مانند ترکیب CNN و RNN یا شبکه‌های سه‌بعدی (برای تشخیص الگوهای رفتاری در توالی فریم‌ها بهره برده‌اند.

در مقابل، گروهی دیگر از روش‌ها از مدل‌های تشخیص شیء برای شناسایی افراد و نشانه‌های ظاهری درگیری استفاده می‌کنند. با وجود پیشرفت‌های چشمگیر، چالش‌هایی همچون هزینه محاسباتی بالا، افت عملکرد در صحنه‌های شلوغ و پویا، و حساسیت نسبت به تغییرات مکانی و پس‌زمینه همچنان پابرجاست. همچنین، قابلیت تعمیم بسیاری از مدل‌ها به محیط‌های جدید و دیده‌نشده محدود بوده و نیازمند راهکارهای مقاوم‌تر است.

با وجود پیشرفت قابل توجه روش‌های مبتنی بر یادگیری عمیق، تشخیص رفتارهای خشونت‌آمیز در ویدئوهای نظارتی همچنان با چالش‌های متعددی روبه‌رو است. بسیاری از روش‌های فضایی-زمانی دارای هزینه محاسباتی بالا بوده و برای کاربردهای بلادرنگ مناسب نیستند. از سوی دیگر، مدل‌های تشخیص شیء در صحنه‌های پرتراکم، دارای هم‌پوشانی شدید و تغییرات زاویه دید با افت عملکرد مواجه می‌شوند. علاوه بر این، محدودیت دسترسی به داده‌های واقعی خشونت‌آمیز موجب کاهش قابلیت تعمیم بسیاری از مدل‌ها شده است. بنابراین توسعه معماری‌هایی که ضمن حفظ سرعت پردازش، توانایی استخراج ویژگی‌های مقاوم و پایدار را داشته باشند، همچنان یک مسئله باز پژوهشی محسوب می‌شود.

به‌منظور پاسخ به این چالش‌ها، پژوهش‌های اخیر بر طراحی معماری‌های سبک‌وزن، بهینه‌سازی ساختار شبکه، افزایش داده و نیز به‌کارگیری سازوکارهای توجه و ماژول‌های تطبیقی متمرکز شده‌اند [۲]. در این میان، استفاده از ماژول‌هایی که امکان تمرکز بر نواحی مهم تصویر و کاهش اثر تغییرات مکانی را فراهم می‌کنند، می‌تواند به افزایش پایداری مدل در شرایط واقعی کمک نماید.

در این پژوهش، با هدف بهبود دقت و پایداری تشخیص رفتارهای خشونت‌آمیز در ویدئوهای نظارتی، یک معماری مبتنی بر YOLOv11ارائه می‌شود که برای عملکرد بهتر در صحنه‌های شلوغ و دارای هم‌پوشانی طراحی شده است. نوآوری‌های اصلی پژوهش به شرح زیر است:

- ادغام شبکه تبدیل مکانی (STN[1]) با YOLOv11: به‌منظور افزایش مقاومت در برابر تغییرات مکانی و تمرکز بهتر بر نواحی کلیدی تصویر.
- جایگزینی ماژول SPPF با ماژول ویژگی‌های پراکنده (SF[2]):با هدف تقویت استخراج ویژگی و کاهش کانال‌های کم‌اهمیت.
- بهبود مسیر استخراج ویژگی با ترکیب DRB[3] و ساختار C3k2: به‌منظور افزایش توان مدل در درک زمینه و مدیریت صحنه‌های پیچیده.

باقی‌مانده مقاله به‌صورت زیر سازمان‌دهی شده است: در بخش ۲ پژوهش‌های مرتبط مرور می‌شود، بخش ۳ روش پیشنهادی ارائه می‌گردد، بخش ۴ نتایج آزمایش‌ها و ارزیابی عملکرد گزارش می‌شود، و در نهایت بخش ۵ به جمع‌بندی و پیشنهادهای آینده اختصاص دارد.

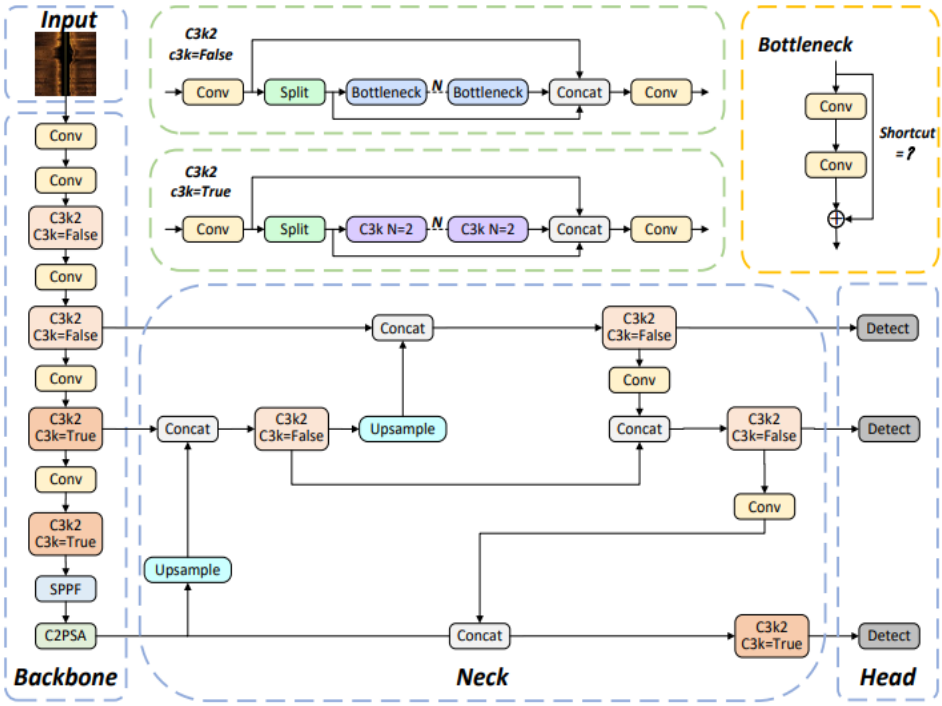
۲- کارهای مرتبط

در حال حاضر، روش‌های تشخیص خشونت عمدتاً به دو دسته تقسیم می‌شوند :تکنیک‌های سنتی یادگیری ماشین و روش‌های بینایی کامپیوتری مبتنی بر یادگیری عمیق .در مطالعات اولیه، روش‌های سنتی یادگیری ماشین بیشتر برای شناسایی خشونت استفاده می‌شدند. این روش‌ها معمولاً بر ویژگی‌های دستی و طبقه‌بندهای از پیش انتخاب‌شده متکی هستند تا وقوع خشونت را تشخیص دهند، که این رویکردها با محدودیت‌های متعددی همراه هستند.

به عنوان مثال، ویژگی‌های دستی اغلب نیاز به دانش قبلی از داده‌های ورودی دارند و بسیاری از الگوهای پیچیده مرتبط با رفتار خشونت‌آمیز به سختی قابل استخراج به صورت دستی هستند. علاوه بر این، دستیابی به داده‌های برچسب‌خورده کافی برای تشخیص خشونت چالش‌برانگیز است و این موضوع عملکرد طبقه‌بندهای آموزش‌دیده را کاهش می‌دهد. همچنین، روش‌های سنتی از قابلیت مقیاس‌پذیری لازم برای پردازش داده‌های ویدئویی

[1] Spatial Transformer Network
[2] Sparse Feature Module
[3] Dilated Residual Block

در مقیاس بزرگ برخوردار نیستند [۴] و بنابراین برای کاربردهای دنیای واقعی کمتر مناسب‌اند.



شکل ۱. معماری شبکهYOLOv11

در سال‌های اخیر، روش‌های یادگیری عمیق در زمینه تشخیص خشونت محبوبیت قابل توجهی یافته‌اند. به عنوان مثال، ژو و همکاران [۵] یک روش مبتنی بر یادگیری عمیق ارائه دادند که با استفاده از معماری شبکه عصبی پیچشی دومسیره[۱] ویژگی‌های مکانی ـ زمانی ویدئوهای نظارتی را استخراج کرده و با دقت طبقه‌بندی ٪۸۹ در یک مجموعه داده نظارتی به نتایج خوبی دست یافت.

به طور مشابه، در مطالعه‌ای دیگر [۶]، مدلی مبتنی بر یادگیری عمیق برای تشخیص خشونت در ویدئوهای شبکه‌های اجتماعی توسعه یافت. این مدل با استفاده از شبکه حافظه طولانی کوتاه‌مدت[۲] ویژگی‌های بصری را استخراج کرده و وابستگی‌های زمانی را ضبط می‌کند و دقت ۹۴/۹ % را بر روی همان مجموعه داده کسب کرده است. افزون بر این، پژوهشگران دیگری [۷,۸] نیز روش‌هایی مبتنی بر یادگیری عمیق برای تشخیص خشونت در ویدئوهای نظارتی شهری ارائه کرده‌اند.

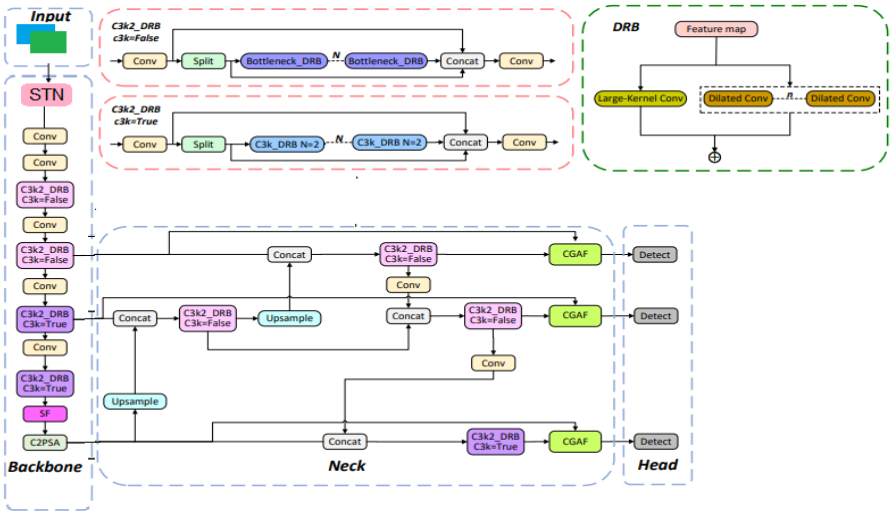
برای تشخیص درگیری‌ها و رویدادهای مشابه، پژوهشگران عمدتاً بر سه رویکرد اصلی یادگیری عمیق متمرکز شده‌اند: شبکه‌های عصبی پیچشی سه‌بعدی ، مدل‌های ترکیبی CNN-RNN و مدل‌های مبتنی بر .YOLO در بخش‌های بعدی، هر یک از این رویکردها به‌طور دقیق‌تر بررسی می‌شوند.

۱-۲- شبکه‌های عصبی پیچشی سه‌بعدی[۳]

یادگیری عمیق به‌طور مؤثری در حال رفع چالش‌ها در حوزه بینایی کامپیوتری است. معماری شبکه عصبی پیچشی سه‌بعدی ، که به‌طور خاص برای وظایف تحلیل ویدئو طراحی شده و نیازمند اطلاعات مکانی و زمانی است، نسخه‌ای توسعه‌یافته از شبکه‌های عصبی پیچشی دوبعدی محسوب می‌شود. این شبکه‌ها دنباله‌های ویدئویی را به‌عنوان ورودی پردازش می‌کنند و امکان یادگیری هم‌زمان از فریم‌های فضایی و روابط زمانی آن‌ها را فراهم می‌سازند. این معماری‌ها معمولاً شامل چندین لایه کانولوشن و پولینگ هستند که ویژگی‌های مکانی ـ زمانی را از دنباله ویدئو استخراج می‌کنند. خروجی این لایه‌های کانولوشن سپس از طریق لایه‌های کاملاً متصل و توابع فعال‌سازی عبور داده شده و پیش‌بینی نهایی تولید می‌شود.

برای نمونه، رمزان و همکاران [۹] از اطلاعات کلیدی استخراج شده از ویژگی‌های HOGبرای نمایندگی هر فریم در یک دنباله استفاده کرده‌اند .لیو و همکاران [۱۰] از معماری 3D-CNNبرای تشخیص صحنه‌های خشونت‌آمیز در ویدئوها بهره برده و نمونه‌برداری فریم‌ها را به‌عنوان یک مرحله پیش‌پردازش اعمال کرده‌اند.

[۱] dual-stream CNN
[2] LSTM
[۳] 3D Convolutional Neural Networks



شکل ۲. معماری مدل پیشنهادی

در همین حال، مطالعه [۱۱] یک چارچوب مبتنی بر Sparkبرای تشخیص صحنه‌های خشونت‌آمیز با استفاده از مدل LSTM دوطرفه پیشنهاد کرده است. به‌طور مشابه، ساموئل و همکاران [۱۲] از تکنیک‌های ادغام مدل‌ها استفاده کردند و در مطالعه [۹] نیز ترکیبی از 3D-CNNو ماشین بردار پشتیبان[۱] [۱۳] برای شناسایی رفتار خشونت‌آمیز در دنباله‌های ویدئویی به‌کار گرفته شده است.

۲-۲- ترکیبCNN-RNN

بسیاری از پژوهشگران معتقدند که داده‌های ویدئویی نمی‌توانند تنها با استفاده از شبکه‌های CNN به‌طور کامل پردازش شوند، زیرا وابستگی‌های زمانی بین فریم‌ها برای درک صحنه‌های پویا اهمیت بالایی دارند. برای رفع این مشکل، مدل‌های ترکیبی CNN-RNNبرای تشخیص رفتارهای غیرعادی در مجموعه داده‌های ویدئویی پیشنهاد شده‌اند.

مگدی و همکاران [۱۴] مدلی به نام Violence 4D معرفی کردند که مبتنی بر معماری 3D-CNN بوده و از ResNet50 به‌عنوان شبکه پایه استفاده می‌کند. در این مدل، جریان نوری متراکم[۲] برای برجسته‌سازی نواحی مهم در فریم‌های RGB به‌کار گرفته شد. مدل 4D-CNNپیشنهادی تعامل بین بخش‌های مکانی مختلف ویدئو را محدود می‌کند تا نمایش سطح بخش‌ها در مدل 3D-CNNبهبود یابد. علاوه بر این، بلوک‌های باقیمانده 4D نیز در معماری ادغام شدند تا قابلیت مدل‌سازی بلندمدت ارتقا یابد.

وانکدری و همکاران [۱۵] مدلی جدید به نام KianNet ارائه دادند که با هدف بهبود عملکرد معماری‌های ResNet50 وConvLSTM توسعه یافته است. با ادغام یک لایه توجه چندسری[۳] ، این مدل توانست فعالیت‌های خشونت‌آمیز در رویدادهای ضبط‌شده را به‌طور مؤثر شناسایی کند.

با این حال، تحلیل جامع روش‌های موجود نشان می‌دهد که بسیاری از رویکردهای فعلی هنوز با چالش‌ها و محدودیت‌های متعددی مواجه هستند. این الگوریتم‌ها غالباً نیازمند محاسبات سنگین هستند که اجرای بلادرنگ آن‌ها را دشوار می‌کند و در نتیجه ممکن است نیازهای عملی سامانه‌های نظارتی را به‌طور کامل برآورده نکنند

YOLO -۲-۳

رحیمه و همکاران [۱۶] مدل YOLOv11را به‌عنوان نسخه بهبودیافته YOLOv8 معرفی کردند که با تغییرات معماری، عملکرد بهتری در تشخیص اشیاء ارائه می‌دهد. همان‌طور که در شکل ۱ نشان داده شده است، این مدل همچنان از ماژول SPPF استفاده می‌کند و همزمان ساختارهای جدیدی مانند C2PSAو C3k2را در خود جای داده است.

در مرحله ورودی، YOLOv11از تکنیک انکر تطبیقی YOLOv8 [۴]به‌همراه روش‌های پیشرفته افزایش داده مانند Mosaicو Mix-up بهره می‌برد. با اتخاذ یک معماری بدون انکر ، مدل قادر است تصاویر با رزولوشن‌های متفاوت را به‌طور انعطاف‌پذیر پردازش کند Backbone. مسئول استخراج ویژگی‌ها در سطوح مختلف است و با جایگزینی ماژول C2f با C3k2و افزودن ماژول توجه فضایی (C2PSA) پس ازSPPF ، توانایی استخراج ویژگی‌های معنادار بهبود می‌یابد. در بخش Neck، مدل از C3k2 برای ادغام

[۱] SVM
[۲] Dense Optical Flow

[۳] Multi-head attention
[۴] Adaptive anchor

ویژگی‌ها در مقیاس‌های مختلف استفاده می‌کند. پارامترهای این ماژول امکان سازگاری با ساختارهای متفاوت مانند C2f یا C3 را فراهم می‌کنند.

در نهایت، ماژول Head با استفاده از تابع توزیع خطای کانونی[1] ، پیش‌بینی‌های کادر محدود[2] را بهینه می‌سازد و با تبدیل مختصات گسسته به فضای پیوسته، کادرهای نهایی تشخیص را تولید می‌کند.

اگرچه بسیاری از روش‌های تشخیص خشونت از مدل‌های فضایی-زمانی نظیر D-CNN۳، ConvLSTMو-CNN RNNاستفاده می‌کنند، هدف پژوهش حاضر ارائه یک چارچوب سبک‌وزن مبتنی بر تشخیص شیء برای کاربردهای بلادرنگ بوده است. از این رو، مقایسه تجربی عمدتاً با معماری‌های تشخیص شیء انجام شده است. با این حال، مقایسه مستقیم با روش‌های فضایی-زمانی می‌تواند در مطالعات آینده تصویر جامع‌تری از نقاط قوت و محدودیت‌های روش پیشنهادی ارائه دهد.

۳- مدل پیشنهادی

در این بخش، معماری شبکه پیشنهادی و اجزای کلیدی آن شامل شبکه تبدیل مکانی(STN) ، ماژول ویژگی‌های پراکنده (SF) و بلوک بازسازی باقیمانده (DRB) معرفی می‌شوند. هدف از افزودن این ماژول‌ها، افزایش پایداری مدل در مواجهه با تغییرات مکانی، هم‌پوشانی افراد، صحنه‌های پرتراکم و تغییرات سریع حرکت است. همچنین با ارائه تحلیل‌های بصری مبتنی بر نقشه‌های حرارتی، تأثیر هر ماژول بر تمرکز مدل بر نواحی مهم و تقویت استخراج ویژگی‌ها بررسی می‌شود.

۳-۱- ساختار کلی مدل پیشنهادی

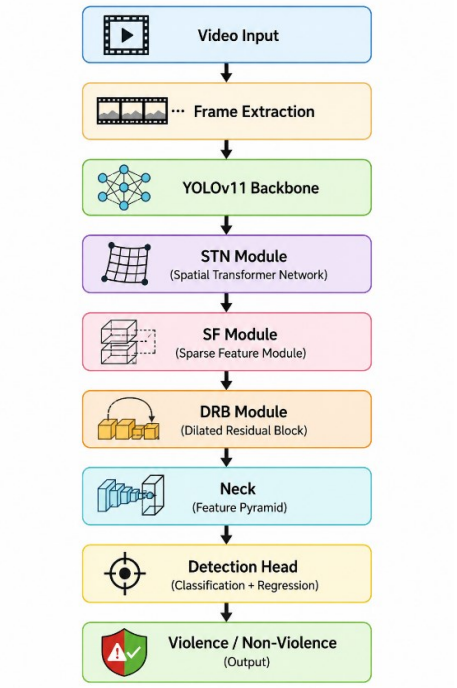
با وجود عملکرد مناسب معماری‌های مبتنی بر YOLO در تشخیص شیء، همچنان چالش‌هایی در سناریوهای نظارتی وجود دارد؛ از جمله کاهش دقت در صحنه‌های شلوغ، حساسیت به تغییرات مکانی و زاویه دید، و افت کیفیت ویژگی‌ها در شرایط دارای نویز و هم‌پوشانی .افزون بر این، در بسیاری از رخدادهای خشونت‌آمیز، نشانه‌های بصری کلیدی به‌صورت موضعی و در تعامل نزدیک افراد ظاهر می‌شوند و نیازمند تمرکز دقیق‌تر شبکه بر نواحی مهم هستند.

برای رفع این محدودیت‌ها، معماری پیشنهادی مطابق شکل ۲ ارائه شده است که سه جزء کلیدی را در ساختار YOLOv11 ادغام می‌کند:

۱) **ماژول STN:** این ماژول به‌منظور کاهش اثر تغییرات مکانی و هم‌ترازسازی ویژگی‌ها به‌کار گرفته شده است. برخلاف اعمال STN بر تصویر خام که می‌تواند هزینه‌بر باشد، در روش پیشنهادی STN بر روی نقشه‌های ویژگی اولیه اعمال می‌شود تا ضمن حفظ کارایی، سربار محاسباتی کاهش یابد.

۲) **ماژول SF:** این ماژول با تمرکز بر انتخاب کانال‌های مؤثر و کاهش افزونگی، موجب افزایش تمایز ویژگی‌ها و تقویت نمایش‌های مفید برای تشخیص می‌شود. در نتیجه، شبکه در لایه‌های میانی به ویژگی‌های کلیدی حساس‌تر شده و اطلاعات کم‌اهمیت تضعیف می‌گردد.

۳) **ماژول DRB:** در مرحله بازسازی و تلفیق ویژگی‌ها، DRBبا استفاده از کانولوشن‌های متسع (Dilated Convolution)، زمینه ادراکی شبکه را تقویت کرده و به مدل کمک می‌کند تعاملات فضایی پیچیده و الگوهای رخداد را بهتر نمایش دهد. این موضوع به‌ویژه در صحنه‌های پرتراکم و دارای هم‌پوشانی، موجب بهبود کیفیت استخراج ویژگی‌ها می‌شود.



شکل ۳ .روند کلی پردازش در مدل پیشنهادی از استخراج فریم تا تشخیص خشونت

در مجموع، این ماژول‌ها به‌گونه‌ای در معماری ادغام شده‌اند که ضمن حفظ ساختار اصلیYOLOv11 ، استخراج ویژگی‌ها تقویت شده و مدل نسبت به تغییرات مکانی و پیچیدگی صحنه مقاوم‌تر

[1] Distributed Focal Loss
[2] Bounding Box

گردد. جزئیات فنی هر ماژول و نحوه ادغام آن در شبکه در بخش‌های بعدی ارائه می‌شود.

لازم به ذکر است که نوآوری پژوهش حاضر صرفاً استفاده از ماژول‌هایSTN ، SF و DRB نیست، بلکه نحوه ادغام هدفمند این ماژول‌ها در معماری YOLOv11 برای افزایش پایداری مدل در صحنه‌های پرتراکم، دارای هم‌پوشانی و تغییرات مکانی است. در این چارچوب، STN برای هم‌ترازسازی ویژگی‌های اولیه، SF برای افزایش تمایزپذیری کانال‌های مؤثر و DRB برای تقویت زمینه ادراکی و نمایش ویژگی‌ها مورد استفاده قرار گرفته‌اند.

۱-۱-۳ راهبرد پردازش ویدئو

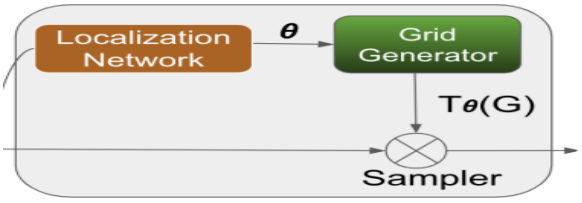
اگرچه هدف نهایی پژوهش حاضر تشخیص رفتارهای خشونت‌آمیز در ویدئوهای نظارتی است، اما معماری پیشنهادی بر پایه پردازش فریم‌های استخراج‌شده از ویدئو طراحی شده است. بدین منظور، هر ویدئو در مرحله پیش‌پردازش به مجموعه‌ای از فریم‌های مجزا تبدیل شده و سپس هر فریم به‌صورت مستقل توسط مدل پیشنهادی مورد تحلیل قرار می‌گیرد. بنابراین، فرآیند تشخیص در این پژوهش مبتنی بر یادگیری ویژگی‌های فضایی بوده و وابستگی‌های زمانی میان فریم‌های متوالی به‌صورت صریح مدل‌سازی نمی‌شوند.

انتخاب این راهبرد با هدف حفظ سرعت پردازش و امکان استفاده از معماری‌های تشخیص شیء بلادرنگ انجام شده است. از آنجا که بسیاری از سامانه‌های نظارتی نیازمند پاسخ سریع و پردازش پیوسته جریان ویدئو هستند، استفاده از یک چارچوب مبتنی بر تشخیص شیء می‌تواند مزایای عملیاتی قابل توجهی فراهم کند.

با این حال، رفتارهای خشونت‌آمیز ذاتاً دارای ماهیت زمانی بوده و بخشی از اطلاعات مرتبط با الگوی حرکت افراد در توالی فریم‌ها نهفته است. بنابراین، اگرچه روش پیشنهادی توانسته است با تقویت استخراج ویژگی‌های فضایی عملکرد مناسبی ارائه دهد، اما استفاده از سازوکارهای مدل‌سازی زمانی نظیرConvLSTM ، شبکه‌های عصبی پیچشی سه‌بعدی(3D-CNN) ، Temporal Shift Module (TSM)و معماری‌های مبتنی بر Transformerمی‌تواند در تحقیقات آینده به‌منظور بهره‌برداری از اطلاعات زمانی و بهبود بیشتر عملکرد مورد استفاده قرار گیرد.

۲-۳- شبکه تبدیل مکانی (Spatial Transformer Network – STN)

شبکه تبدیل مکانی (STN) یک ماژول قابل یادگیری و مشتق‌پذیر است که با تخمین یک تبدیل هندسی مناسب، امکان هم‌ترازسازی مکانی را برای ورودی فراهم می‌کند. این ماژول می‌تواند تغییرات ناشی از جابه‌جایی، چرخش و مقیاس را تا حدی جبران کرده و باعث شود شبکه در ادامه مسیر پردازش، بر نواحی مهم‌تر تمرکز کند.



شکل ۳. ماژول STN اعمال شده بر نقشه ویژگی اولیه

در این پژوهش، STNبر روی نقشه‌های ویژگی اولیه اعمال می‌شود تا ضمن بهبود پایداری مکانی، سربار محاسباتی نسبت به اعمال بر تصویر خام کاهش یابد.

درSTN ، پارامترهای تبدیل به‌صورت پویا و متناسب با ورودی تخمین زده می‌شوند و خروجی به شکل یک نقشه ویژگی هم‌ترازشده تولید می‌گردد. در صورتی که ورودی دارای چندین کانال باشد، تبدیل مکانی یکسان بر روی تمامی کانال‌های آن اعمال می‌شود.

مطابق شکل ۳، ماژول STN از سه بخش اصلی تشکیل شده است:

۱) **شبکه مکان‌یابی:(Localization Network)**

این شبکه با دریافت نقشه ویژگی ورودی، پارامترهای تبدیل هندسی را تخمین می‌زند. در این پژوهش، تبدیل از نوع affineدر نظر گرفته شده است و پارامترهای آن با θ نمایش داده می‌شوند. تبدیل affine به‌صورت یک ماتریس 2×3 مدل می‌شود که شامل مؤلفه‌های انتقال، چرخش و مقیاس است. پارامترهای θ برای هر نمونه ورودی تخمین زده می‌شوند و سپس روی تمامی کانال‌های نقشه ویژگی اعمال می‌گردند.

۲) **مولد شبکه نمونه‌برداری:(Grid Generator)**

با استفاده از پارامترهای θ، یک شبکه نمونه‌برداری تولید می‌شود که مشخص می‌کند هر نقطه از خروجی از کدام مختصات متناظر در ورودی استخراج شود.

نمونه‌بردار[۱]:

در این مرحله، با استفاده از شبکه نمونه‌برداری، مقادیر نقشه ویژگی خروجی از ورودی استخراج می‌شوند. این فرآیند معمولاً با درون‌یابی دوخطی انجام می‌شود تا کل مسیر محاسباتی مشتق‌پذیر باقی بماند و امکان آموزش انتهابه‌انتها فراهم گردد.

۱-۲-۳- شبکه مکان‌یابی (Localization Network) :

شبکه مکان‌یابی، نقشه ویژگی ورودی را دریافت می‌کند. این ورودی به‌صورت $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$ در نظر گرفته می‌شود که در آن

[۱] Sampler

H، W و C به‌ترتیب نمایانگر ارتفاع، عرض و تعداد کانال‌ها هستند. هدف شبکه مکان‌یابی، تخمین پارامترهای تبدیل هندسی θ به‌صورت پویا و وابسته به ورودی است. بدین منظور، ابتدا ویژگی‌های استخراج‌شده از $\mathbf{U}$ از طریق چند لایه کانولوشن و/یا لایه‌های تمام‌متصل پردازش شده و در نهایت به یک لایه رگرسیون منتقل می‌شوند تا پارامترهای تبدیل به‌صورت زیر تولید گردد: $\theta = f_{\text{loc}}(\mathbf{U})$ که در آن f_{loc} شبکه مکان‌یابی و θ پارامترهای تبدیل affine است. سپس θ برای تولید شبکه نمونه‌برداری و اعمال تبدیل بر روی نقشه ویژگی ورودی مورد استفاده قرار می‌گیرد. لازم به ذکر است که پارامترهای تبدیل θ برای هر نمونه ورودی تخمین زده شده و همان تبدیل به‌صورت یکسان بر روی تمام کانال‌های نقشه ویژگی اعمال می‌شود.

ابعاد θ بسته به نوع تبدیل انتخابی متفاوت است. در این پژوهش، تبدیل از نوع affine در نظر گرفته شده و در نتیجه θ شامل شش پارامتر است. تابع f_{loc} می‌تواند با یک شبکه تمام‌متصل یا یک شبکه عصبی پیچشی سبک‌وزن پیاده‌سازی شود، با این شرط که در انتهای آن یک لایه رگرسیون برای تولید پارامترهای تبدیل وجود داشته باشد.

۳-۲-۲- شبکه نمونه‌برداری پارامتری (Parametric Sampling Grid)

برای اعمال تبدیل مکانی، هر عنصر از نقشه ویژگی خروجی به یک موقعیت متناظر در نقشه ویژگی ورودی نگاشت می‌شود. در اینجا منظور از «پیکسل»، یک عنصر (Element) در نقشه ویژگی است و لزوماً معادل پیکسل تصویر خام نیست. فرض می‌شود نقشه ویژگی ورودی به‌صورت $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$ و خروجی به‌صورت $\mathbf{V} \in \mathbb{R}^{H' \times W' \times C}$ تعریف می‌شود، که در آن H' و W' به‌ترتیب ارتفاع و عرض خروجی بوده و تعداد کانال‌ها ثابت و برابر با C در نظر گرفته می‌شود .شبکه نمونه‌برداری هدف به‌صورت یک شبکه منظم از نقاط تعریف می‌شود:

$$G = \{(x_i^t, y_i^t) \mid i = 1, \dots, H'W'\} \quad (۱)$$

که در آن (x_i^t, y_i^t) مختصات نقاط روی نقشه ویژگی خروجی هستند. سپس برای هر نقطه هدف، مختصات متناظر آن روی نقشه ویژگی ورودی (x_i^s, y_i^s) با استفاده از پارامترهای تبدیل θ محاسبه می‌شود.

در این پژوهش، تبدیل از نوع Affine دوبعدی در نظر گرفته شده است و رابطه نگاشت مختصات به صورت زیر تعریف می‌شود:

$$\begin{bmatrix} x_i^s \\ y_i^s \end{bmatrix} = \mathbf{T}_\theta \begin{bmatrix} x_i^t \\ y_i^t \\ 1 \end{bmatrix} \quad (۲)$$

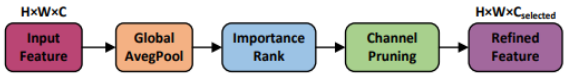
که در آن $\mathbf{T}_\theta$ یک ماتریس 2×3 شامل پارامترهای θ بوده و مختصات (x_i^s, y_i^s) نقاط متناظر در نقشه ویژگی ورودی را تولید می‌کند. بدین ترتیب، شبکه نمونه‌برداری تعیین می‌کند هر عنصر از خروجی از کدام موقعیت در ورودی استخراج شود.

استفاده از این سازوکار باعث می‌شود ویژگی‌های مکانی در نقشه ویژگی ورودی به‌صورت پویا هم‌تراز شده و اثر تغییرات مکانی نظیر جابه‌جایی، چرخش یا تغییر مقیاس کاهش یابد. در نتیجه، شبکه در مراحل بعدی استخراج ویژگی، تمرکز دقیق‌تری بر نواحی مهم داشته و پایداری مدل در صحنه‌های پرتراکم و دارای هم‌پوشانی افزایش می‌یابد.

۳-۳- ماژول ویژگی‌های پراکنده (Sparse Feature – SF)

در صحنه‌های پرتراکم و دارای هم‌پوشانی، بخشی از کانال‌های ویژگی در مدل‌های تشخیص شیء می‌توانند اطلاعات کم‌اهمیت یا تکراری تولید کنند که موجب کاهش تمرکز شبکه بر نشانه‌های کلیدی رخداد می‌شود. به‌منظور افزایش تمایزپذیری ویژگی‌ها و کاهش افزونگی کانالی، ماژول ویژگی‌های پراکنده (SF) در این پژوهش معرفی شده است. این ماژول با تقویت کانال‌های مؤثر و تضعیف کانال‌های کم‌اهمیت، کیفیت نمایش ویژگی‌ها را بهبود داده و باعث می‌شود شبکه در ادامه مسیر پردازش، حساسیت بیشتری نسبت به نواحی مهم داشته باشد. ساختار کلی ماژول SF در شکل ۴ نشان داده شده است.

فرض می‌شود نقشه ویژگی ورودی به صورت زیر تعریف شود: $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ که در آن H و W ابعاد مکانی و C تعداد کانال‌ها است.



شکل ۴. ساختار ماژولSF

ماژول SF از سه گام اصلی تشکیل شده است:

۱. تجمیع ویژگی‌های سراسری
۲. رتبه‌بندی اهمیت کانال‌ها
۳. انتخاب کانال‌های مؤثر و بازسازی نقشه ویژگی

مرحله ۱: تجمیع ویژگی‌های سراسری (Global Feature Aggregation)

شکل ۵. مجموعه داده تشخیص رفتار خشونت‌آمیز: نمونه‌هایی از مجموعه‌داده هاکی

در این مرحله، عملیات میانگین‌گیری سراسری بر روی هر کانال اعمال می‌شود تا یک نمایش فشرده از پاسخ کانال به‌دست آید. ویژگی سراسری کانال c به صورت زیر تعریف می‌شود:

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j) \quad (۳)$$

که در آن $X_c(i,j)$ مقدار عنصر مکانی (i,j) در کانال c است. بدین ترتیب بردار $\mathbf{z} \in \mathbb{R}^C$ به‌عنوان توصیف‌گر اولیه اهمیت کانال‌ها حاصل می‌شود.

مرحله ۲: رتبه‌بندی اهمیت کانال‌ها (Channel Importance Ranking)

برای رتبه‌بندی کانال‌ها، یک امتیاز اهمیت برای هر کانال بر اساس مقادیر $\mathbf{z}$ محاسبه می‌شود. در این پژوهش، به‌منظور افزایش پایداری، میانگین‌گیری بر روی نمونه‌های یک دسته (در زمان آموزش) به‌صورت زیر در نظر گرفته می‌شود:

$$\bar{z}_c = \frac{1}{B}\sum_{b=1}^{B} z_c^{(b)} \quad (۴)$$

که در آن B اندازه دسته، b شاخص نمونه و $\bar{z}_c$ میانگین ویژگی سراسری کانال c در کل دسته است.

در مرحله آزمون/استنتاج که معمولاً اندازه دسته برابر ۱ است، مقدار $\bar{z}_c$ با مقدار همان نمونه جایگزین می‌شود و بنابراین روش بدون وابستگی به اندازه دسته قابل استفاده است.

مرحله ۳: انتخاب کانال و بازسازی نقشه ویژگی (Channel Selection & Reconstruction)

پس از رتبه‌بندی، کانال‌ها بر اساس مقدار $\bar{z}_c$ مرتب شده و k کانال با بالاترین مقدار انتخاب می‌شوند:

$$\mathcal{J} = \text{Top-}k(\bar{\mathbf{z}}) \quad (۵)$$

که در آن $\mathcal{J}$ مجموعه شاخص‌های کانال‌های منتخب است. سپس نقشه ویژگی خروجی با نگه‌داشتن کانال‌های انتخاب‌شده بازسازی می‌شود:

$$\mathbf{Y} = \text{Gather}(\mathbf{X}, \mathcal{J}), \mathbf{Y} \in \mathbb{R}^{k\times H\times W} \quad (۶)$$

در نهایت، برای حفظ سازگاری ابعادی با مسیر اصلی شبکه، خروجی ماژول به شکل زیر در ادامه پردازش استفاده می‌شود:

$$\Upsilon = \phi(\mathbf{Y}) \quad (۷)$$

که در آن $\phi(\cdot)$ یک نگاشت بازسازی (مانند کانولوشن 1×1) جهت تنظیم ابعاد کانالی و ادغام در معماری اصلی است.

این فرآیند تضمین می‌کند تنها کانال‌های مهم‌تر در نقشه ویژگی باقی بمانند و بدین ترتیب، تمرکز شبکه بر ویژگی‌های کلیدی افزایش یابد.

۳-۴- ماژول DRB (Dilated Residual Block)

در سناریوهای نظارتی، رخدادهای خشونت‌آمیز اغلب در صحنه‌های پرتراکم و دارای هم‌پوشانی رخ می‌دهند و تشخیص صحیح آن‌ها نیازمند درک بهتر زمینه و تعاملات فضایی بین افراد است. ازاین‌رو، برای رفع محدودیت میدان دید مؤثر در بخش‌هایی از YOLOv11 و تقویت نمایش ویژگی‌ها در شرایط پیچیده، ماژول بلوک بازسازی باقیمانده متسع (DRB) معرفی شده است. این ماژول، کانولوشن‌های متسع [۱۶] (Dilated Convolution) را با راهبرد بازپارامتردهی [۱۷] ترکیب می‌کند تا ضمن بهبود توان استخراج ویژگی، هزینه محاسباتی در مرحله استنتاج کنترل شود.

کانولوشن متسع با ایجاد فاصله بین عناصر هسته کانولوشن، میدان دید را افزایش می‌دهد بدون آنکه تعداد پارامترهای هسته به‌صورت مستقیم افزایش یابد. همچنین، بازپارامتردهی امکان تبدیل ساختار چندشاخه در زمان آموزش به یک ساختار تک‌شاخه بهینه در زمان استنتاج را فراهم می‌کند.

در طول آموزش، ماژول DRB از چند شاخه کانولوشن موازی تشکیل می‌شود که هر شاخه از نرخ اتساع r و اندازه هسته متفاوت استفاده می‌کند. شاخه اصلی با Conv_0 نمایش داده می‌شود که دارای وزن $\mathbf{W}_0$ و بایاس $\mathbf{b}_0$ بوده و خروجی آن پس از عبور از نرمال‌سازی دسته‌ای (BN) استفاده می‌شود. شاخه‌های دیگر با $\text{Conv}_{k,r}$ نمایش داده می‌شوند که بیانگر شاخه k-ام با نرخ اتساع r و پارامترهای $\mathbf{W}_{k,r}$ و $\mathbf{b}_{k,r}$ هستند.[۱۸]

فرض می‌شود ورودی و خروجی ماژول به‌ترتیب با $\mathbf{x}$ و $\mathbf{y}$ نمایش داده شوند: $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{B\times C\times H\times W}$ که در آن B اندازه دسته، C تعداد کانال‌ها و H و W ابعاد مکانی نقشه ویژگی هستند.

در مرحله آموزش، خروجی DRB به صورت زیر تعریف می‌شود:

$$\mathbf{y} = \mathbf{x} + \mathrm{BN}_0(\mathrm{Conv}_0(\mathbf{x})) + \sum_{k,r} \mathrm{BN}_{k,r}\,(\mathrm{Conv}_{k,r}(\mathbf{x})) \quad (۸)$$

هر عمل کانولوشن به صورت زیر مدل می‌شود:

$$\mathrm{Conv}(\mathbf{x}; \mathbf{W}, \mathbf{b}) = \mathbf{W} * \mathbf{x} + \mathbf{b} \quad (۹)$$

که در آن $*$عمل کانولوشن را نشان می‌دهد.

برای افزایش کارایی در مرحله استنتاج، لایه BN هر شاخه در کانولوشن متناظر آن ادغام می‌شود. خروجی BN برای ورودی $\mathbf{u}$ به صورت زیر است:

$$\mathrm{BN}(\mathbf{u}) = \gamma \frac{\mathbf{u} - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \quad (۱۰)$$

که در آن μ و σ^2 میانگین و واریانس برآوردشده، γ و β پارامترهای یادگرفتنی BN و ε یک مقدار کوچک برای جلوگیری از تقسیم بر صفر است.

با ادغام BN در کانولوشن، پارامترهای معادل کانولوشن به صورت زیر به‌دست می‌آیند:

$$\mathbf{W}' = \frac{\gamma}{\sqrt{\sigma^2 + \varepsilon}} \mathbf{W}, \mathbf{b}' = \frac{\gamma}{\sqrt{\sigma^2 + \varepsilon}} (\mathbf{b} - \mu) + \beta \quad (۱۱)$$

در مرحله استنتاج، شاخه‌های موازی ادغام شده و به یک کانولوشن منفرد تبدیل می‌شوند. برای این منظور، هسته‌های شاخه‌های متسع در یک هسته بزرگ‌تر جاسازی شده و سپس با یکدیگر جمع می‌شوند. وزن و بایاس نهایی به صورت زیر تعریف می‌شوند:

$$\mathbf{W}_d = \mathbf{W}'_0 + \sum_{k,r} \mathcal{E}_{k,r}\,(\mathbf{W}'_{k,r}), \mathbf{b}_d = \mathbf{b}'_0 + \sum_{k,r} \mathbf{b}'_{k,r} \quad (۱۲)$$

که در آن $\mathcal{E}_{k,r}(\cdot)$ عملگر جاسازی[۱] است که هسته کانولوشن متسع با نرخ r را به‌صورت صحیح در هسته بزرگ‌تر $\mathbf{W}_d$ قرار می‌دهد تا اثر فضایی شاخه‌ها در قالب یک کانولوشن معادل حفظ شود.

در نهایت، خروجی ماژول DRB در مرحله استنتاج به صورت زیر محاسبه می‌شود:

$$\mathbf{y} = \mathbf{x} + \mathrm{Conv}(\mathbf{x}; \mathbf{W}_d, \mathbf{b}_d) \quad (۱۳)$$

این بازپارامتردهی موجب می‌شود ماژول DRB در زمان استنتاج با یک مسیر کانولوشن منفرد اجرا شود و در نتیجه، ضمن حفظ مزایای چندشاخه بودن در آموزش، کارایی اجرا افزایش یابد.

[۱] Embedding

۴- آزمایش‌ها و تحلیل نتایج

۴-۱- مجموعه داده

هدف نهایی این پژوهش کاربرد در محیط زندان است؛ بااین‌حال، داده‌های زندان به‌علت مسائل امنیتی و حقوقی قابل انتشار و استفاده عمومی نیستند و این موضوع مانع از گزارش نتایج به‌صورت بازتولیدپذیر می‌شود. بنابراین، برای جلوگیری از سوگیری ارزیابی و فراهم‌کردن امکان تکرار آزمایش‌ها توسط سایر پژوهشگران، از مجموعه‌داده عمومی Hockey Fight استفاده شد که در ادبیات موضوع به‌عنوان یک بنچمارک رایج برای تشخیص درگیری شناخته می‌شود. به این ترتیب، روش پیشنهادی در یک چارچوب استاندارد سنجیده شده و امکان مقایسه مستقیم با روش‌های موجود فراهم می‌گردد.

هرچند سناریوی مجموعه‌داده منتخب دقیقاً منطبق بر محیط زندان نیست، اما از آنجا که مسئله اصلی (تشخیص الگوهای بصریِ درگیری و خشونت) مشترک است، این بنچمارک امکان سنجش اولیه و مقایسه‌پذیر روش را فراهم می‌کند .

بدین منظور، داده‌های آزمایش بر اساس مجموعه‌داده Hockey Fight Dataset [۱۹] آماده‌سازی شد. این مجموعه‌داده شامل ۱۰۰۰ کلیپ ویدئویی از مسابقات NHL است که از زوایای مختلف ثبت شده‌اند و شامل ۵۰۰ کلیپ دارای درگیری/خشونت و ۵۰۰ کلیپ بدون رخداد خشونت‌آمیز می‌باشد.

به‌منظور ارزیابی بهتر قابلیت تعمیم مدل و کاهش وابستگی نتایج به یک مجموعه‌داده خاص، علاوه بر مجموعه‌دادهHockey Fight، از مجموعه‌داده RWF-2000 [۲۷] نیز استفاده شد.این مجموعه‌داده شامل ۲۰۰۰ کلیپ ویدئویی واقعی ثبت‌شده توسط دوربین‌های نظارتی است که بدون اعمال تغییرات چندرسانه‌ای جمع‌آوری شده‌اند و شرایط واقعی‌تری از محیط‌های نظارتی را بازنمایی می‌کنند.

مجموعه‌داده RWF-2000 شامل دو کلاس خشونت‌آمیز و غیرخشونت‌آمیز است و چالش‌هایی نظیر کیفیت پایین تصویر، نور نامناسب، تراکم افراد، هم‌پوشانی و فاصله زیاد افراد از دوربین را در بر می‌گیرد. به همین دلیل، این مجموعه‌داده به‌عنوان یکی از معیارهای استاندارد ارزیابی سامانه‌های تشخیص خشونت شناخته می‌شود.

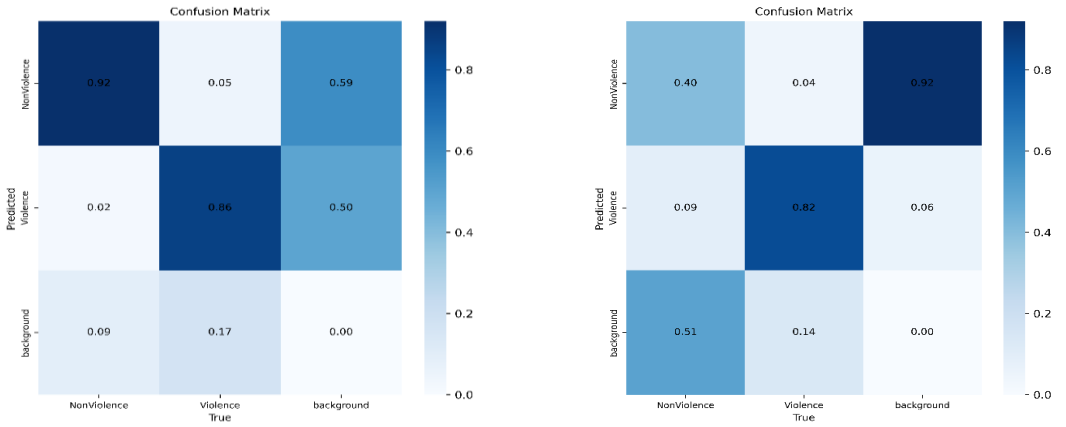
در این پژوهش، همانند مجموعه‌دادهHockey Fight ، فریم‌ها با استفاده از کتابخانه OpenCV استخراج شده و از هر پنج فریم یک فریم نمونه‌برداری شد

۴-۱-۱- استخراج فریم و آماده‌سازی داده

ابتدا کلیپ‌های دارای کیفیت مناسب انتخاب شدند و استخراج فریم با استفاده از کتابخانه OpenCV انجام گرفت؛ به‌گونه‌ای که

جدول ۱. مقایسه عملکرد مدل‌های مختلف بر روی مجموعه داده هاکی

مدل	mAP@50 (%)	mAP@[0.5:0.95] (%)	Parameters/M	FLOPs/G	FPS
[۲۱]YOLOv8l	۸۴/۷	۶۹/۳	۴۳/۶	۱۶۴/۸	۴۲/۶
[۲۱]YOLOv8m-World	۸۳/۵	۶۶/۳	۲۹	۸۹/۹	۵۷/۸
[۲۰]YOLOv5m	۸۲/۴	۶۲/۸	۲۵	۶۴	۷۳/۵
[۲۲]YOLOv3-tiny	۷۷/۶	۵۳/۲	۱۲/۱	۱۸/۹	۱۹۲/۳
[۲۳]RT-DETR-L	۹۰/۴	۷۵/۴	۳۱/۹	۱۰۸	۶۳/۳
[۲۴]GELAN-C	۹۱/۷	۷۴/۹	۲۵/۲	۱۰۱/۸	۴۷/۶
[۲۵]YOLOv9-C	۹۰/۶	۷۰/۶	۵۰/۷	۲۳۶/۶	۲۴/۸
[۲۶]Improvement GENLAN-C	۹۲/۶	۷۵	۲۲/۱	۸۸/۹	۴۷/۸
روش پیشنهادی	**۹۲/۸**	**۷۶/۱**	**۱۲/۵**	**۹۱/۱**	**۴۵/۹**



(الف). روش پیشنهادی

(ب).YOLOv11s

شکل ۶. ماتریس سردرگمی

از هر ۵ فریم، یک فریم نمونه‌برداری شد. در ادامه، فریم‌های استخراج‌شده از نظر وضوح رخداد و قابل‌مشاهده بودن تعاملات کلیدی (مانند برخورد فیزیکی) پالایش شدند. این مرحله منجر به انتخاب ۱۶۰۰ فریم اولیه قابل استفاده به‌عنوان هسته اصلی داده‌های حاشیه‌نویسی شد.

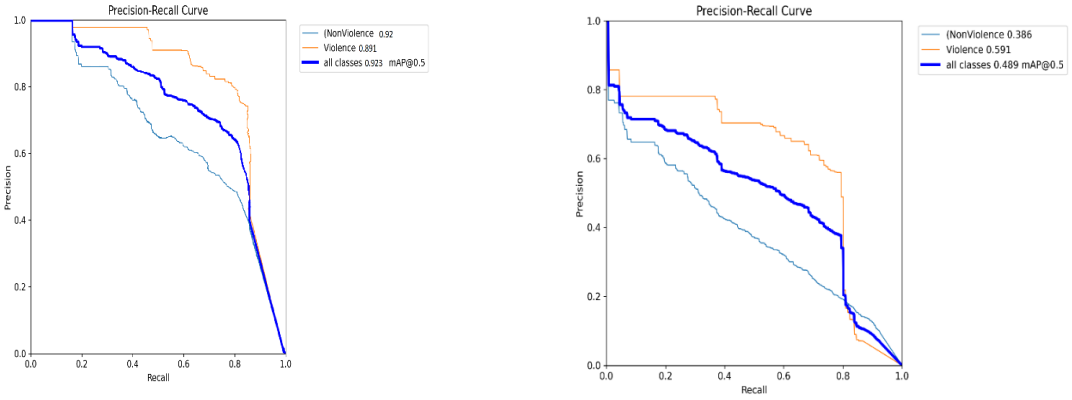
سپس برای افزایش پوشش و تنوع نمونه‌ها، استخراج فریم در بازه‌های زمانی مختلف کلیپ‌ها ادامه یافت و پس از حذف نمونه‌های تکراری یا کم‌کیفیت، در مجموع ۵۹۸۷ فریم نهایی برای برچسب‌گذاری انتخاب گردید. بنابراین:

- **۱۶۰۰ نمونه** اشاره به فریم‌های اولیه منتخب پس از پالایش اولیه دارد؛
- **۵۹۸۷ نمونه** تعداد نهایی فریم‌های حاشیه‌نویسی‌شده و مورد استفاده در آزمایش‌ها است.

۲-۱-۴- حاشیه‌نویسی و جلوگیری از نشت اطلاعات

برای حاشیه‌نویسی، از ابزار LabelImg نسخه ۱/۸/۶ استفاده شد و برای هر فریم، کلاس‌های خشونت‌آمیز و غیرخشونت‌آمیز برچسب‌گذاری شدند و فایل‌های حاشیه‌نویسی متناظر تولید گردید. نمونه‌ای از تصاویر مجموعه‌داده در شکل ۵ نمایش داده شده است.

از آنجا که فریم‌های متوالی در یک کلیپ بسیار مشابه‌اند، تقسیم تصادفی در سطح فریم می‌تواند منجر به نشت اطلاعات و ایجاد نتایج خوش‌بینانه شود. به همین دلیل، تقسیم‌بندی داده‌ها به‌صورت ویدئو-محور انجام شد.



(a) Proposed Method (b) YOLOv11s

شکل ۷ .منحنی های Precision–Recall

به‌گونه‌ای که فریم‌های مربوط به هر کلیپ تنها در یکی از مجموعه‌های آموزش، اعتبارسنجی یا آزمون قرار گرفتند و هیچ کلیپی به‌طور هم‌زمان در چند مجموعه حضور ندارد. بدین ترتیب، شباهت زمانی بین آموزش و آزمون کنترل شده و ارزیابی منصفانه‌تر انجام می‌شود.

در نهایت، داده‌ها به سه بخش آموزش، اعتبارسنجی و آزمون تقسیم شدند. به‌منظور کاهش احتمال بیش‌برازش و افزایش قابلیت تعمیم مدل، علاوه بر تقسیم‌بندی ویدئو-محور داده‌ها، از راهبردهای افزایش داده موجود در چارچوب YOLOv11 شامل Mosaic، MixUp، تغییرات مقیاس، چرخش و تغییرات روشنایی استفاده شد. همچنین مجموعه اعتبارسنجی به‌صورت مستقل از مجموعه آموزش در نظر گرفته شد تا روند یادگیری مدل به‌طور مستمر پایش شود. این اقدامات موجب کاهش نشت اطلاعات زمانی و افزایش اعتبار نتایج ارزیابی شده است. با وجود این تمهیدات، ارزیابی مدل تنها بر روی یک مجموعه‌داده عمومی انجام شده است. بنابراین، بررسی عملکرد روش پیشنهادی بر روی مجموعه‌داده‌های متنوع‌تر و ارزیابی بین‌داده‌ای (Cross-Dataset Evaluation) می‌تواند در آینده تصویر دقیق‌تری از قابلیت تعمیم مدل ارائه دهد.

۴-۲- معیارهای ارزیابی

برای ارزیابی عملکرد مدل، از معیارهای دقت (Precision) و میانگین دقت متوسط (mAP) استفاده شده است. با در نظر گرفتن TP تشخیص درست، FP تشخیص نادرست مثبت و FN عدم تشخیص نمونه مثبت، دقت به صورت زیر تعریف می‌شود:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (۱۴)$$

برای محاسبه Average Precision (AP) هر کلاس، ابتدا منحنی Precision-Recall تشکیل شده و مقدار AP به صورت انتگرال سطح زیر منحنی تعریف می‌گردد:

$$AP = \int_0^1 P(R)\, dR \quad (۱۵)$$

با توجه به وجود N کلاس، معیار mAP به صورت میانگین AP در تمامی کلاس‌ها محاسبه می‌شود:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (۱۶)$$

در این پژوهش، تعداد کلاس‌ها برابر با $N = 2$ بوده و شامل خشونت‌آمیز و غیرخشونت‌آمیز است. علاوه بر این، دو معیار زیر گزارش شده‌اند:

- **mAP@0.5**: میانگین AP در آستانه همپوشانی $\text{IoU} = 0.5$، که بیانگر عملکرد مدل در تشخیص با همپوشانی نسبتاً زیاد است.
- **mAP@0.5:0.95**: میانگین AP در بازه آستانه‌های IoU از ۰/۵ تا ۰/۹۵ با گام ۰/۰۵، که ارزیابی دقیق‌تر و سخت‌گیرانه‌تری از کیفیت مکان‌یابی و تشخیص ارائه می‌دهد:

$$mAP@[0.5{:}\,0.95] = \frac{1}{10}\sum_{t=0.5}^{0.95} mAP@t \quad (۱۷)$$

۴-۳- آزمایش‌های مقایسه‌ای

به‌منظور اعتبارسنجی اثربخشی روش پیشنهادی، عملکرد آن با چند مدل شناخته‌شده تشخیص شیء بر روی مجموعه‌داده Hockey مورد استفاده در این پژوهش مقایسه شد.

شکل ۸ .مقایسه عملکرد تشخیص در مجموعه داده هاکی

برای رعایت انصاف در مقایسه، آموزش و ارزیابی مدل‌ها تا حد امکان تحت تنظیمات یکسان (شامل داده آموزش/آزمون، اندازه ورودی و معیارهای ارزیابی) انجام شده است.

نتایج جدول ۱ نشان می‌دهد روش پیشنهادی به مقدار mAP@0.5 برابر ۹۲/۸ % و mAP@0.5:0.95برابر ۷۶/۱ % دست یافته است.

این اعداد بیانگر بهبود عملکرد مدل در تشخیص رخدادهای خشونت‌آمیز نسبت به اغلب روش‌های مقایسه‌شده است. همچنین مشاهده می‌شود برخی مدل‌های سنگین‌تر مانند RT-DETR-L و YOLOv9-C با وجود توانایی بالا، در این مسئله به عملکردی پایین‌تر یا نزدیک‌تر به روش پیشنهادی دست یافته‌اند.

۱-۳-۴ ارزیابی قابلیت تعمیم بر روی مجموعه‌داده RWF-2000

به‌منظور بررسی قابلیت تعمیم مدل پیشنهادی و ارزیابی عملکرد آن در شرایط متفاوت، آزمایش‌های تکمیلی بر روی مجموعه‌داده RWF-2000 نیز انجام شد. این مجموعه‌داده شامل ۲۰۰۰ کلیپ ویدئویی واقعی ثبت‌شده توسط دوربین‌های نظارتی بوده و چالش‌هایی نظیر کیفیت پایین تصویر، نور نامناسب، تراکم افراد، هم‌پوشانی و زوایای متنوع تصویربرداری را در بر می‌گیرد.

نتایج ارزیابی در جدول ۲ ارائه شده است. همان‌گونه که مشاهده می‌شود، مدل پیشنهادی بر روی مجموعه‌داده Hockey Fight به مقدار mAP@0.5 برابر با ۹۲/۸% و mAP@0.5:0.95 برابر با ۷۶/۱% دست یافته است. همچنین بر روی مجموعه‌داده RWF-2000نیز مقادیر ۹۲/۴% و ۷۷/۸% به‌دست آمده است.

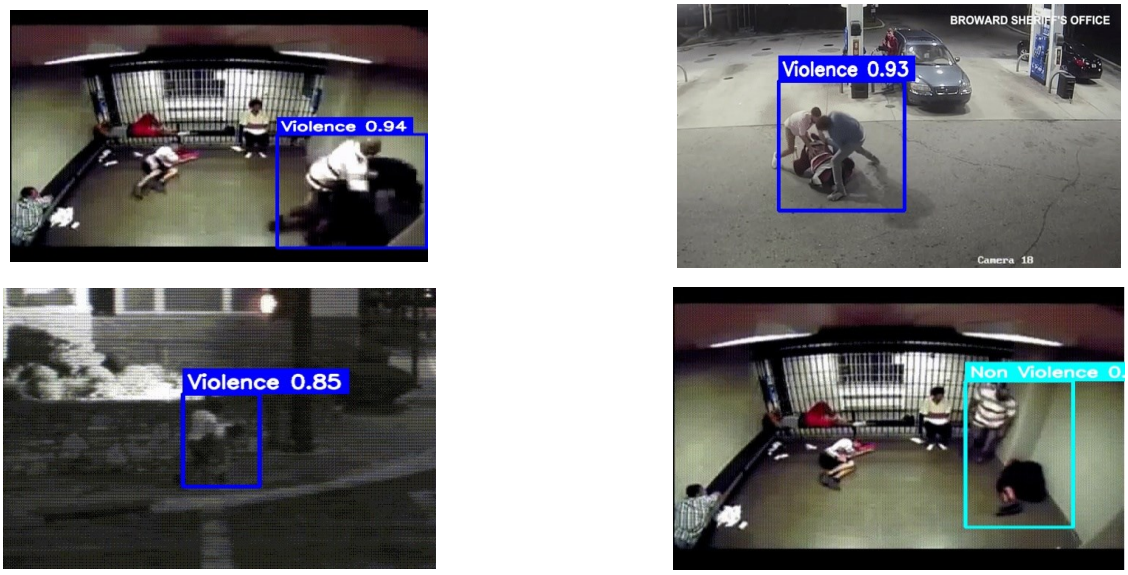
جدول ۲. ارزیابی قابلیت تعمیم مدل پیشنهادی بر روی مجموعه‌داده‌های مختلف

مجموعه‌داده	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Hockey Fight	۹۲/۸	۷۶/۱
RWF-2000	۹۲/۴	۷۷/۸

اگرچه عملکرد مدل بر روی RWF-2000 اندکی کمتر از مجموعه‌داده Hockey Fight است، اما اختلاف محدود نتایج نشان می‌دهد که معماری پیشنهادی توانایی مناسبی در استخراج ویژگی‌های مرتبط با رفتارهای خشونت‌آمیز در سناریوهای نظارتی متفاوت دارد.

جدول ۳. مطالعه حذف مؤلفه‌ها برای بررسی اثر ماژول‌های پیشنهادی

مدل	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FLOPs/G	FPS
YOLOv11 Baseline	۹۰/۲	۷۲/۵	۸۶/۴	۵۲/۳
YOLOv11 + STN	۹۱/۱	۷۳/۴	۸۸/۰	۵۰/۱
YOLOv11 + STN + SF	۹۱/۹	۴۷/۷	۸۹/۲	۴۸/۳
YOLOv11 + STN + SF + DRB	۹۲/۸	۷۶/۱	۹۱/۱	۴۵/۹



شکل ۹.نمونه خروجی مدل پیشنهادی روی مجموعه داده RWF-2000

این موضوع بیانگر پایداری روش پیشنهادی و کاهش احتمال بیش‌برازش به یک مجموعه‌داده خاص است.

نمونه‌هایی از خروجی مدل پیشنهادی بر روی مجموعه‌داده RWF-2000در شکل ۹ نمایش داده شده‌اند. همان‌طور که مشاهده می‌شود، مدل قادر است رخدادهای خشونت‌آمیز را در شرایط دشوار شامل کیفیت پایین تصویر، محیط‌های کم‌نور، تراکم افراد و صحنه‌های واقعی نظارتی با اطمینان بالا شناسایی کند.

۲-۳-۴ مطالعه حذف مؤلفه‌ها

به‌منظور بررسی سهم هر یک از ماژول‌های پیشنهادی در بهبود عملکرد مدل، مطالعه حذف مؤلفه‌ها انجام شد. در این آزمایش، مدل پایه YOLOv11 به‌عنوان خط مبنا در نظر گرفته شد و سپس ماژول‌های شبکه تبدیل مکانی(STN) ، ماژول ویژگی‌های پراکنده (SF) و بلوک بازسازی باقیمانده متسع (DRB) به‌صورت مرحله‌ای به معماری اضافه شدند. نتایج این ارزیابی در جدول ۳ ارائه شده است.

همان‌طور که در جدول ۳ مشاهده می‌شود، افزودن ماژول STNموجب افزایش mAP@0.5 از ۹۰/۲% به ۹۱/۱% و mAP@0.5:0.95از ۷۲/۵% به ۷۳/۴% شده است. این بهبود نشان می‌دهد که هم‌ترازسازی مکانی نقشه‌های ویژگی اولیه می‌تواند مقاومت مدل را در برابر تغییرات زاویه دید، جابه‌جایی و چرخش افزایش دهد.

با اضافه شدن ماژولSF ، مقدار mAP@0.5 به ۹۱/۹% و mAP@0.5:0.95به ۷۴/۷% افزایش یافته است. این نتیجه بیانگر آن است که انتخاب و تقویت کانال‌های مؤثر موجب کاهش افزونگی ویژگی‌ها و تمرکز بهتر مدل بر نواحی مهم تصویر می‌شود.

در نهایت، افزودن ماژول DRB باعث دستیابی مدل به mAP@0.5برابر با ۹۲/۸% و mAP@0.5:0.95 برابر با ۷۶/۱% شده است. این افزایش نشان می‌دهد که تقویت میدان دید مؤثر و بازسازی ویژگی‌ها در صحنه‌های پرتراکم و دارای هم‌پوشانی نقش مهمی در بهبود کیفیت تشخیص دارد.

از نظر بار محاسباتی، افزایش تدریجی FLOPs و کاهش محدود FPS مشاهده می‌شود. با این حال، مدل نهایی همچنان سرعت ۴۵/۹ فریم بر ثانیه را حفظ کرده است که برای کاربردهای نظارتی بلادرنگ قابل قبول است. بنابراین، نتایج مطالعه حذف

جدول ۴. مقایسه روش پیشنهادی با روش‌های موجود تشخیص رفتارهای خشونت‌آمیز

روش	دقت Hockey Fight (%)	دقت RWF-2000 (%)	تعداد پارامترها	سرعت / FLOPs
Violence 4D [۲۸]	۱۰۰/۰۰	۹۴/۶۷	–	–
Lightweight Bi-LTMA + TSM [۲۹]	۹۸/۵۰	۹۰/۲۵	M۴/۴۸	GFLOPs۱/۲۱
Efficient Human Violence Recognition [۳۰]	۹۸/۲۰	۸۸/۵۰	M۳/۵۱	GFLOPs۳/۱۵
LAVID [۳۱]	۹۵/۳۳	۹۱/۰۶	M۰/۵۷	ms۳۲
Pose + Motion Fusion [۳۲]	۹۸/۵۰	۹۵/۰۵	۱۱۸/۷۵۵	s۰/۳۳۳
روش پیشنهادی	۹۸/۶۰	۹۵/۱۰	M۱۲/۵	۴۵/۹ GFLOPs / ۹۱/۱ FPS

مؤلفه‌ها نشان می‌دهد که هر یک از ماژول‌هایSTN ، SFو DRBسهم مشخصی در بهبود عملکرد داشته و ترکیب آن‌ها بهترین تعادل میان دقت و سرعت را فراهم کرده است.

۴-۴ مصورسازی و تحلیل

برای بررسی دقیق‌تر عملکرد طبقه‌بندی، ماتریس سردرگمی برای مدل پایه YOLOv11s و روش پیشنهادی مطابق شکل ۶ ارائه شده است. در این ماتریس، محور افقی نشان‌دهنده برچسب واقعی و محور عمودی نشان‌دهنده برچسب پیش‌بینی‌شده است. مقادیر قطری ماتریس بیانگر نسبت پیش‌بینی‌های صحیح هر کلاس بوده و مقادیر خارج از قطر به خطاهای طبقه‌بندی مربوط می‌شوند.
مطابق شکل ۶، روش پیشنهادی نسبت به مدل پایه، توانایی بالاتری در تمایز بین کلاس‌های «خشونت‌آمیز» و «غیرخشونت‌آمیز» نشان می‌دهد. به‌طور مشخص، نرخ پیش‌بینی صحیح کلاس «خشونت‌آمیز» از ۰/۴۰ در مدل پایه به ۰/۹۲ در روش پیشنهادی افزایش یافته است. همچنین، نرخ پیش‌بینی صحیح کلاس «غیرخشونت‌آمیز» از ۰/۸۲ به ۰/۸۶ بهبود پیدا کرده است. این نتایج نشان می‌دهد افزودن ماژول‌های پیشنهادی موجب کاهش خطای طبقه‌بندی و افزایش پایداری مدل در تشخیص رخدادهای خشونت‌آمیز شده است.

۴-۴-۱ منحنی‌هایPrecision–Recall

به‌منظور تحلیل رفتار مدل در سطوح مختلف بازیابی(Recall) ، منحنی‌های Precision–Recall برای مدل پایه و روش پیشنهادی مطابق شکل ۷ ترسیم شده‌اند. مقایسه منحنی‌ها نشان می‌دهد روش پیشنهادی در اغلب نقاط، مقدار Precision بالاتری را در سطح Recall یکسان ارائه می‌دهد که بیانگر کاهش تشخیص‌های مثبت کاذب و افزایش دقت پیش‌بینی است.
به‌طور ویژه، برای کلاس «غیرخشونت‌آمیز»، مقدار Precision از حدود ۰/۳۸ در مدل پایه به حدود ۰/۹۲ در روش پیشنهادی افزایش یافته است. همچنین، مقدار mAP@0.5به‌عنوان شاخص کلی عملکرد تشخیص، از مقدار پایین‌تر در مدل پایه به حدود ۰/۹۲۳ در روش پیشنهادی ارتقا یافته است که بهبود قابل توجه کیفیت تشخیص را تأیید می‌کند.

۴-۴-۲ تحلیل کیفی تشخیص‌ها

برای تحلیل کیفی عملکرد مدل، چند فریم شامل صحنه‌های دارای درگیری انتخاب شد و خروجی مدل پایه با روش پیشنهادی مقایسه گردید (شکل ۸). نتایج نشان می‌دهد روش پیشنهادی در تشخیص رخدادهای خشونت‌آمیز، به‌ویژه در شرایط دشوار مانند کیفیت پایین تصویر، تراکم بالای افراد، هم‌پوشانی و تغییرات سریع حرکت، خروجی پایدارتر و دقیق‌تری ارائه می‌دهد. این بهبود را می‌توان به تقویت استخراج ویژگی و افزایش تمرکز مدل بر نواحی مهم توسط ماژول‌هایSTN ، SFو DRB نسبت داد.

۴-۴-۳ تحلیل خطاهای تشخیص

به‌منظور بررسی دقیق‌تر محدودیت‌های مدل پیشنهادی، نمونه‌های خطادار نیز مورد تحلیل قرار گرفتند. نتایج نشان داد که خطاهای مثبت کاذب (False Positive) عمدتاً در شرایطی رخ می‌دهند که تعامل نزدیک افراد شباهت ظاهری زیادی به درگیری فیزیکی داشته باشد. برای مثال، در آغوش گرفتن، فعالیت‌های ورزشی، تجمع‌های متراکم و حرکات سریع گروهی ممکن است به‌اشتباه

به‌عنوان رفتار خشونت‌آمیز شناسایی شوند .همچنین خطاهای منفی کاذب (False Negative) بیشتر در شرایط دارای کیفیت پایین تصویر، فاصله زیاد افراد از دوربین، انسداد شدید افراد و تاری ناشی از حرکت مشاهده شدند. در این شرایط، بخشی از نشانه‌های بصری مرتبط با خشونت از دست رفته و فرآیند تشخیص دشوارتر می‌شود .بررسی این موارد نشان می‌دهد که اگرچه معماری پیشنهادی در استخراج ویژگی‌های فضایی عملکرد مطلوبی دارد، اما بهره‌گیری از اطلاعات زمانی و استفاده از مجموعه‌داده‌های متنوع‌تر می‌تواند در آینده موجب کاهش نرخ خطا و افزایش قابلیت تعمیم مدل شود.

٤-٥ مقایسه با روش‌های موجود تشخیص رفتارهای خشونت‌آمیز

به‌منظور تعیین جایگاه روش پیشنهادی در میان روش‌های موجود، نتایج آن با روش‌های منتخب ارزیابی‌شده بر روی مجموعه‌داده‌های Hockey Fight و RWF-2000 از نظر دقت، تعداد پارامترها و هزینه محاسباتی یا سرعت استنتاج مقایسه شده است. نتایج این مقایسه در جدول ٤ ارائه شده است.

جدول ٤. مقایسه روش پیشنهادی با روش‌های موجود تشخیص رفتارهای خشونت‌آمیز

نتایج جدول ٤ نشان می‌دهد که روش پیشنهادی از نظر دقت، عملکرد رقابتی و بالایی در هر دو مجموعه‌داده دارد. در مجموعه‌داده Hockey Fight، روش پیشنهادی با دستیابی به Accuracy %٩٨/٦٠ عملکردی نزدیک به روش‌های Lightweight Bi-LTMA + TSM و Pose + Motion Fusion با ٩٨/٥٠% داشته و تنها از روش Violence 4D با ١٠٠% پایین‌تر است. در مجموعه‌داده RWF-2000 نیز روش پیشنهادی با ٩٥/١٠% Accuracy بالاترین مقدار دقت را در میان روش‌های مورد مقایسه به دست آورده است، هرچند اختلاف آن با روش Pose + Motion Fusion با ٩٥/٠٥% بسیار اندک است. از نظر پیچیدگی، روش پیشنهادی با ١٢/٥ میلیون پارامتر و ٩١/١ GFLOPs سبک‌ترین روش مورد مقایسه نیست؛ بااین‌حال، با دستیابی به ٤٥/٩ FPS قابلیت پردازش بلادرنگ را حفظ می‌کند. علاوه بر این، برخلاف بسیاری از روش‌های موجود که مسئله را در سطح ویدئو به‌صورت طبقه‌بندی حل می‌کنند، روش پیشنهادی در چارچوب تشخیص شیء توسعه یافته و علاوه بر تشخیص رفتار خشونت‌آمیز، امکان مکان‌یابی رخدادهای خشونت‌آمیز را نیز فراهم می‌کند. بنابراین، مزیت اصلی روش پیشنهادی در دستیابی هم‌زمان به دقت تشخیص بالا، قابلیت مکان‌یابی و پردازش بلادرنگ است، نه در کمینه‌سازی تعداد پارامترها یا هزینه محاسباتی. همچنین، از آنجا که روش‌های موجود عمدتاً Accuracy و روش پیشنهادی علاوه بر Accuracy از معیارهای mAP@0.5 و mAP@0.5:0.95 استفاده می‌کند، این معیارها به‌عنوان معیارهای کاملاً معادل در نظر گرفته نشده‌اند.

۶-۴ تحلیل پیچیدگی محاسباتی

یکی از موضوعات مهم در به‌کارگیری مدل‌های تشخیص خشونت در سامانه‌های نظارتی، حفظ تعادل میان دقت تشخیص و سرعت پردازش است. از آنجا که در معماری پیشنهادی سه ماژول شبکه تبدیل مکانی(Spatial Transformer Network - STN) ، ماژول ویژگی‌های پراکنده (Sparse Feature - SF) و بلوک بازسازی باقیمانده متسع (Dilated Residual Block - DRB) به ساختار پایه YOLOv11 افزوده شده‌اند، بررسی تأثیر این ماژول‌ها بر بار محاسباتی مدل ضروری است.

نتایج جدول ١ نشان می‌دهد که روش پیشنهادی با وجود اضافه شدن ماژول‌های یادشده، دارای مقدار ٩١/١ GFLOPs و سرعت پردازش ۴۵/۹ فریم بر ثانیه است. این مقدار نشان می‌دهد که مدل پیشنهادی همچنان قابلیت استفاده در بسیاری از کاربردهای نظارتی بلادرنگ را حفظ می‌کند. در مقایسه با برخی مدل‌های سنگین‌تر مانند RT-DETR-L وYOLOv9-C ، روش پیشنهادی ضمن دستیابی به دقت بالاتر، پیچیدگی محاسباتی کنترل‌شده‌تری دارد.

اگرچه افزودنSTN ، SFو DRB موجب افزایش نسبی هزینه محاسباتی نسبت به مدل پایه می‌شود، اما این افزایش در برابر بهبود دقت تشخیص، افزایش پایداری در صحنه‌های پرتراکم و کاهش خطا در شرایط دارای هم‌پوشانی قابل توجیه است. همچنین، در طراحی معماری پیشنهادی تلاش شده است تا سربار محاسباتی کنترل شود؛ به‌گونه‌ای که STN بر روی نقشه‌های ویژگی اولیه و نه تصویر خام اعمال شده و DRB نیز با بهره‌گیری از راهبرد بازپارامتردهی، در مرحله استنتاج به ساختاری کاراتر تبدیل می‌شود. بنابراین، روش پیشنهادی تعادل مناسبی میان دقت، پایداری و سرعت پردازش برقرار می‌کند.

۵- نتیجه‌گیری

تشخیص رخدادهای خشونت‌آمیز در ویدئوهای نظارتی، به‌ویژه در صحنه‌های پرتراکم و دارای هم‌پوشانی، یکی از چالش‌های مهم در کاربردهای امنیتی محسوب می‌شود. در این پژوهش، به‌منظور بهبود دقت و پایداری تشخیص، یک نسخه تقویت‌شده از YOLOv11ارائه شد که شامل سه ماژول کلیدی شبکه تبدیل مکانی(STN) ، ماژول ویژگی‌های پراکنده (SF) و بلوک بازسازی باقیمانده متسع (DRB) است.

ماژول STN با هم‌ترازسازی مکانی نقشه‌های ویژگی اولیه، موجب افزایش مقاومت مدل در برابر تغییرات مکانی و اختلافات

زاویه دید شد. ماژول SF با تقویت کانال‌های مؤثر و کاهش افزونگی ویژگی‌ها، تمرکز شبکه را بر نواحی مهم در صحنه‌های شلوغ افزایش داد. همچنین ماژول DRB با تقویت زمینه ادراکی و بهبود نمایش ویژگی‌ها در شرایط پیچیده، به استخراج بهتر الگوهای رخداد کمک کرد.

با وجود نتایج امیدوارکننده، پژوهش حاضر دارای محدودیت‌هایی نیز است. نخست آنکه تشخیص رخدادهای خشونت‌آمیز بر اساس تحلیل فریم‌های منفرد انجام شده و وابستگی‌های زمانی میان فریم‌های متوالی به‌صورت صریح مدل‌سازی نشده است. در نتیجه، بخشی از اطلاعات رفتاری موجود در توالی زمانی ویدئوها مورد استفاده قرار نگرفته است. همچنین ارزیابی مدل تنها بر روی یک مجموعه‌داده عمومی انجام شده است. بنابراین، استفاده از معماری‌های فضایی-زمانی و ارزیابی بر روی مجموعه‌داده‌های متنوع‌تر می‌تواند به‌عنوان یکی از مسیرهای اصلی توسعه پژوهش در آینده دنبال شود.

نتایج آزمایش‌ها بر روی مجموعه‌داده عمومی Hockey نشان داد روش پیشنهادی به mAP@0.5 برابر با ۹۲/۸ % و mAP@0.5:0.95 برابر با ۷۶/۱ % دست یافته است که بهبود معناداری نسبت به مدل پایه ارائه می‌دهد. این نتایج نشان می‌دهد ادغام هدفمند ماژول‌های پیشنهادی در چارچوب YOLOv11 می‌تواند عملکرد تشخیص رخدادهای خشونت‌آمیز را در صحنه‌های پرتراکم ارتقاء دهد.

با توجه به محرمانه بودن داده‌های واقعی زندان، ارزیابی این پژوهش بر پایه مجموعه‌داده‌های عمومی انجام شده است تا امکان مقایسه منصفانه فراهم گردد.

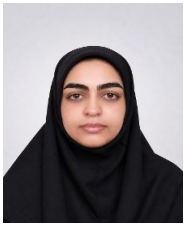
قدردانی

این پژوهش با حمایت مالی اداره کل زندان‌های استان کرمان تحت قرارداد شماره ۱۷۹۳۲/۱۰/۴۳۷/۰۲ انجام شده است.

منابع:

[1].O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A review of video surveillance systems, Journal of Visual Communication and Image Representation, 2021.
[2]. S. K. Thampan, J. Joseph, Intelligent Surveillance System for Crime Prevention Using Deep Learning, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2023.
[3]. M. Salvi, U. R. Acharya, F. Molinari, K. M. Meiburger, The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis, Computers in Biology and Medicine, 2021.
[4]. P. Wang, P. Wang, E. Fan, Violence detection and face recognition based on deep learning, Pattern Recognition Letters, 2021.
[5]. X. Zhou, Y. Chen, Q. Zhang, Trajectory Analysis Method Based on Video Surveillance Anomaly Detection, China Automation Congress (CAC), 2021.
[6]. M. Guo, X. Zeng, D. Chen, N. Yang, Deep-learning-based earth fault detection using continuous wavelet transform and convolutional neural network in resonant grounding distribution systems, IEEE Sensors Journal, 2017.
[7]. F. Barros, S. Aguiar, P. J. Sousa, A. Cachaço, P. J. Tavares, P. M. G. P. Moreira, D. Ranzal, N. Cardoso, N. Fernandes, et al., Displacement monitoring of a pedestrian bridge using 3D digital image correlation, Procedia Structural Integrity, 2022.
[8]. M. Khan, W. Gueaieb, A. El Saddik, G. De Masi, F. Karray, An efficient violence detection approach for smart cities surveillance system, IEEE International Smart Cities Conference (ISC2), 2023.
[9]. M. Ramzan, A. Abid, H. Ullah Khan, S. Mahmood Awan, A. Ismail, M. Ahmed, A. A review on state-of-the-art violence detection techniques, IEEE Access, 2019.
[10]. G. Liu, Z. Wang, H. Zhang, X. Guo, Y. Wang, C. Zhang, A novel violent video detection method based on improved C3D and transfer learning, International Conference on Computer Information and Big Data Applications (CIBDA), 2022.
[11]. E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, A. Ahilan, Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM, Computer Networks, 2019.
[12]. D. Jackson Samuel R., F. E., G. Manogaran, V. G. N., T. T., S. Jeeva, A. Ahilan, Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets, Neurocomputing, 2019.
[13]. S. Accattoli, P. Sernani, N. Falcionelli, D. Neway Mekuria, Violence detection in videos by combining 3D convolutional neural networks and support vector machines, Applied Artificial Intelligence, 2020.
[14]. M. Magdy, M. W. Fakhr, F. A. Maghraby, Violence 4D: Violence detection in surveillance using 4D convolutional neural networks, IET Computer Vision, 2023.
[15]. N. Waddenkery, S. Soma, An efficient convolutional neural network for detecting the crime of stealing in videos, Entertainment Computing, 2024.
[16]. K. Rahima, H. Muhammad, YOLOv11: An overview of the key architectural enhancements, arXiv, 2024.
[17]. G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, version 8.0.0, GitHub, 2024.
[18]. Z. Wang, S. Ji, Smoothed dilated convolutions for improved dense prediction, ACM SIGKDD

International Conference on Knowledge Discovery & Data Mining, 2019.
[19]. W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, Advances in Neural Information Processing Systems, 2017.
[20]. G. Jocher, et al., YOLOv5, Ultralytics, 2020.
[21]. G. Jocher, et al., YOLOv8: Ultralytics Open-Source YOLO Models, 2023.
[22]. J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement, arXiv preprint arXiv:1804.02767, 2018.
[23]. Y. Zhang, H. Peng, D. Zhang, J. Wang, T. Tan, RT-DETR: Real-Time Detection Transformer, arXiv preprint arXiv:2304.08069, 2023.
[24]. S. Liu, L. Jin, W. Wang, GELAN: A Generalized Efficient Layer Aggregation Network for Object Detection, arXiv preprint arXiv:2303.09355, 2023.
[25]. G. Jocher, A. Chaurasia, J. Qiu, YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information, arXiv preprint arXiv:2402.13616, 2024.
[26]. W. Xu, D. Zhu, R. Deng, K. L. Yung, A. W. H. Ip, Violence-YOLO: Enhanced GELAN Algorithm for Violence Detection, Applied Sciences, 2024.
[27]. M. Cheng, K. Cai, M. Li, RWF-2000: An open large scale video database for violence detection, Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021.
[28]. M. Magdy, M. W. Fakhr, F. A. Maghraby, Violence 4D: Violence detection in surveillance using 4D convolutional neural networks, IET Computer Vision, 2023.
[29]. Z. Wang, S. Ji, Lightweight Violence Detection Model Based on 2D CNN with Bi-Directional Motion Attention, Applied Sciences, 2024.
[30]. Efficient Human Violence Recognition for Surveillance in Real Time, Sensors, 2024.
[31]. LAVID: A Lightweight and Autonomous Smart Camera System for Urban Violence Detection and Geolocation, Computers, 2025.
[32]. S. T. Y. Aung, W. Kusakunniran, K. H. Yew, Leveraging Fusion Methods of Human Pose and Motion Dynamics for Accurate Violence Detection in Video Surveillance, IEEE Access, 2025.



افسانه شمس‌الدینی فرسنگی، دانش‌آموخته مهندسی کامپیوتر با گرایش هوش مصنوعی از دانشگاه شهید باهنر کرمان است. حوزه‌های پژوهشی و علاقه‌مندی ایشان شامل پردازش تصویر، هوش مصنوعی و بینایی ماشین می‌باشد. همچنین تمرکز ایشان بر به‌کارگیری روش‌های یادگیری ماشین در تحلیل و تفسیر داده‌های تصویری و توسعه سامانه‌های هوشمند است.

مصطفی قاضی‌زاده-احسائی، مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه فردوسی مشهد دریافت کرد. ایشان مدرک کارشناسی ارشد را در سال ۱۳۷۸ در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه فردوسی مشهد اخذ نمود و دکترای خود را نیز در سال ۱۳۹۵ در رشته نرم‌افزار از دانشگاه فردوسی مشهد دریافت کرد. وی در حال حاضر استادیار دانشگاه شهید باهنر کرمان (دانشکده مهندسی کامپیوتر) است و حوزه‌های تخصصی ایشان شامل یادگیری ماشین، داده‌کاوی، یادگیری عمیق و پردازش تصویر می‌باشد.