دانشکده مهندسی کامپیوتر

توسعه ابزار برچسب‌گذاری هوشمند تصویر در کاربرد بازشناسی انسان

پروژه کارشناسی مهندسی کامپیوتر گرایش هوش مصنوعی

الناز رضائی خانیکی

استاد راهنما

دکتر محمدرضا محمدی

بهمن ۱۴۰۲


بسم الله الرحمن الرحيم

تأییدیه هیأت داوران جلسه دفاع از پروژه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: الناز رضائی خانیکی

عنوان پروژه : توسعه ابزار برچسب‌گذاری هوشمند تصویر در کاربرد بازشناسی انسان

تاریخ دفاع: بهمن ۱۴۰۲

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضـــــــا
۱	استاد راهنما	دکتر محمدرضا محمدی	استادیار	دانشگاه علم و صنعت ایران	

تأییدیه صحت و اصالت نتایج

باسمه تعالی

اینجانب الناز رضائی خانیکی به شماره دانشجویی ۹۸۴۱۱۳۸۷ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه نتایج این پروژه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: الناز رضائی خانیکی

تاریخ و امضا:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.

☐ بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه تا تاریخ ..... ممنوع است.

استاد راهنما: دکتر محمدرضا محمدی

تاریخ:

امضا:

قدردانی

خود را موظف می‌دانم از تمامی افرادی که در طول این مسیر پشتیبان و راهنمای من بودند، صمیمانه تشکر کنم. ابتدا، از خانواده‌ام، به‌ویژه پدر و مادر عزیزم که همواره در کنارم بودند و حمایت‌شان را از من دریغ نکردند بسیار سپاسگزارم.

از اساتید دانشکده مهندسی کامپیوتر، به ویژه جناب آقای دکتر محمدی، که نه تنها در تکمیل این پروژه به اینجانب کمک فراوانی کردند، بلکه در طول دوران تحصیل همیشه به عنوان یک راهنما و استاد برجسته برایم حضور داشتند، صمیمانه قدردانی می‌نمایم. تجارب و آموخته‌های بی‌نظیری که در کنار ایشان بهره‌مند شدم، گره‌گشای بزرگی برایم بوده و مسیر تحصیلی‌ام را روشن‌تر ساخته است. برای ایشان و خانواده گرامی‌شان، آرزوی سلامتی، خوشبختی و موفقیت دارم.

همچنین از جناب آقای دکتر مزینی که زحمت داوری این پروژه را بر عهده گرفته و وقت ارزشمند خود را در اختیار اینجانب قرار دادند، تشکر و قدردانی می‌نمایم.

از خانم انوری نیز بخاطر راهنمایی‌ها و پشتیبانی‌هایش در طول انجام این پروژه، صمیمانه سپاسگزارم.

الناز رضائی خانیکی

بهمن ۱۴۰۲

چکیده

شناسایی مجدد انسان یکی از مسائل حوزه بینایی کامپیوتر است که در آن هدف پیدا کردن یک فرد خاص در تصاویر گرفته شده از چندین زاویه دوربین است. با پیشرفت روزافزون روش‌های یادگیری عمیق، به ویژه شبکه‌های عصبی همگشتی، این حوزه به یکی از بخش‌های مهم و رو به رشد در زمینه‌های نظارت و امنیت تبدیل شده است. در حال حاضر، یکی از چالش‌های مهم در شناسایی مجدد انسان، موجودیت‌یابی و برچسب‌گذاری داده‌ها است. تولید داده‌های برچسب‌گذاری شده دقیق و کافی برای آموزش مدل‌های یادگیری عمیق از جمله این چالش‌ها است. در این زمینه، توسعه ابزارهای برچسب‌گذاری هوشمند می‌تواند بهبودهای قابل توجهی را در فرآیند برچسب‌گذاری و در نتیجه در عملکرد مدل‌های شناسایی مجدد انسان ایجاد کند. اکثر روش‌های فعلی، به خصوص مدل‌های یادگیری عمیق، بر پایه یادگیری نظارت شده هستند که نیاز به تعداد زیادی تصویر دسته‌بندی شده از هر فرد در دوربین‌های مختلف دارند. از این رو، استفاده از یادگیری بدون نظارت برای شناسایی مجدد فرد به تازگی مورد توجه قرار گرفته است. تاکنون روش‌هایی که برای وظیفه شناسایی مجدد انسان ارائه شده‌اند، بر پایه استفاده از ویژگی‌های ظاهری و معنایی بوده است. بنابراین، در مجموعه داده‌هایی که شامل داده سخت هستند، مانند مجموعه داده‌ای که در این پژوهش استفاده شده است، نتیجه خوبی ندارند. هدف اصلی این پژوهش، ارائه یک مدل زمانی است که بتوان با استفاده از اختلاف زمانی ظاهر شدن فرد در دو دوربین، او را در دوربین متناظر تشخیص داد. این روش امکان افزایش دقت شناسایی و در نظر گرفتن ویژگی‌های اضافی به جز ظاهری و معنایی را فراهم می‌کند و می‌تواند منجر به کاهش خطا و ارتقاء عملکرد کلی سیستم شناسایی مجدد انسان گردد.

واژگان کلیدی: شناسایی مجدد انسان، روش‌های یادگیری عمیق، شبکه‌های عصبی همگشتی، ردیابی فرد، مدل‌های یادگیری بدون نظارت، مدل زمانی

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمه

۱ـ۱ مقدمه

شناسایی مجدد انسان [۱] یکی از مسائل بینایی کامپیوتر [۲] است که هدف آن تطابق [۳] یک فرد هدف [۴] در سراسر چندین زاویه دید [۵] دوربین است. در سال‌های اخیر، این موضوع به دلیل توسعه و افزایش روش‌های یادگیری عمیق [۶] ، به‌ویژه شبکه‌های عصبی همگشتی [۷]،شاهد پیشرفت سریعی بوده است. این حوزه، به ویژه در زمینه نظارت و امنیت، به یکی از زمینه‌های تحقیقاتی مهم و در حال رشد تبدیل شده است. انگیزه اصلی شناسایی مجدد انسان، امکان ردیابی [۸] موثر افراد در محیط‌های پیچیده و شلوغ مانند فرودگاه‌ها، ایستگاه‌های قطار و مکان‌های عمومی است [۹]. به دلیل افزایش تقاضا برای امنیت عمومی و توسعه سریع شبکه‌های دوربین، شناسایی مجدد انسان جلب توجه بیشتری کرده است. این مطالعات به هدف صرفه‌جویی در منابع انسانی و یافتن کارآمد فرد مورد نظر، مانند کودک گم‌شده در فرودگاه، از بین هزاران تصویر داوطلب [۹]، انجام می‌شوند [۱۰]. با این حال، هنوز مشکلات زیادی در استفاده از مدل‌های فعلی شناسایی مجدد انسان در دنیای واقعی

[1] Person Re-identification
[2] Computer Vision (CV)
[3] Match
[4] Target
[5] Perspectives
[6] Deep Learning (DL)
[7] Convolutional Neural Network (CNN)
[8] Tracking
[9] Candidate

وجود دارد.

به طور کلی، شناسایی مجدد انسان بسیار چالش‌برانگیز است زیرا تصاویری که توسط دوربین‌های مختلف گرفته می‌شوند، معمولاً شامل تغییرات قابل توجه داخل کلاسی [۱۰] هستند که ناشی از تغییرات در پس‌زمینه [۱۱]، دامنه تصویر [۱۲]، حالت انسان [۱۳] و غیره می‌باشد. به عبارت دیگر، طراحی یا یادگیری بازنمایی‌هایی [۱۴] که در برابر تغییرات داخل کلاسی تا حداکثر امکان مقاوم باشند، یکی از اهداف اصلی در شناسایی مجدد انسان بوده است [۱۱]. به طور مثال، اگر دوربین ابتدا با یک فرد از یک زاویه خاص روبرو شود، ویژگی‌ها [۱۵] یا بردارهای تعبیه [۱۶] شخص را در پایگاه داده ذخیره می‌کند. سپس، هنگامی که دوربین با همان فرد از زاویه متفاوتی روبرو می‌شود، آن ویژگی‌ها یا بردارهای ابتدایی ممکن است برای شناسایی مجدد شخص کافی نباشد و به همین دلیل احتمال اینکه شخصیت جدید دیده شده به عنوان یک فرد متفاوت شناخته شود، بسیار بالاست.

اکثر روش‌های موجود شناسایی مجدد انسان، به خصوص مدل‌های یادگیری عمیق شبکه‌های عصبی[۱۷]، متکی بر یادگیری نظارتی [۱۸] هستند. این مدل‌های عمیق نظارتی فرض می‌کنند که تعداد زیادی جفت تصاویر دسته‌بندی شده دستی از هر فرد در دوربین‌های مختلف برای هر جفت دوربین در دسترس است. در نتیجه، این امکان را فراهم می‌کند که یک بازنمایی ویژگی و یا یک تابع معیار [۱۹] فاصله بهینه برای هر جفت دوربین به دست آوریم. اما چنین فرضیه‌ای محدودیت‌هایی ذاتی برای عمومی‌سازی [۲۰] مدل به شبکه‌های دوربین مختلف ایجاد می‌کند. این کار به این دلیل است که، برچسب‌گذاری دستی جفت تصاویر مثبت و منفی هر فرد برای هر جفت دوربین به صورت کامل، هزینه‌بر است، زیرا تعداد دوربین‌ها در یک شبکه نظارتی به صورت مربعی افزایش می‌یابد. از این رو، استفاده از یادگیری بدون نظارت [۲۱] برای شناسایی مجدد انسان، به تازگی مورد

[10] Intra-class
[11] Background (BG)
[12] Viewpoint
[13] Human Pose
[14] Representation
[15] Feature
[16] Embedding
[17] Neural Network
[18] Supervised Learning
[19] Metric Function
[20] Generalizing
[21] Unsupervised Learning

تمرکز تحقیقات واقع شده است. در این شرایط، داده‌های آموزشی[۲۲] برچسب‌دار[۲۳] تشخیص هویت برای هر جفت دوربین دیگر لازم نیست.

به هر حال، مدل‌های یادگیری بدون نظارت موجود در شناسایی مجدد انسان، از لحاظ دقت[۲۴] به طرز چشمگیری ضعیف‌تر عمل می‌کنند. این امر به این دلیل است که عدم وجود داده‌های برچسب‌گذاری شده جفتی از طریق دامنه دید دوربین‌های مختلف، از توانایی مدل برای یادگیری اطلاعات متمایز قوی جلوگیری می‌کند. با این وجود، مدل های یادگیری بدون نظارت برای مقابله با تغییرات قابل توجه در ظاهر فرد در طول زمان در دوربین‌ها بسیار حیاتی است [۱۲].

هدف از انجام این پژوهش، ارائه یک مدل زمانی است که با استفاده از اختلاف زمانی ظاهر شدن فرد در دو دوربین، این فرد را با شخص متناظر در دوربین دیگر تطبیق می‌دهد. این کار به این منظور صورت می‌گیرد که مدل نهایی، توانایی قابل توجهی در شناسایی افراد روی مجموعه‌داده شخصی‌سازی شده خودمان نیز داشته باشد. این روش امکان افزایش دقت شناسایی را فراهم می‌کند و ویژگی‌های اضافی به جز ظاهری و معنایی را در نظر می‌گیرد. از طریق این روش، قادر خواهیم بود با استفاده از ویژگی‌های بیشتر و ترکیب روش‌های خلاقانه با روش‌های موجود، دقت شناسایی را بهبود بخشیم و به نتایج دقیق‌تری دست پیدا کنیم. این ترکیب از ویژگی‌ها و روش‌ها، می‌تواند منجر به کاهش خطا و ارتقای عملکرد کلی سیستم شناسایی مجدد انسان شود.

در ادامه ابتدا به مقایسه روش‌های پیشین در حوزه شناسایی مجدد انسان می‌پردازیم. سپس ایده ارائه شده را به همراه نتایج مربوطه بررسی خواهیم کرد. در فصل سوم، به بررسی جزئیات مدل شناسایی مجدد انسان، نظیر ساختار شبکه، تابع ضرر[۲۵]، و معیارهای ارزیابی[۲۶] می‌پردازیم. در فصل چهارم، نتایج الگوریتم پیشنهادی به همراه تحلیل آن‌ها ارائه می‌شود. در انتها، جمع‌بندی و کارهای آینده برای بهبود وظیفه[۲۷] شناسایی مجدد انسان آورده شده است.

[22] Training Data
[23] Labelled
[24] Accuracy
[25] Loss Function
[26] Evaluation Metric
[27] Task

فصل ۲

بررسی مطالعات قبلی و کارهای مرتبط

تجزیه و تحلیل تصاویر متمرکز بر انسان [۱] نقش مهمی در برنامه‌های گسترده‌ای مانند نظارت، ورزش، واقعیت افزوده [۲] و تولید ویدیو دارد. شناسایی مجدد انسان، تشخیص ویژگی‌های عابر پیاده [۳]، جستجوی فرد [۴]، تشخیص عابر پیاده [۵]، تجزیه و تحلیل انسان [۶] و تخمین حالت انسان [۷] در سال‌های اخیر پیشرفت‌های قابل توجهی کسب کرده‌اند. از دید دیگر، تصاویر انسانی بسیاری در جامعه بینایی کامپیوتری فعلی موجود است. به عنوان مثال، حتی یک مجموعه‌داده برای شناسایی مجدد انسان بدون برچسب مانند LUPerson (#Img≈4.18M) ۴ برابر اندازه مجموعه‌داده ImageNet (#Img≈1M) است [۱۳]. چگونگی استفاده از داده‌های بدون برچسب برای ساخت بازنمایی انسانی چالش برانگیز است، به ویژه زمانی که نیاز به استفاده از آن در وظایف مختلفی وجود دارد.

یادگیری خودنظارتی [۸] با استفاده از داده‌های بدون برچسب برای یادگیری بازنمایی‌ها پیشرفت‌های بزرگی کسب کرده است. بسیاری از وظایف پیش‌بینی مطرح شده [۹] از جمله یادگیری متقابل [۱۰] و مدل‌سازی تصویر با

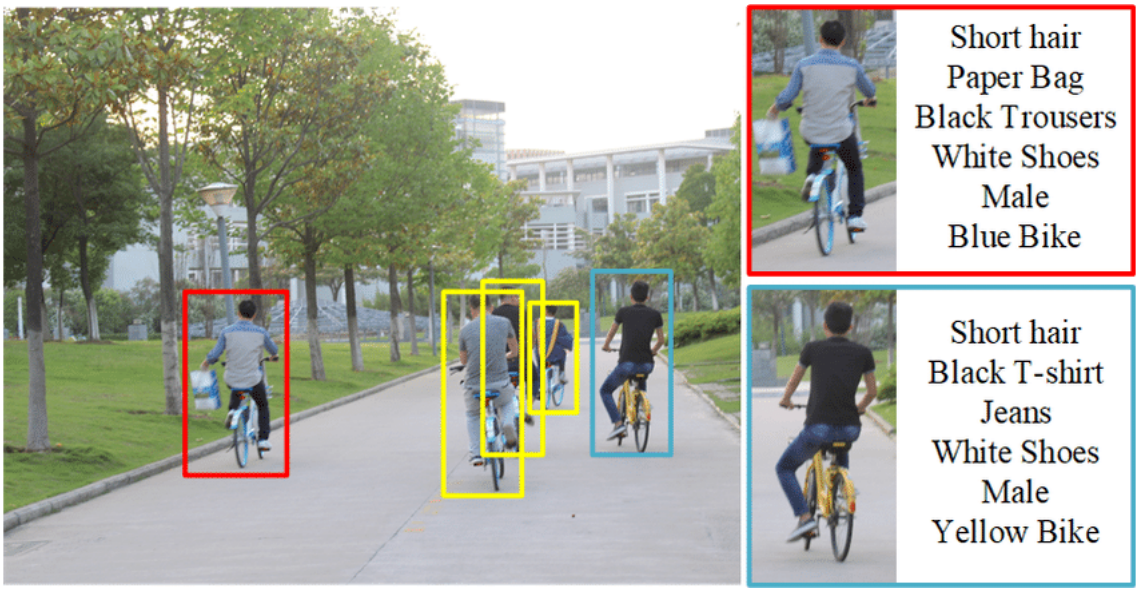
[1] Human-centric
[2] Augmented Reality (AR)
[3] Pedestrian Attribute Recognition (PAR)
[4] Person Search
[5] Pedestrian Detection
[6] Human Parsing
[7] Human Pose Estimation
[8] Self-supervised Learning
[9] Pretext
[10] Contrastive Learning

ماسک‌گذاری [11] طراحی شده‌اند. اگرچه این روش‌ها موفقیت‌های بزرگی در یادگیری بازنمایی‌های کلی تصویر کسب کرده‌اند، اما همچنان کمبودهایی در طراحی مدل‌ها با هدف وظایف متمرکز بر انسان وجود دارد.

در ادامه این فصل، ابتدا به شرح کوتاه و مختصر هر یک از این وظایف متمرکز بر انسان پرداخته شده و سپس موضوع اصلی این پروژه، یعنی شناسایی مجدد انسان و روش‌های پیشین مورد استفاده و مقایسه این روش‌ها با یکدیگر به طور جامع مورد بررسی قرار می‌گیرد.

۲ـ۱ تشخیص ویژگی‌های عابر پیاده

تشخیص ویژگی‌های عابر پیاده یک وظیفه مهم در حوزه بینایی کامپیوتر است، زیرا نقش اساسی در نظارت تصویری ایفا می‌کند؛ مانند اینکه آیا آنها در حال صحبت کردن با گوشی هستند، آیا کوله‌پشتی دارند و غیره. برای پیش‌بینی وجود یک ویژگی خاص، احتیاج به محل‌یابی نواحی مرتبط با ویژگی است. تشخیص ویژگی‌های عابر پیاده، مانند جنسیت، سن و سبک لباس، به دلیل پتانسیل زیاد در برنامه‌های نظارت ویدیویی مانند تایید چهره [12]، بازیابی فرد [13] و شناسایی مجدد انسان، توجه فراوانی را به خود جلب کرده است [۱۴].



شکل ۲ـ۱: مثالی از تشخیص ویژگی‌های عابر پیاده [۱]

[11] Masking Image Modeling
[12] Face Verification
[13] Person Retrieval

با توجه به تصویر فرد، تشخیص ویژگی‌های عابر پیاده به هدف پیش‌بینی یک گروه از ویژگی‌ها برای توصیف ویژگی‌های این فرد از یک لیست ویژگی‌های پیش‌تعریف شده می‌پردازد، به عنوان مثال همانطور که در تصویر ۲ـ۱ مشاهده می‌شود، ویژگی‌های مرد در جعبه قرمز عبارت است از: موهای کوتاه، همراه با کیف کاغذی، شلوار سیاه و غیره.

۲ـ۲ جستجوی فرد

جستجوی فرد یک وظیفه است که هدف آن تطابق یک فرد خاص در میان تعداد زیادی از تصاویر صحنه کلی است [۱۵]. جستجوی فرد در ویدیوها در حالات واقعی بسیار ضروری است. برای نمونه، جهت دستگیری یک جنایتکار مورد نظر، پلیس ممکن است بخواهد با تنها یک پرتره [۱۴]، از هزاران ساعت فیلم جمع‌آوری‌شده از چندین دوربین نظارتی مجرم را پیدا کند. البته در چنین شرایطی، از دوربین‌ها در محیط‌های متفاوت استفاده می‌شود تا تصاویر مرجع متمایز و بهتری داشته باشیم. همانطور که در شکل ۲ـ۲ نیز مشخص است، تصویر هدف با تصاویر مرجع تفاوت‌های قابل توجهی از نظر حالت، لباس، روشنایی و غیره دارد [۲]. بنابراین، حتی روش‌های تشخیص به‌روز [۱۵] نیز دشواری‌های زیادی در تشخیص قابل اعتماد شخص دارند.

شکل ۲ـ۲: یک نمونه از تصاویر موجود مجموعه‌داده CSM جهت جستجوی فرد [۲]

۲ـ۳ تشخیص عابر پیاده

تشخیص عابر پیاده، وظیفه شناسایی فرد پیاده از طریق دوربین است. این وظیفه یک چالش بزرگ در بینایی کامپیوتر است که کاربردهای مهمی دارد، مانند نظارت ویدیویی، کمک به رانندگی و ربات‌های هوشمند [۱۶].

[14] Portrait
[15] State-of-the-Art

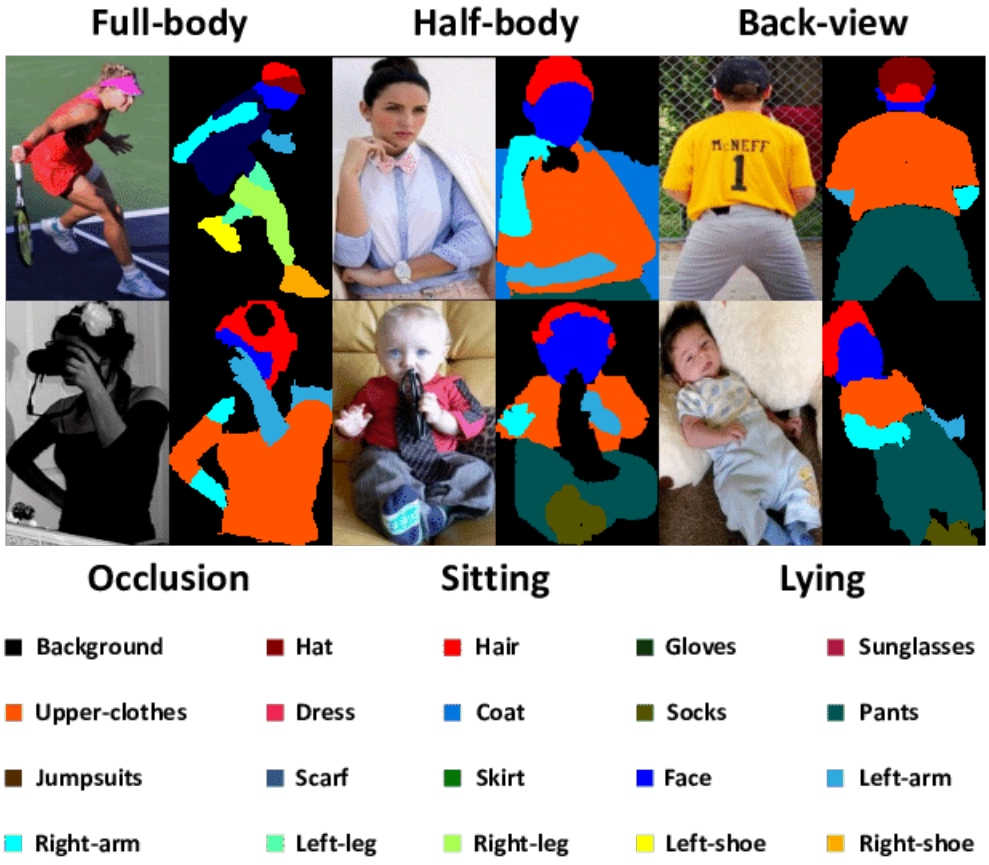
تشخیص‌گرهای [16] عمیق عابران پیاده، که از یادگیری ویژگی‌های عالی برای تصاویر توسط شبکه‌های عصبی عمیق بهره می‌برند، در سال‌های اخیر پیشرفت‌های بزرگی کرده‌اند [۱۷]. با این حال، تشخیص عابر پیاده ممکن است توسط مشکلاتی مانند مقیاس‌های مختلف ویژگی‌های عابر پیاده و پس‌زمینه شلوغ، تحت تاثیر قرار گیرد [۱۸].

شکل ۲ـ۳: مثالی از تشخیص عابران پیاده

۲ـ۴ تجزیه و تحلیل انسان

همان‌طور که در شکل ۲ـ۵ نشان داده شده است، تجزیه و تحلیل انسان، که همچنین به عنوان ناحیه‌بندی معنایی انسان [17] شناخته می‌شود، به مسئله اختصاص برچسب‌های معنایی دقیق (مانند "مو"، "صورت"، "لباس" و غیره) به هر پیکسل در تصویر مربوط می‌شود [۱۹]. این وظیفه یک وظیفه بسیار چالش‌برانگیز در حوزه بینایی کامپیوتر است و یکی از مراحل بسیار اساسی در راستای درک دقیق تصویر برای تحلیل متمرکز بر انسان است. روش‌های موفق تجزیه و تحلیل انسان می‌توانند در ارتقای برنامه‌های هوش مصنوعی [18] سطح بالاتر، مانند تحلیل رفتار انسان، تشخیص و بازیابی سبک لباس و پیشنهاد خودکار محصول، مفید باشند [۳].

[16] Detector
[17] Human Semantic Segmentation
[18] Artificial Intelligence (AI)

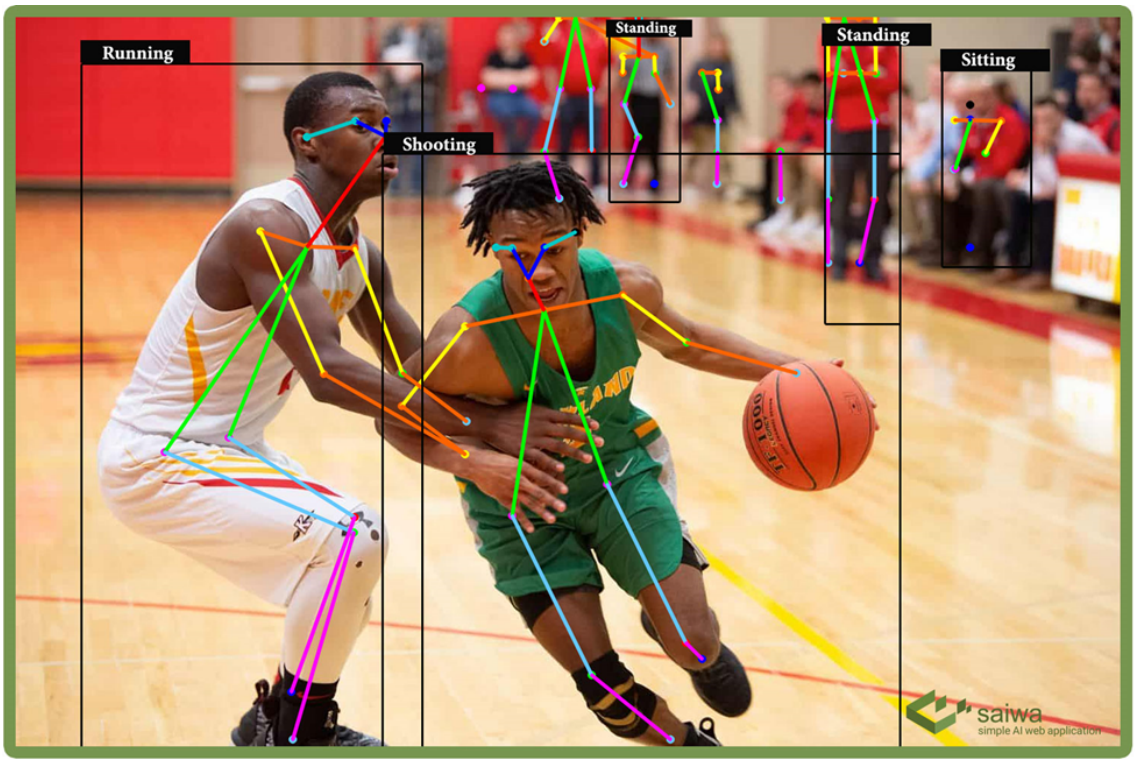


شکل ۲ـ۴: نمونه‌ای از وظیفه تجزیه و تحلیل انسان [۳]

۲ـ۵ تخمین حالت انسان

تخمین حالت یک وظیفه بینایی کامپیوتر است که هدف آن شناسایی موقعیت و جهت یک فرد یا یک شیء است. معمولاً، این کار با پیش‌بینی موقعیت نقاط خاصی مانند دست‌ها، سر، زانوها و غیره برای تخمین حالت انسان انجام می‌شود. تخمین حالت می‌تواند کاربردهای بسیاری مانند تعامل انسان و کامپیوتر [۱۹]، تطبیق حرکت [۲۰] و کنترل آواتار مجازی استفاده شود [۲۰].

[19] Human-Computer Interaction
[20] Motion Retargeting



شکل ۲ـ۵: یک مثال از تخمین حالت انسان

۲ـ۶ شناسایی مجدد انسان

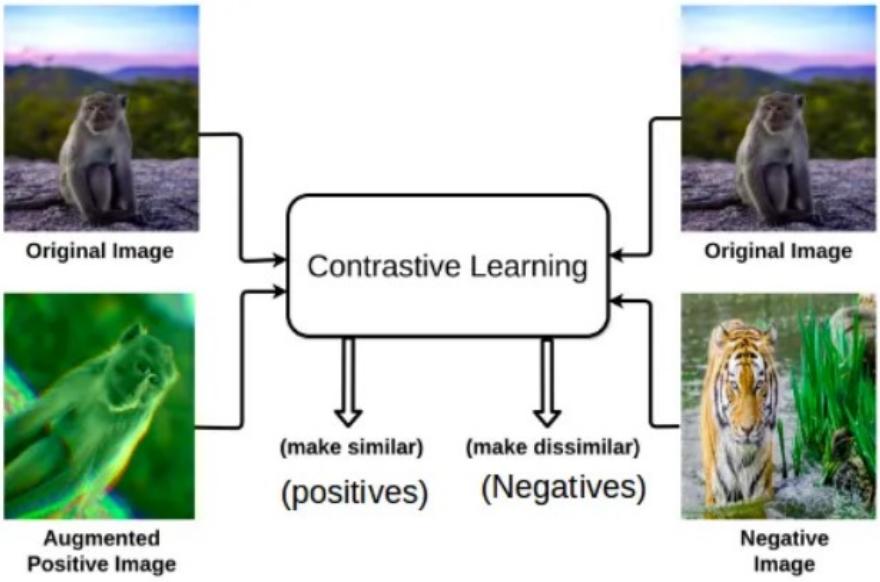
همانطور که می‌دانید، رویکردهای یادگیری خودنظارتی به میزان قابل توجهی در حوزه بینایی کامپیوتر به ویژه موضوع شناسایی مجدد جلب توجه کرده‌اند، به ویژه زمانی که به عنوان وظیفه پیشنهادی برای یادگیری بازنمایی تصویر استفاده می‌شوند. در این بخش، ابتدا به بررسی مختصر برخی از این مدل‌ها می‌پردازیم. سپس، روشی که به عنوان خط پایه [۲۱] مورد استفاده قرار می‌گیرد را به‌طور جامع معرفی کرده و راهبرد آن را در حل چالش‌های قبلی شرح می‌دهیم.

۲ـ۶ـ۱ روش‌های مقایسه‌ای

در یادگیری خودنظارتی کنونی، روش‌های مقایسه‌ای [۲۲] موفقیت‌های قابل توجهی را کسب کرده‌اند و بهترین عملکردها را برای بازنمایی تصویر ارائه کرده‌اند [۲۱]. اثبات شده است که این روش، قابلیت خوبی در

[21] Baseline
[22] Contrastive Methods

یادگیری ویژگی‌های تمییز دهنده [۲۳] دارد. ایده اصلی پشت یادگیری مقایسه‌ای این است که مدل را آموزش می‌دهد تا بین نمونه‌ها یا اشیا مشابه و مختلف تفاوت قائل شود.



شکل ۲ـ۶: نحوه انجام یادگیری مقایسه‌ای [۴]

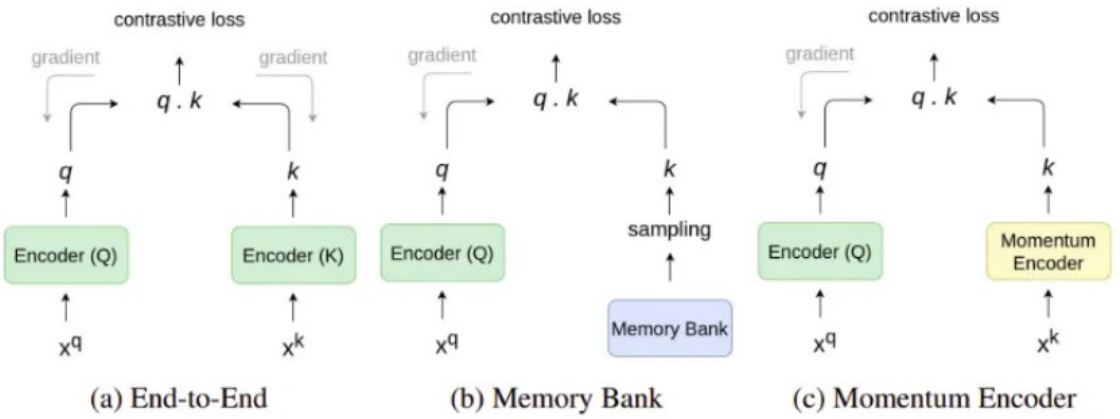
همانطور که در شکل ۲ـ۶ نشان داده شده است، نمونه‌های مشابه به عنوان نمونه‌های مثبت در نظر گرفته می‌شوند در حالی که نمونه‌های متفاوت به عنوان نمونه‌های منفی در نظر گرفته می‌شوند.

یادگیری مقایسه‌ای بیشتر بر روی تولید تعداد بیشتری از نمونه‌های منفی تمرکز دارد تا بازنمایی با کیفیت بالاتر ایجاد کند. بنابراین، معماری‌های مختلفی معرفی شدند تا تعداد نمونه‌های منفی را افزایش دهند [۴]. در ادامه به بررسی سه مدل کلی از روش‌های یادگیری مقایسه‌ای می‌پردازیم.

۲ـ۶ـ۱ـ۱ معماری یادگیری انتها به انتها

معماری انتها به انتها [۲۴]،مطابق با شکل ۲ـ۷، از دو کدگذار [۲۵]، یعنی کدگذار پرس‌وجو [۲۶] (Q) و کدگذار کلید (K) تشکیل شده است که هر دو می‌توانند هر معماری از شبکه عصبی همگشتی باشند (معمولاً از معماری

[23] Discriminative
[24] End-to-End
[25] Encoder
[26] Query



شکل ۲ـ۷: معماری‌های عمومی استفاده شده در روش‌های یادگیری مقایسه‌ای [۴]

ResNet50 استفاده می‌شود) تا ویژگی‌های مورد نیاز را استخراج کنند. این دو کدگذار به صورت جداگانه آموزش داده می‌شوند تا نمایش‌های متمایزی تولید شود. ویژگی‌های q از تصاویر اصلی استخراج می‌شود، در حالی که ویژگی‌های k از تصاویر تغییریافته و تصاویر منفی در دسته [۲۷] استخراج می‌شود. در نهایت، از ضرر متضاد [۲۸] برای محاسبه شباهت بین q و k استفاده می‌شود [۲۲]. یکی از مثال‌های این معماری، چارچوب کاری [۲۹] SimCLR (a Simple framework for Contrastive LeaRning) است [۲۳]. SimCLR نشان می‌دهد که در صورتی که دسته بسیار بزرگ باشد، بانک حافظه [۳۰] می‌تواند کاملاً با عناصر از همان دسته جایگزین شود.

نقطه ضعف این معماری این است که مدل نیاز به یک اندازه دسته بسیار بزرگ از نمونه‌های منفی دارد که حافظه را به طور زیادی مصرف می‌کند.

۲ـ۶ـ۱ـ۲ معماری بانک حافظه

برای حل نقاط ضعف موجود در ساختار انتها به انتها، استفاده از بانک حافظه پیشنهاد می‌شود. بانک حافظه شامل بازنمایی ویژگی برای تمام تصاویر در مجموعه‌داده است، این نمایش ویژگی برای هر تصویر به صورت

[27] Batch
[28] Contrastive Loss
[29] Framework
[30] Memory Bank

میانگین متحرک نمایی [۳۱] در دوره‌های [۳۲] قبلی محاسبه می‌شود. در دوره بعدی، بانک حافظه به‌روزرسانی و با نمونه‌های منفی و بازنمایی ویژگی‌شان جایگزین می‌شود [۲۴]. یک مثال از این ساختار، چارچوب PIRL است.

نقطه ضعف این ساختار این است که نیاز به به‌روزرسانی منظم نمایش ویژگی‌ها در بانک حافظه است که ممکن است منابع محاسباتی را به خود اختصاص دهد.

۲ـ۶ـ۱ـ۳ کدگذار سرعتی [۳۳]

ایده اصلی پشت ساختار کدگذار سرعتی استفاده از یک مدل آموزشی به جای آموزش دو مدل مستقل است. کدگذار سرعتی یک دیکشنری [۳۴] تولید می‌کند که یک صف از بازنمایی ویژگی‌های کلیدها (کلیدهای کدگذاری شده) را در اختیار دارد. این کلیدهای کدگذاری شده در ریزدسته [۳۵] فعلی در صف قرار داده می‌شوند، در حالی که همزمان قدیمی‌ترین ریزدسته از صف خارج می‌شود [۲۵].

کدگذار سرعتی دارای همان پارامترهای کدگذار Q است، اما به جای به‌روزرسانی پس‌انتشار [۳۶]، توسط به‌روزرسانی سرعتی زیر به‌روزرسانی می‌شود.

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \qquad (۱ـ۲)$$

در رابطه ۲ـ۱، m ضریب سرعتی می‌باشد، در حالیکه θ_k و θ_q پارامترهای کدگذار هستند.

یک مثال از این معماری، چارچوب Momentum Contrast (Moco) است که با ذخیره بازنمایی‌ها از یک کدگذار سرعتی به جای شبکه آموزش دیده، بهبودی در آموزش روش‌های مقایسه‌ای ایجاد می‌کند. در این روش، نمونه‌های داده در یک دیکشنری نگهداری می‌شود که کلیدهای این دیکشنری، تصاویر هستند و توسط یک شبکه کدگذار نمایش داده می‌شوند. این کدگذار وظیفه تولید بردارهای بازنمایی برای هر تصویر را دارد. هدف این است که یک پرس‌وجو کدگذاری شده به شکلی باشد که به کلید متناظر با آن شبیه باشد و از دیگران متمایز باشد.

[31] Exponential Moving Average (EMA)
[32] Epoch
[33] Momentum Encoder
[34] Dictionary
[35] Mini-Batch
[36] Backpropagation

۲ـ۶ـ۲ شبکه عصبی DINO

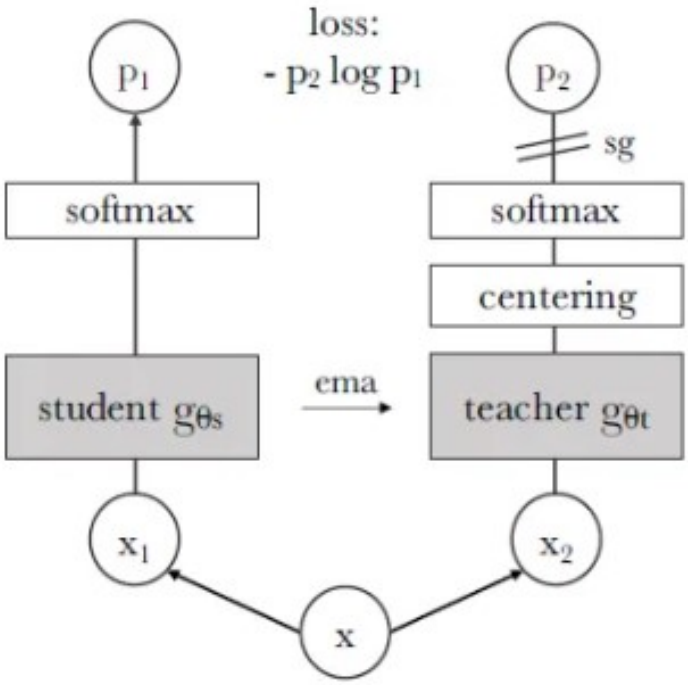
Dino (self-DIstillation with NO labels)، روش‌های بیشتری را ترکیب می‌کند، شامل کدگذار سرعتی، آموزش چند برشی [۳۷] [۲۶] و استفاده از قطعات [۳۸] کوچک با مبدل تصویر [۳۹] که یک پایه بسیار بهتری را ایجاد می‌کند. در واقع، روش خودنظارتی بر روی مبدل تصویر استفاده می‌شود و باعث تشکیل DINO می‌شود. همانطور که در تصویر ۲ـ۸ نشان داده شده است، ویژگی‌های خودنظارتی ViT اطلاعات صریحی درباره تقسیم‌بندی معنایی تصویر دارند. همچنین ویژگی‌های استخراج شده به عنوان طبقه‌بندهای [۴۰] k-NN عالی عمل می‌کنند.

شکل ۲ـ۸: Self-attention یک مبدل تصویر با قطعه‌های ۸×۸ که بدون نظارت آموزش دیده است. [۵]

۲ـ۶ـ۲ـ۱ معماری شبکه DINO

در DINO، مدل دو تبدیل [۴۱] تصادفی متفاوت از تصویر ورودی را به شبکه دانش‌آموز [۴۲] (g_{θ_s}) و شبکه معلم [۴۳] (g_{θ_t}) منتقل می‌کند. هر دو شبکه دانش‌آموز و معلم، معماری یکسان اما با پارامترهای متفاوتی دارند. خروجی شبکه معلم با میانگین محاسبه شده روی دسته، مرکزگرا [۴۴] [۲۶] می‌شود. هر دو شبکه خروجی یک ویژگی با ابعاد K را تولید می‌کنند که با P_s و P_t نشان داده می‌شود، به عبارت دیگر، توزیع‌های احتمال خروجی، که با

[37] Multi-Crop
[38] Patch
[39] Vision Transformer (ViT)
[40] Classifier
[41] Transformation
[42] Student
[43] Teacher
[44] Centralize



شکل ۲ـ۹: معماری شبکه عصبی DINO [۵]

temperature softmax (τ_s) روی بعد ویژگی نرمال شده‌اند طبق رابطه ۲ـ۲ می‌باشد.

$$P_s(x)^{(i)} = \frac{exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\Sigma_{k=1}^{K} exp(g_{\theta_s}(x)^{(k)}/\tau_s)} \qquad (۲ـ۲)$$

با یک معلم ثابت، شباهت آن‌ها سپس با یک تابع ضرر متقابل [۴۵] مطابق رابطه ۲ـ۳ اندازه‌گیری می‌شود:

$$\min_{\theta_s} H(P_t(x), P_s(x)) \quad , \quad where \quad H(a,b) = -a \log^b \qquad (۲ـ۳)$$

قابل ذکر است که از یک تصویر داده شده، مجموعه‌ای از نماهای مختلف (V) تولید می‌شود. این مجموعه دو نمای سراسری [۴۶]، xg1 و xg2 و چندین نمای محلی [۴۷] با وضوح [۴۸] کمتر را شامل می‌شود. تمام برش‌ها از شبکه دانش‌آموز عبور می‌کنند در حالی که فقط دیدگاه‌های سراسری از شبکه معلم عبور می‌کنند، بنابراین روش

[45] Cross-Entropy Loss
[46] Global
[47] Local
[48] Resolutuin

"تطابق‌های محلی به سراسری [۴۹]" انجام می‌شود.

همچنین تابع ضرر با استفاده از رابطه ۲ـ۴ کمینه می‌شود تا آموزش به درستی صورت بگیرد [۵].

$$\min_{\theta_s} \sum_{x \in x_1^g, x_2^g} \sum_{\substack{x' \in V \\ x' \neq x}} H(P_t(x), P_s(x')) \qquad (۴ـ۲)$$

۲ـ۶ـ۲ـ۲ نقش میانگین متحرک نمایی (EMA) در جلوگیری از فروپاشی

مرکزگذاری از غالب شدن یک بعد جلوگیری می‌کند اما فروپاشی را به توزیع یکنواخت [۵۰] تشویق می‌کند، در حالی که تیزکردن [۵۱] اثر عکس دارد. بنابراین، اعمال هر دو عملیات تعادل اثرات آنها را تضمین می‌کند [۲۷].

توقف گرادیان [۵۲] است که بر روی مدل معلم اعمال می‌شود تا گرادیان‌ها فقط از طریق دانش‌آموز منتقل شوند. پارامترهای معلم با میانگین متحرک نمایی (ema) از پارامترهای دانش‌آموز به‌روزرسانی می‌شوند.

عملیات مرکزگذاری تنها به آمار دسته اولیه بستگی دارد و می‌توان آن را به عنوان افزودن یک بایاس [۵۳] (c) به معلم تفسیر کرد:

$$g_t(x) \leftarrow g_t(x) + c \qquad (۵ـ۲)$$

مرکز c با یک میانگین متحرک نمایی (EMA) به‌روزرسانی می‌شود که این امکان را برای این روش فراهم می‌کند تا به خوبی در اندازه‌های دسته‌های مختلف عمل کند:

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i) \qquad (۶ـ۲)$$

در رابطه ۲ـ۶، $m > 0$ پارامتر نرخ است و B اندازه دسته را تعیین می‌کند.

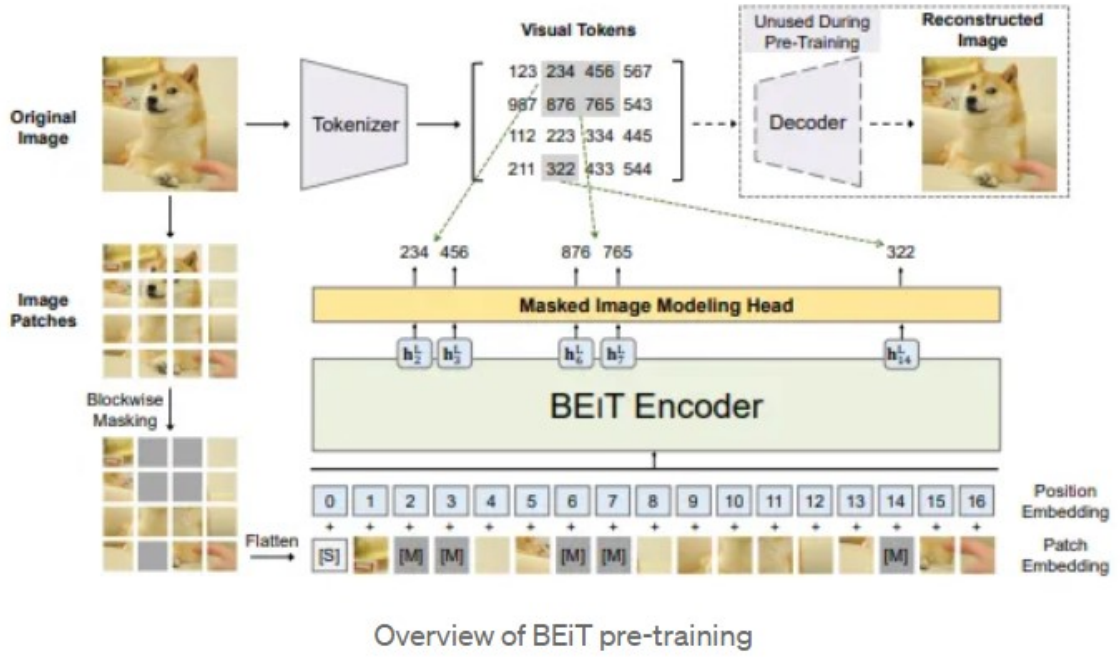
[49] Local-to-Global
[50] Uniform Distribution
[51] Sharpening
[52] Stop-Gradient (SG)
[53] Bias

۲ـ۶ـ۳ مدل‌سازی تصویر ماسک‌شده

علاوه بر روش‌های مقایسه‌ای، روش‌های مدل‌سازی تصویر ماسک‌شده [۵۴] توجه گسترده‌ای از سوی پژوهش‌گران به دنبال داشتند. این روش، شامل پیش‌بینی پیکسل‌های [۵۵] گم‌شده در یک تصویر با استفاده از پیکسل‌های اطراف به عنوان زمینه است. این روش اغلب در پرکردن تصاویر، جایگزین‌کردن بخش‌های گم‌شده یا آسیب‌دیده تصویر با استفاده از اطلاعات از مناطق اطراف استفاده می‌شود. در ادامه به بررسی ۲ مدل کلی از این روش می‌پردازیم.

۲ـ۶ـ۳ـ۱ BEiT

BEiT (Bidirectional Encoder representation from Image Transformers) از یک خودکدگذار برای کدگذاری قطعه‌های تصویر به عنوان توکن‌های [۵۶] گسسته استفاده می‌کند. سپس، یک مدل مبدل پیش‌آموزش داده می‌شود تا مقادیر توکن‌های گسسته برای توکن‌های ماسک‌شده را پیش‌بینی کند [۶].



شکل ۲ـ۱۰: نمای کلی پیش‌آموزش BEiT [۶]

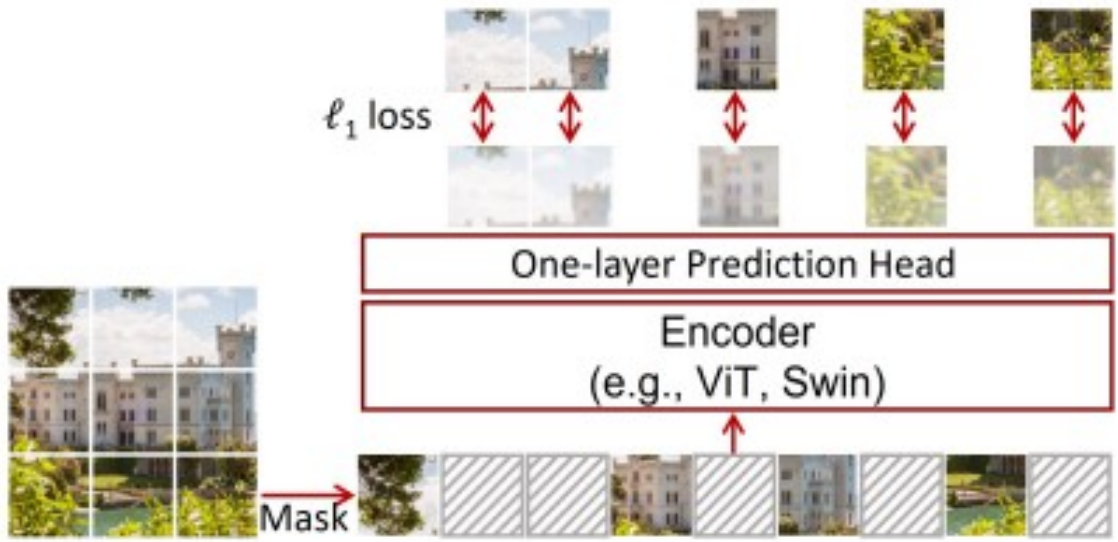
[54] Masked Image Modeling
[55] Pixle
[56] Token

این روش، از BERT [۲۸] الهام گرفته است. ابتدا برای هر تصویر، دو نما تولید می‌شود که شامل قطعات تصویر و توکن‌های تصویری [۵۷] هستند. قطعات تصویر از تصویر جدا می‌شوند و سپس برخی از آن‌ها به صورت تصادفی مسدود می‌شوند. این قطعات به طور خطی جاسازی می‌شوند و توکن‌های موقعیت [۵۸] به آن‌ها اضافه می‌شوند. توکن‌های جاسازی شده به شبکه اصلی مبدل منتقل می‌شوند. درنتیجه، این شبکه یاد می‌گیرد تا توکن‌های تصویری گم‌شده در تصویر را بازسازی کند.

این معماری یک تطبیق هوشمندانه از BERT بر روی داده‌های تصویری است. اما به نظر می‌رسد به دلیل ذخیره و پردازش تعداد زیادی توکن، نیاز به قدرت محاسباتی بزرگی دارد.

۲ـ۶ـ۳ـ۲ SimMIM

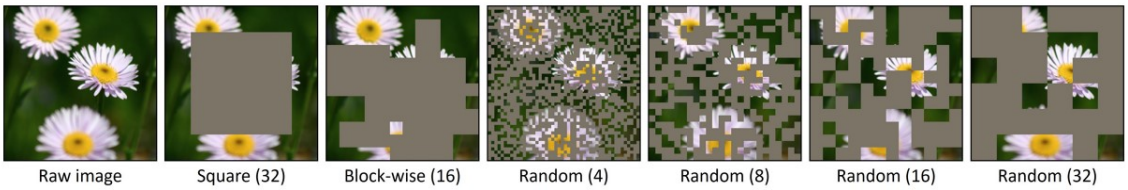
SimMIM (a simple framework for Masked Image Modeling) به جای بازسازی توکن‌های گسسته تصویر که از کدگذار استخراج شده‌اند، به طور مستقیم بخش‌های تصویر ماسک‌شده را بازسازی می‌کند [۷]. ایده این روش، ماسک کردن تصادفی بخش‌های تصویر و استفاده از یک لایه خطی [۵۹] برای بازگرداندن مقادیر خام پیکسل منطقه ماسک‌شده با یک هزینه $\ell 1$ می‌باشد. (تصویر ۲ـ۱۱)



شکل ۲ـ۱۱: نمای کلی SimMIM [۷]

[57] Visual Token
[58] Position Embedding
[59] Linear Layer

این مدل، از طراحی‌های متفاوتی برای ماسک کردن تصویر استفاده می‌کند. انواع این روش‌ها را می‌توانید در تصویر ۲ـ۱۲ ببنید.



شکل ۲ـ۱۲: روش‌های مختلف ماسک کردن مدل SimMIM [۷]

۲ـ۶ـ۴ شبکه عصبی SOLIDER

اگرچه مدل‌سازی تصویر ماسک شده موفق به وارد کردن اطلاعات معنایی به بازنمایی‌های تصویری می‌شود، اما نمی‌تواند به طور صریح اطلاعات معنایی را از تصویر استخراج کند تا به آموزش نظارت شود. بنابراین، در اینجا از مدل SOLIDER (a Semantic cOntrollable seLf-supervised lEaRning) [۸]، به عنوان مدل شناسایی مجدد که قرار است با مدل زمانی ما ترکیب شود، استفاده می‌کنیم. در این مدل، ابتدا توکن‌ها را خوشه‌بندی [۶۰] کرده و از دانش قبلی انسان برای اختصاص برچسب‌های معنایی به این توکن‌ها استفاده می‌شود که می‌تواند یک بازنمایی معنایی قوی‌تر را برای انسان آموزش دهد.

همانطور که در بخش ۲ـ۶ـ۲ نیز اشاره شد، DINO [۵] یک روش پیشرفته و به‌روز برای یادگیری بازنمایی تصویر می‌باشد. به عنوان یک روش مبتنی بر یادگیری مقایسه‌ای، اطلاعات ظاهری تصویر در آن به خوبی یاد گرفته می‌شود. بنابراین، از DINO به عنوان پایه در SOLIDER استفاده می‌شود و هدف اضافه کردن اطلاعات معنایی بیشتر به نمایش DINO می‌باشد.

۲ـ۶ـ۴ـ۱ الگوریتم

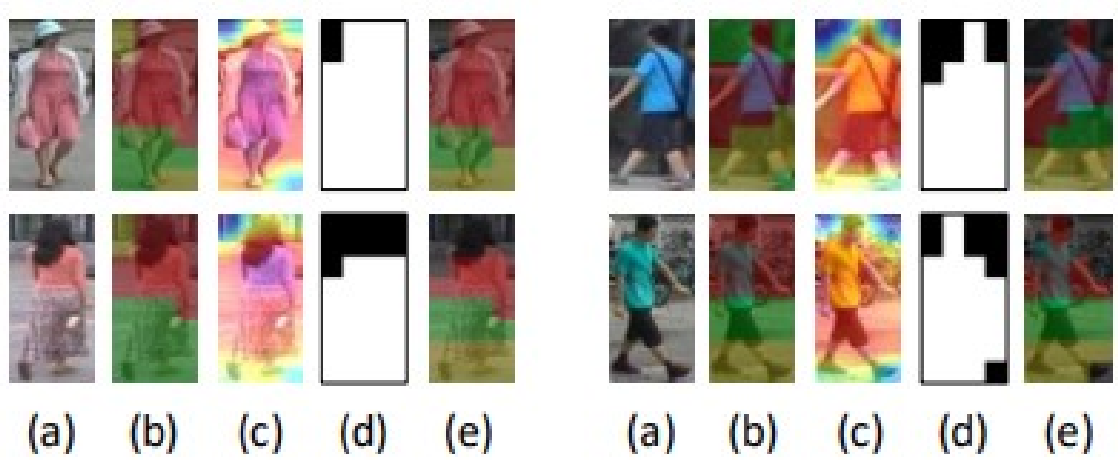
تحقیقات نشان داده‌اند که در تصاویر بدون برچسب انسان الگوی ثابتی وجود دارد؛ بدن افراد به صورت عمودی تصویر را پوشانده و سر همواره در بالای تصویر قرار دارد [۲۹]. بر اساس این مشاهده، برچسب‌های معنایی برای بخش‌های خوشه‌بندی شده بر اساس ترتیب مختصات محور y آن‌ها، اختصاص داده شد. شبه

[60] Clustering

برچسب‌های معنایی اطلاعات معنایی را برای هر بردار توکن فراهم می‌کنند. در این مقاله، برای استفاده بهتر از این برچسب‌ها، تغییرات اضافی ایجاد می‌شود که در ادامه آمده است:

- در نظر گرفتن خوشه‌بندی پس‌زمینه و پیش‌زمینه [61]

برای رفع نویز از پس زمینه در فرآیند خوشه‌بندی، یک مرحله خوشه‌بندی اضافی قبل از خوشه‌بندی معنایی معرفی می‌شود. با توجه به اینکه توکن‌های پس‌زمینه همیشه پاسخ‌های کوچکتری نسبت به توکن‌های پیش‌زمینه دارند، بردارهای توکن به دو دسته‌ی پیش‌زمینه و پس‌زمینه براساس بزرگی بردار آن‌ها خوشه‌بندی می‌شود [۳۰]. نتایج در شکل ۲ـ۱۳ (c) نشان داده شده است. سپس خوشه‌بندی معنایی فقط روی توکن‌های پیش‌زمینه انجام شده و نتایج جدید در شکل ۲ـ۱۳ (d) نشان داده شده است. می‌توان دید که توکن‌های پیش‌زمینه به خوبی در سه بخش معنایی مانند بالا تنه، پایین تنه و کفش‌ها خوشه‌بندی شده‌اند. a تصویر ورودی، b تصویر پس از خوشه‌بندی معنایی اولیه و e نتیجه نهایی می‌باشد.



شکل ۲ـ۱۳: مراحل انجام خوشه‌بندی [۸]

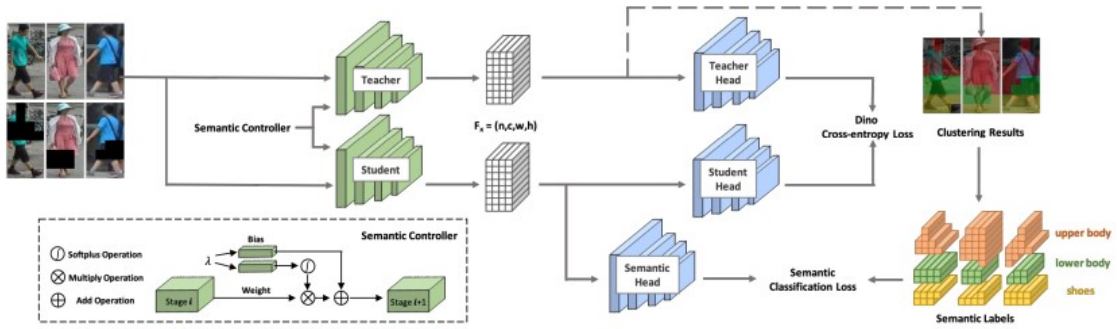
- در نظر گرفتن مدل‌سازی تصویر ماسک شده

در بخش ۲ـ۶ـ۳ در مورد مدل‌سازی تصویر ماسک شده و روش کار آن، توضیح داده شد؛ اما در این مقاله، هدف افزودن ویژگی‌های معنایی بیشتر به بازنمایی تصویر است. بنابراین، فرض می‌شود که

[61] Foreground (FG)

اگر یک قسمت معنایی از تصویر انسان گم‌شده یا پوشیده شده باشد، مدل همچنان قادر خواهد بود تا معنای آن را بر اساس سایر قسمت‌های اطراف پیش‌بینی کند. برای این منظور در این مقاله، نظارت معنایی را به نظارت معنایی ماسک شده ارتقا می‌دهند. به طور خاص، یک قسمت معنایی را به صورت تصادفی از تصویر x پنهان می‌شود و این تصویر ماسک شده $\tilde{x}$ را دوباره به مدل داده می‌شود. سپس توکن‌های خروجی توسط برچسب‌های معنایی اصلی نظارت می‌شوند، زیرا انتظار می‌رود که مدل بتواند با کمک توکن‌های دیگر، برچسب‌های معنایی واقعی را برای توکن‌های ماسک شده ارائه دهد.

مدل خودنظارتی معنایی در تصویر ۲ـ۱۴ نشان داده شده است. به طور خاص، در طول آموزش، برای هر دوره، نقشه‌های ویژگی خروجی (F) از مدل پایه با اندازه (n, c, h, w) به دست می‌آید. برای هر تصویر x، نقشه‌های ویژگی آن F_t [۶۲] به عنوان بردارهای توکن w * h با اندازه c در نظر گرفته می‌شود. سپس از K-means [۳۱] استفاده می‌شود تا آن‌ها به دو دسته براساس مقدار c خوشه‌بندی شوند و دسته با بزرگترین مقدار به عنوان ماسک زمینه (پیش‌زمینه) M در نظر گرفته می‌شود. پس از آن، K-means دیگری بر روی توکن‌ها در ماسک پیش‌زمینه M به منظور به دست آوردن N دسته معنایی پیش‌تعریف شده و اختصاص برچسب معنایی y به هر توکن، برای انجام خوشه‌بندی معنایی اعمال می‌شود. در عین حال، یک سری semantic head جهت دسته‌بندی بردارهای F_s از شاخه دانش‌آموز بر اساس این برچسب‌های معنایی y مشغول به کار است.



شکل ۲ـ۱۴: نمای کلی SOLIDER [۸]

[۶۲] نقشه ویژگی‌های F_t از شبکه معلم برای خوشه‌بندی معنایی انتخاب می‌شود. نقشه ویژگی‌های F_s از شبکه دانش‌آموز انتخاب می‌شود.

تابع ضرر طبقه‌بندی معنایی [۶۳]، توسط رابطه ۲ـ۷ تعریف می‌شود:

(۲ـ۷)

$$L_{sm} = \frac{1}{w \times h} \sum_{\substack{u \in w \\ v \in h}} \sum_{i=1}^{N+1} -y^{(u,v)} \log \frac{f_s(u,v)^{(i)}}{\sum_{k=1}^{N+1} f_s(u,v)^{(k)}} \quad where \quad f_s = h_{sm(Flatten(F_s))}$$

$flatten()$ به منظور تغییر شکل F_s از (n, c, h, w) به $(n * c * h * w)$ استفاده می‌شود. همچنین h_{sm} نیز نمایانگر semantic head می‌باشد. N تعداد بخش‌های معنایی خوشه‌بندی شده را بیان می‌کند. $f_s(u,v)^{(i)}$ احتمال پیش‌بینی شده برای توکن (u,v) بر روی بخش i را نشان می‌دهد. پس از تصویر x، به طور تصادفی یک قسمت از x ماسک می‌شود تا تصویر $\tilde{x}$ بدست بیاید و آن هم به معادله ۲ـ۷ ارسال شود.

مقدار نهایی ضرر چارچوب SOLIDER، توسط رابطه زیر به‌دست می‌آید:

$$L = \alpha L_{dino} + (1 - \alpha) L_{sm} \qquad (۲ـ۸)$$

α ضریب تعادل است و با تجربه ۰.۵ تنظیم شده است.

- کنترل‌گر معنایی

مدل SOLIDER، یک مدل جامع برای تمامی وظیفه‌های متمرکز بر انسان که به برخی از آن‌ها در بخش ۲ پرداخته شد، می‌باشد. با توجه به این موضوع، اطلاعات ظاهری و معنایی برای هر وظیفه باید با نسبت متناسبی ترکیب شوند. برای نمونه، در وظیفه شناسایی مجدد انسان اطلاعات ظاهری از اطلاعات معنایی مهم‌تر می‌باشد، درحالی که در وظیفه تشخیص عابر پیاده اطلاعات معنایی نقش مهم‌تری ایفا می‌کنند. برای حل این مشکل، این مقاله از یک شبکه شرطی استفاده می‌کند. همانطور که در تصویر ۲ـ۱۴ نیز مشاهده می‌شود، ورودی کنترل‌گر معنایی، نقشه ویژگی تصویر و یک مقدار پیوسته $\lambda \in [0, 1]$ است. خروجی این شبکه، نقشه ویژگی با در نظر گرفتن نسبت مورد نیاز می‌باشد. در کنترل‌گر معنایی، مقدار λ به یک بردار وزن و یک بردار انحراف کدگذاری می‌شود. پس از استفاده از تابع فعال‌سازی Softplus، بردار وزن بر روی نقشه‌های ویژگی اصلی ضرب می‌شود و بردار انحراف برای خروجی‌های نهایی اضافه می‌شود.

[63] Semantic Classification Loss

کنترل‌گر معنایی پس از هر بلوک [۶۴] از شبکه اصلی Swin اعمال می‌شود و نقشه‌های ویژگی $F(\lambda)$ با نسبت جدید λ به بلوک بعدی ارسال می‌شود. از معادله ۲ـ۹ برای ایجاد مدل کنترل‌پذیر معنایی استفاده می‌شود.

$$L = \alpha L_{dino}(F(\lambda)) + \lambda(1-\alpha)L_{sm}(F(\lambda)) \qquad (۹ـ۲)$$

۲ـ۶ـ۴ـ۲ مجموعه‌داده

برای وظایف پیش‌بینی، مجموعه‌داده LUPerson [۳۲] برای آموزش استفاده می‌شود. این مجموعه‌داده شامل ۱۸.۴ میلیون تصویر انسان بدون برچسب است. از هر وظیفه پایین‌دست، آزمایش‌هایی بر روی مجموعه‌داده‌های معمولی آن‌ها انجام می‌شود. به‌طور خاص، در شناسایی مجدد انسان، آزمایش‌ها بر روی مجموعه‌داده‌های Market1501 و MSMT17 انجام می‌شود. در تشخیص ویژگی‌ها، مجموعه‌داده‌های PETAzs، RAPzs و PA100k مورد استفاده قرار می‌گیرند. در جستجوی فرد، از CUHK-SYSU و PRW استفاده می‌شود. CityPerson برای تشخیص عابرین پیاده به‌کار می‌رود. در تجزیه و تحلیل انسان و تخمین حالت، به ترتیب از مجموعه‌داده‌های LIP و COCO استفاده می‌شود.

با توجه به اینکه موضوع پروژه در مورد شناسایی مجدد انسان است، به بررسی کوتاه مجموعه‌داده Market1501 می‌پردازیم.

مجموعه‌داده Market1501 یکی از مجموعه‌های معروف در زمینه شناسایی مجدد انسان است. این مجموعه‌داده شامل تصاویری از افراد در محیط‌های مختلف مانند خیابان‌ها، پارک‌ها و مراکز خرید است. این مجموعه‌داده شامل ۱۵۰۱ هویت است که توسط شش دوربین مختلف ضبط شده‌اند و ۶۶۸،۳۲ جعبه محدوده تصویر افراد را با استفاده از آشکارساز عابر پیاده دریافت کرده است. هر فرد به طور میانگین ۳.۶ تصویر در هر نما دارد. این مجموعه‌داده به دو بخش تقسیم شده است: ۷۵۰ هویت برای آموزش [۶۵] و ۷۵۱ هویت باقی‌مانده برای گالری [۶۶] استفاده می‌شود.

مجموعه‌های پرس‌وجو و گالری ممکن است دیدهای دوربین مشابهی داشته باشند، اما برای هر هویت پرس‌وجو فردی، نمونه‌های گالری از همان دوربین مربوطه حذف می‌شوند. در واقع، تنها یک نمونه به صورت

[64] Block
[65] Train
[66] Gallery

تصادفی برای هر هویت گالری نمونه‌برداری نمی‌شود. این به این معناست که پرس‌وجو همیشه با نزدیک‌ترین نمونه مثبت در گالری همخوانی خواهد داشت، در حالی که نمونه‌های مثبت دیگری که دشوارتر هستند در محاسبه معیار CMC مورد توجه قرار نمی‌گیرند.

فرمت نام‌گذاری هر عکس به صورت XXXX_cYsZ_TTTTT_RR.jpg می‌باشد. بخش اول (XXXX)، یک عدد چهار رقمی است که هویت فرد را مشخص می‌کند. بخش دوم، شماره دوربین (Y) و دنباله [۶۷] (Z) را نشان می‌دهد. در قسمت سوم نیز شماره قاب [۶۸] آمده است و بخش نهایی، شماره جعبه محدود کننده را مشخص می‌کند.

جدول ۲ـ۱: مجموعه‌داده Market1501

subset	# ids	# images	# cameras
train	751	12936	6
query	750	3368	6
gallery	751	15913	6

در جدول ۳ـ۱، تعداد نمونه‌های موجود در هر دسته از مجموعه‌داده را مشاهده می‌نمایید.

۲ـ۶ـ۴ـ۳ معیار ارزیابی

در شناسایی مجدد شخص و جستجوی شخص، mAP/Rank1 به عنوان معیارهای ارزیابی انتخاب می‌شوند. برای تشخیص ویژگی‌ها، معیارهای ارزیابی دقت میانگین (mA) [۶۹] به کار می‌روند. در حالت تجزیه و تحلیل انسان، معیار mIoU و برای تخمین وضعیت، معیارهای دقت متوسط [۷۰] / بازخوانی [۷۱] (AP/AR) به عنوان معیارهای ارزیابی استفاده می‌شوند.

در ادامه، بیشتر به بررسی این معیارهای ارزیابی می‌پردازیم.

[67] Sequence
[68] Frame
[69] Mean Accuracy
[70] Average Precision
[71] Average Recall

۲ـ۶ـ۴ـ۴ نتیجه‌گیری

این مقاله [۸] یک چارچوب یادگیری خودنظارتی با قابلیت کنترل معنایی به نام SOLIDER پیشنهاد می‌دهد. این چارچوب قادر است از دانش پیشین درباره تصاویر انسان برای آموزش بازنمایی‌های با اطلاعات معنایی بیشتر استفاده کند. علاوه بر این، مدل پیش‌آموزش داده شده از SOLIDER می‌تواند توسط یک مقدار ورودی از طریق کنترل‌گر معنایی تنظیم شود، که می‌تواند بازنمایی‌های با نسبت‌های مختلفی از اطلاعات معنایی تولید کند و نیاز وظایف خردتر را برآورده سازد. بازنمایی‌های انسانی از SOLIDER بر روی شش وظیفه بصری متمرکز بر انسان تأیید شده است، که می‌تواند توسعه این وظایف متمرکز بر انسان در جامعه بینایی کامپیوتر را ترویج کند.

۲ـ۶ـ۵ معیارهای ارزیابی

معیار ارزیابی در واقع یک مقیاس یا سنجه است که برای ارزیابی کارایی یک مدل یا الگوریتم در حوزه خاصی از پردازش داده‌ها استفاده می‌شود. در حوزه شناسایی مجدد انسان، که به طور خاص در پردازش تصاویر و فیلم‌ها مورد استفاده قرار می‌گیرد، ارزیابی کارایی مدل‌ها بسیار حیاتی است. در این حوزه، ما با تمرکز بر روی شناسایی مجدد یا بازیابی اطلاعات، به دنبال این هستیم که مدل‌ها به درستی اجسام، اشیا یا انسان‌ها را در تصاویر شناسایی کنند. در ادامه، به بررسی پرکاربردترین معیارهای ارزیابی در حوزه شناسایی مجدد انسان می‌پردازیم.

۲ـ۶ـ۵ـ۱ معیار **CMC (cumulative matching characteristic)**

این معیار نشان می‌دهد که برای هر پرس‌وجو، چه تعداد تصاویر گالری در رتبه‌بندی اولین تصویر گالری مطابقت دارد.

$$CMC = \frac{\text{تعداد موارد مطابقت تا تصویر اول}}{\text{تعداد کل پرس‌وجوها}} \qquad (۲ـ۱۰)$$

۲ـ۶ـ۵ـ۲ معیار MAP (Mean Average Precision)

این معیار میزان دقت رتبه‌بندی شده را بر اساس نرخ میانگین تشخیص محاسبه می‌کند.

$$MAP = \frac{1}{N}\sum_{k=1}^{N}\frac{\sum_{i=1}^{k}\text{Precision}(i)\times\text{Relevance}(i)}{\text{Documents Relevant of Number}} \qquad (۱۱ـ۲)$$

- N تعداد تمامی نتایج بازیابی است
- $Precision(i)$ دقت $(Precision)$ در رتبه i ام است.
- $Relevance(i)$ میزان ارتباط موجود بین نتیجه i ام و مسئله مورد بررسی است.

مثال: اگر دو مسئله داشته باشیم و اولین نتیجه بازیابی با دقت $\frac{1}{1}$ و دومین نتیجه با دقت $\frac{1}{2}$ باشد و همچنین تمام نتایج بازیابی ارتباطی نداشته باشند، در این صورت MAP برابر با میانگین دقت دو نتیجه مذکور خواهد بود، یعنی:

$$\frac{\frac{1}{2}+\frac{1}{1}}{2} = \frac{3}{4} \qquad (۱۲ـ۲)$$

۲ـ۶ـ۵ـ۳ معیار Precision

این معیار، بیانگر نسبت تعداد نمونه‌هایی که به درستی شناسایی شده به کل نمونه‌هایی که توسط مدل به عنوان مثبت تشخیص داده شده‌اند، می‌باشد.

$$Precision = \frac{TP}{TP+TF} \qquad (۱۳ـ۲)$$

۲ـ۶ـ۵ـ۴ معیار Recall

این معیار، نمایانگر نسبت نمونه‌هایی که به درستی شناسایی شده به کل نمونه‌های مثبت موجود در مجموعه‌داده است.

$$Recall = \frac{TP}{TP+FN} \qquad (۱۴ـ۲)$$

۲ـ۶ـ۵ـ۵ معیار F1 (F1-score)

این معیار، از میانگین هارمونیک Precision و Recall که نشان دهنده تعادل میان Precision و Recall است، به دست می‌آید.

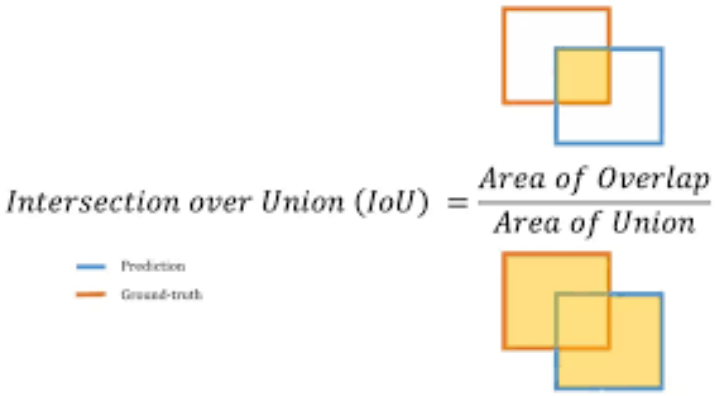
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (۱۵ـ۲)$$

۲ـ۶ـ۵ـ۶ معیار IoU

این معیار، نسبت مساحت تقاطع ناحیه پیش‌بینی شده توسط مدل و ناحیه واقعی اجسام به مجموع مساحت اجسام پیش‌بینی شده و واقعی را نشان می‌دهد.

$$IoU = \frac{TP}{TP + FP + FN} \quad (۱۶ـ۲)$$

در شکل ۲ـ۱۵، نحوه محاسبه این معیار نشان داده شده است.

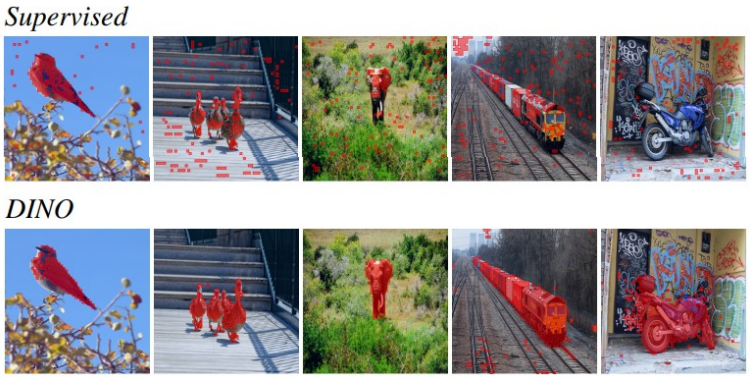


شکل ۲ـ۱۵: نحوه کارکرد معیار IOU

فصل ۳

روش پیشنهادی

در حوزه پردازش تصویر و هوش مصنوعی، شناسایی مجدد انسان یکی از مسائل مهم و چالش‌برانگیز است که به دلیل کاربردهای گسترده‌ای از جمله نظارت امنیتی، خودروهای هوشمند و پزشکی، توجه بسیاری از پژوهش‌گران را به خود جلب کرده است. در این پروژه، ما به بهبود مدل شناسایی خود می‌پردازیم. در این قسمت، بر مدل SOLIDER که در بخش ۲ـ۶ـ۴ توضیح داده شد و گفته شد بر پایه معماری DINO است، تمرکز کرده و به منظور افزایش دقت و کارایی آن، قصد داریم تا به این مدل ابزارهای جدیدی اضافه کنیم. همانطور که در تصویر ۳ـ۱ نیز مشاهده می‌شود، روش DINO نسبت به سایر روش‌های خودنظارتی عملکرد بهتری دارد و اشیا، حیوانات، وسایل نقلیه و غیره را بهتر تشخیص می‌دهد. به همین دلیل، روش SOLIDER که خود به بهبود روش DINO پرداخته بود را به عنوان مدل شناسایی انتخاب کردیم.



شکل ۳ـ۱: مقایسه مدل DINO با روش نظارت شده [۵]

یکی از رویکردهایی که در این پروژه استفاده می‌شود، اضافه کردن اطلاعات زمانی به مدل است. با در نظر گرفتن زمان ظاهر شدن هر فرد در دوربین، ما می‌توانیم از این اطلاعات برای بهبود دقت شناسایی استفاده کنیم. این امکان به مدل اجازه می‌دهد تا با ترکیب اطلاعات زمانی و ویژگی‌های تصویری، عملکرد بهتری داشته باشد و نتایج دقیق‌تری را ارائه دهد.

اطلاعات زمانی معمولاً زمانی معتبر هستند که یک فرد در دو دوربین مختلف قطعی مشاهده شود. به عبارت دیگر، فرض می‌شود که هنگامی که یک فرد از دوربین اول عبور می‌کند، با احتمال بالا از دوربین دوم نیز عبور کند. این فرض برای محاسبه‌ی اختلاف زمانی بین دو دوربین ضروری است، زیرا به زمان ظاهر شدن فرد در هر دو دوربین نیاز داریم. به عنوان مثال، اگر یک فرد از دوربین اول عبور کرده و مسیری را انتخاب کند که از دوربین دوم عبور نکند، این اطلاعات قابل اعتماد نیست. در اینجا چون مجموعه‌داده شامل تصاویر افرادی است که از درب ۱ حراست بانوان دانشگاه علم و صنعت عبور می‌کنند، فرض بر این است که اگر یک فرد از دوربین اول عبور کند، حتماً از دوربین دوم نیز عبور خواهد کرد. اما در مواردی که دو دوربین در دانشکده در نظر گرفته شود، ممکن است یک فرد از دوربین اول عبور کند اما از دوربین دوم عبور نکند.

در این بخش، ما به توضیح روش‌های پیشنهادی برای بهبود مدل SOLIDER و افزودن اطلاعات زمانی به آن می‌پردازیم. با استفاده از این رویکردها، امیدواریم که بهبودی قابل ملاحظه‌ای در دقت و کارایی مدل شناسایی افراد در تصاویر داشته باشیم و به عملکرد بهتری برای انواع مختلفی از سناریوهای [1] کاربردی دست یابیم.

۳ـ۱ مشکلات روش‌های قبلی

۳ـ۱ـ۱ کارایی نامناسب بر روی مجموعه‌داده با حجم کم

یکی از مشکلاتی که در این پروژه مطرح شده است، محدودیت تعداد داده‌های موجود در مجموعه‌داده است. تعداد کم داده‌ها می‌تواند باعث کاهش دقت و کارایی مدل‌های پیش‌بینی شده شود. این مشکل به دلیل این است که مدل‌ها عموماً نیاز به حجم بالایی از داده‌ها برای آموزش دارند تا بتوانند الگوهای معمول و پیچیده‌تری را که در داده‌ها وجود دارد، تشخیص دهند. اگر داده‌های آموزشی کم باشند، مدل ممکن است دقت کمتری در تشخیص و پیش‌بینی داشته باشد و نتواند عملکرد مطلوبی نداشته باشد.

[1] Scenario

با توجه به محدودیت تعداد داده‌ها، ممکن است مدل‌های آموزش دیده بر روی این مجموعه‌داده‌ها نتوانند الگوهای کلی و کارا برای شناسایی مجدد انسان‌ها را یاد بگیرند. این مشکل می‌تواند باعث شود که مدل‌ها به دقت کمتری در تشخیص اجسام و اشیا در تصاویر دست یابند و نتایج نهایی ناامیدکننده‌تر باشند.

۳ـ۱ـ۲ در نظر نگرفتن اطلاعات زمانی

اطلاعات معنایی و زمانی به تنهایی ممکن است کافی نباشند و در برخی موارد به خطاها منجر شوند، به خصوص در اینجا که مجموعه‌داده ما از دوربین‌های حراست بانوان دانشگاه علم و صنعت گرفته شده است. این داده‌ها اغلب از افرادی که لباس‌های مشابهی به تن دارند و ممکن است ویژگی‌های مشابهی داشته باشند، تشکیل شده است. به عنوان مثال، بیشتر بانوان ممکن است مقنعه داشته باشند یا برخی از آن‌ها چادر بپوشند. همچنین، اغلب از لباس‌های رنگ تیره استفاده می‌شود که می‌تواند به ایجاد اشتباه‌ها در شناسایی افراد منجر شود.

شکل ۳ـ۲: تصویر هویت 0000

شکل ۳ـ۳: تصویر هویت 0006

برای نمونه، همانطور که در تصویر ۳ـ۲ و ۳ـ۳ می‌بینید، به علت شباهت ظاهری شخص 0000 و 0006، مدل دچار خطا می‌شود و آن‌ها را به عنوان یک شخص تشخیص می‌دهد.

با این حال، با استفاده از مدل زمانی، ما می‌توانیم از اختلاف‌های زمانی بین ظاهر شدن اشخاص در دو

دوربین مختلف نیز استفاده کنیم. این اختلاف‌ها می‌تواند به ما کمک کند تا بهترین تطابق بین تصویر پرس‌وجو و تصاویر موجود در گالری را پیدا کنیم. به عبارت دیگر، با توجه به اینکه هر فرد در زمان‌های مختلف در دو دوربین ظاهر می‌شود، ما می‌توانیم بر اساس این اطلاعات زمانی، حدس بزنیم که هر فرد با کدام تصویر در گالری مطابقت دارد. این اطلاعات زمانی می‌تواند مشکلات احتمالی اطلاعات معنایی و ظاهری جبران کند و دقت و کارایی الگوریتم را بهبود بخشد.

۳ـ۲ ویژگی‌های الگوریتم ارائه شده

ویژگی‌های الگوریتم ارائه شده باید با رویکردی که به حل مشکلات مطرح شده مرتبط است، مطابقت داشته باشند. به همین دلیل، اصلی‌ترین توجه و تمرکز این الگوریتم باید بر روی دو عنصر اصلی باشد: مدل زمانی و ارزیابی مناسب بر روی مجموعه‌داده شخصی.

از آنجا که هدف اصلی این پروژه بهبود شناسایی مجدد انسان است، امتیاز حاصل از دو مدل زمانی و شناسایی مجدد انسان ترکیب شده و به عنوان امتیاز نهایی در نظر گرفته می‌شود. این ترکیب امتیازها به ما امکان می‌دهد تا یک ارزیابی جامع و کامل‌تر از عملکرد الگوریتم در اختیار داشته باشیم.

در این پروژه با توجه به نکات، مشکلات و اهداف مشخص شده، از مدل SOLIDER به عنوان مدل اصلی شناسایی مجدد انسان استفاده می‌شود. در ادامه، بخش‌های مختلف این الگوریتم با جزئیات بیشتری مورد بررسی و تحلیل قرار خواهند گرفت.

۳ـ۳ جزئیات الگوریتم

در بخش ۲ـ۶ـ۴ مدل SOLIDER به طور کامل بررسی شد. حال در این بخش، الگوریتم پیاده‌سازی شده برای مدل زمانی مورد بررسی قرار می‌گیرد و بخش‌های آن به صورت مجزا تفسیر می‌گردند.

۳ـ۳ـ۱ آماده‌سازی مجموعه‌داده

با توجه به اینکه در مدل زمانی، اولین زمان ظاهر شدن شخص در دوربین‌ها حائز اهمیت است، یک الگوریتم برای انتخاب این لحظه اولیه ظاهر شدن شخص در هر دوربین برای تمامی تصاویر پرس‌وجو تدوین شده است.

در این روش، تصاویر یک دوربین به عنوان گالری و تصاویر دوربین دیگر به عنوان پرس‌وجو در نظر گرفته می‌شود. در اینجا، تصاویر دوربین ۳ را به عنوان پرس‌وجو و دوربین ۱ را به عنوان گالری در نظر گرفته‌ایم. در این الگوریتم، به ازای هر هویت، دو تصویر از هر دوربین انتخاب می‌شود که یکی برای پرس‌وجو و دیگری برای گالری می‌باشد. این انتخاب بر اساس شماره قاب هر تصویر صورت می‌گیرد، به طوری که تصاویر با کمترین شماره قاب در هر دوربین برای هر شخص انتخاب می‌شوند. بنابراین ابعاد این مجموعه‌داده مطابق جدول زیر می‌شود.

جدول ۳ـ۱: مجموعه‌داده IUST برای مدل زمانی

subset	# ids	# images	# cameras
query	32	32	1
gallery	32	32	1

۳ـ۳ـ۲ محاسبه اختلاف زمانی تصاویر پرس‌وجو و گالری به ازای هر شخص

اختلاف زمانی بین هر دو قاب پرس‌وجو و گالری را برای هر شخص محاسبه کرده و یک مدل KDE (Kernel Density Estimation) بر روی آن اعمال می‌کنیم.

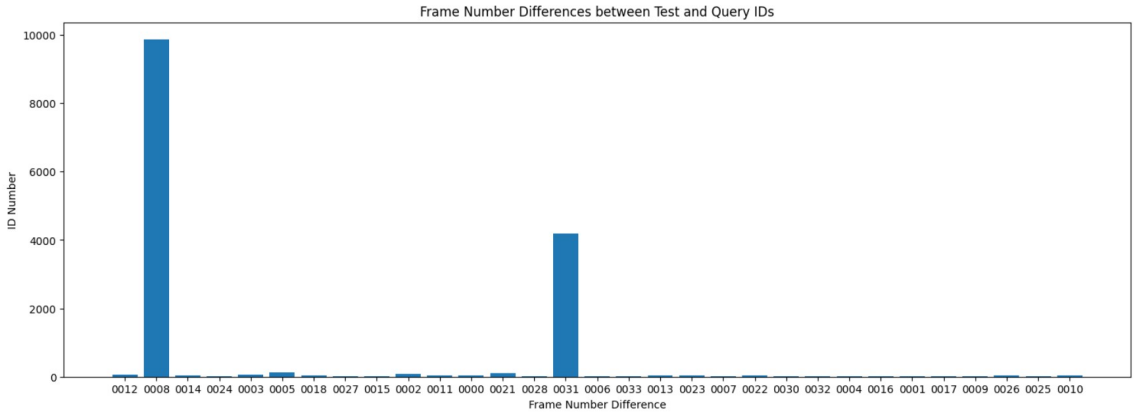
روش KDE یک روش غیرپارامتری برای تخمین توزیع احتمال یک متغیر تصادفی است. این روش برای تخمین توزیع احتمال از داده‌های مشاهده‌شده استفاده می‌کند و به طور خاص برای ارزیابی توزیع احتمال یک متغیر تصادفی، اغلب در زمینه‌های مختلفی مانند آمار، پردازش تصویر، مهندسی و غیره استفاده می‌شود.

در این روش، یک تابع هسته (مانند گاوسی [۲]) به هر نقطه از داده‌ها اختصاص داده می‌شود. سپس این توابع هسته به داده‌های مشاهده‌شده اعمال می‌شود و در نتیجه، توزیع چگالی احتمال مورد نظر را محاسبه می‌کند. تنظیم پارامترهای مختلف هم‌چون پهنای باند [۳] و نوع تابع هسته بر اساس خواسته‌ها و شرایط مساله می‌تواند تأثیر مستقیمی بر کیفیت و دقت تخمین‌ها داشته باشد.

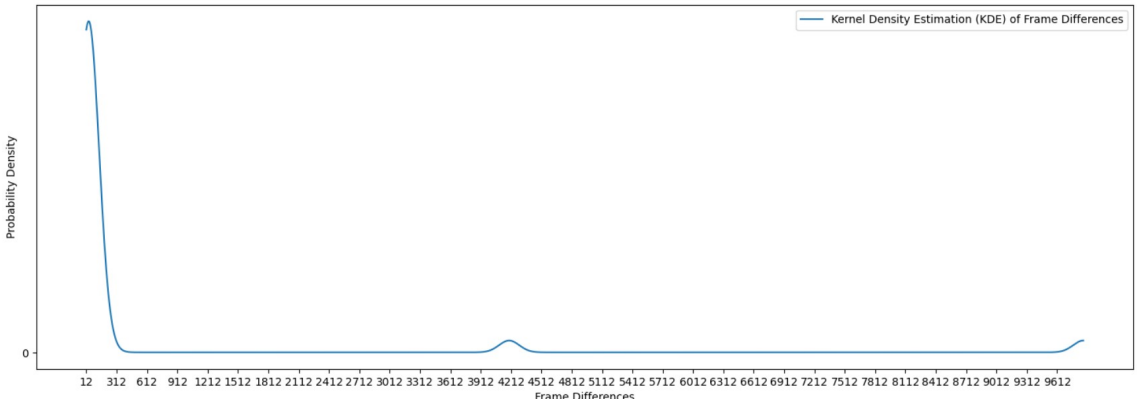
در اینجا، تابع گاوسی به عنوان تابع هسته و اندازه پهنای باند با استفاده از روش سعی و خطا، ۱۰۰ انتخاب شده است.

همانطور که در شکل ۳ـ۵ نیز مشخص است، اختلاف زمانی بین تصویر پرس‌وجو و گالری به ازای هر

[2] Gaussian
[3] Bandwidth



شکل ۳ـ۴: اختلاف زمانی تصاویر پرس‌وجو و گالری به ازای هر شخص



شکل ۳ـ۵: تخمین احتمال با استفاده از KDE بر روی اختلاف قاب‌ها

شخص، به طور میانگین حدودا ۳۰ می‌باشد. تنها دو تصویر خاص هستند که اختلاف زمانی طبیعی ندارند که یکی مربوط به شخص 0008 با اختلاف قاب ۹۸۷۰ و دیگری مربوط به شخص 0031 با اختلاف ۴۱۹۳ می‌باشد که در تصویر ۳ـ۴ به وضوح مقادیر اختلاف زمانی زیاد تصاویر پرس‌وجو و گالری این دو شخص با سایر افراد دیده می‌شوند. پس از بررسی تصاویر مربوط به این افراد، متوجه شدیم که علت اختلاف زیاد هویت 0008 به این علت است که شخص زمانی را در قسمت حراست می‌نشیند و سپس از حراست خارج می‌شود. در مورد شخص 0031 نیز، این هویت مربوط به کارمند حراست بانوان می‌باشد و مدتی مشغول آب دادن به گل‌ها در خارج قسمت حراست می‌باشد و سپس به اتاق حراست برمی‌گردد.

۳ـ۳ـ۳ محاسبه اختلاف قاب برای هر پرس‌وجو به ازای تمامی تصاویر گالری

در این بخش، اختلاف قاب هر پرس‌وجو با تمامی تصاویر گالری محاسبه می‌شود. این کار به این منظور صورت می‌گیرد که زمانی که یک پرس‌وجو وارد شبکه می‌شود، ما تصویر گالری متناظر با آن را نمی‌دانیم. بنابراین، اختلاف قاب هر پرس‌وجو با تک تک تصاویر گالری محاسبه کرده و هر یک را به KDE می‌دهیم تا احتمال متناسب با آن را تخمین بزند. این مقدار احتمال، امتیاز شباهت [۴] ما محسوب می‌شود. درنتیجه، هر یک از تصاویر گالری که امتیاز بیشتری داشت، به عنوان پیش‌بینی مدل KDE به عنوان گالری متناظر با پرس‌وجو داده شده، انتخاب می‌گردد.

۳ـ۳ـ۴ محاسبه Rank 1, Rank 5, Rank 10

محاسبه معیارهای Rank 1، Rank 5 و Rank 10 در مسائل بازیابی اطلاعات و ارزیابی عملکرد مدل‌های بازیابی اطلاعات مورد استفاده قرار می‌گیرد. این معیارها نشان دهندهٔ میزان موفقیت مدل در توانایی تشخیص تصویر شخص مورد نظر از میان گالری مرتب شده بر اساس امتیازهای شباهت به پرس‌وجو است. در معیار Rank 1، اگر نمونه مورد نظر اولین نمونه مرتب شده باشد، به عنوان یک موفقیت در نظر گرفته می‌شود. Rank 5 نشان می‌دهد که آیا شخص مورد نظر در بین پنج تصویر اول وجود دارد. Rank 10 نیز مشابه Rank 5 عمل می‌کند، اما در میان ده نمونه اول بررسی می‌شود.

ابتدا برای هر پرس‌وجو، امتیازهای شباهت به تصاویر گالری مرتب شده و سپس بر اساس این رتبه‌بندی، معیارهای Rank محاسبه می‌شوند. سپس نتایج به صورت درصد تجمعی گزارش می‌شوند.

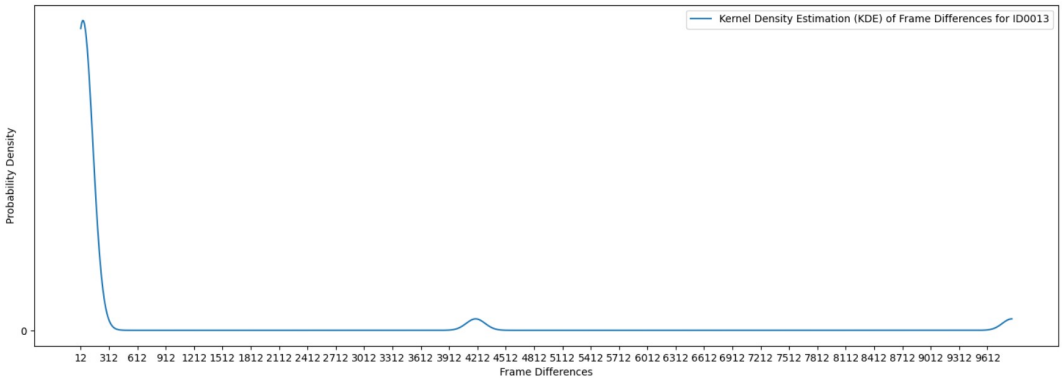
۳ـ۳ـ۵ روش Leave-One-Out

در این روش، به ازای هر پرس‌وجو، آن پرس‌وجو را از مجموعه‌داده حذف می‌کنیم. سپس، اختلاف قاب پرس‌وجوهای باقی‌مانده با گالری‌های متناظرشان را به دست آورده و یک KDE جدید روی آن اعمال می‌کنیم. سپس، اختلاف قاب پرس‌وجو حذف شده را با تک تک گالری‌ها محاسبه می‌کنیم. در ادامه، این اختلاف قاب‌ها را به KDE جدید داده و امتیاز هر تصویر گالری را طبق این KDE محاسبه می‌کنیم.

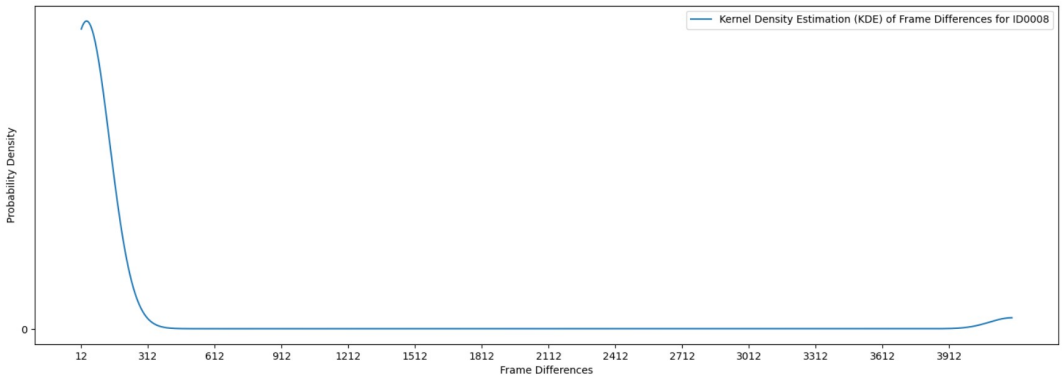
در این روش، اکثر KDEهای به دست آمده از حذف پرس‌وجو، نمایی مانند شکل ۳ـ۶ را دارند. تنها

[4] Similarity Score

تصاویر مربوط به اشخاص با هویت 0008 و 0031 با بقیه متفاوت است. همانطور که مشاهده می‌شود در شکل ۳ـ۷، KDE در نقطه ۴۱۹۳ دیگر مقدار قابل توجهی ندارد. این موضوع به این علت است که پرس‌وجو مربوط به هویت 0008 از مجموعه‌داده حذف شده است و در نتیجه، چنین اختلاف قابی به KDE اضافه نمی‌شود تا باعث برهم‌خوردن شکل KDE شود. همچنین، این موضوع در مورد هویت 0031 نیز صدق می‌کند.



شکل ۳ـ۶: محاسبه KDE بدون پرس‌وجو ۱۳



شکل ۳ـ۷: محاسبه KDE بدون پرس‌وجو ۸

شکل ۳ـ۸: محاسبه KDE بدون پرس‌وجو ۳۱

۳ـ۳ـ۶ ترکیب امتیاز مدل شناسایی مجدد انسان و مدل زمانی

در این بخش که بخش نهایی پروژه نیز هست، امتیازهای به دست آمده از مدل زمانی و مدل شناسایی مجدد طبق رابطه ۳ـ۱ ترکیب می‌شوند.

$$Score = \alpha Score_{ReId} + (1 - \alpha) Score_{Temporal} \quad (۱ـ۳)$$

مقدار α که بیانگر ضریبی برای تعیین میزان تاثیر مدل شناسایی مجدد و مدل زمانی است، باید یک عدد بین ۰ و ۱ باشد. براساس آزمایشاتی که توسط روش سعی و خطا انجام شده، مشخص شده است که مقدار ۰.۵ برای این ضریب بسیار مناسب است.

پس از اینکه امتیازها را ترکیب کردیم، امتیاز نهایی را به تابع محاسبه کننده Rank می‌دهیم تا روش خود را ارزیابی کنیم.

۳ـ۳ـ۷ هزینه محاسباتی

با توجه به اندازه کوچک مجموعه‌داده در اینجا با تنها ۳۲ هویت، هزینه محاسباتی پایینی خواهیم داشت. اما در صورتی که مجموعه‌داده اندازه بزرگتری داشته باشد، مثلاً مجموعه‌داده Market1501 با حدود ۱۵۰۱ هویت، هزینه محاسباتی به طور قابل توجهی افزایش خواهد یافت.

با در نظر گرفتن کد ارائه شده برای مدل زمانی و نوع عملیات‌هایی که در آن انجام می‌شود، می‌توان

گفت که سرعت اجرای الگوریتم‌ها به نسبت خوب است. زیرا الگوریتم‌های استفاده شده از نظر پیچیدگی محاسباتی نسبتاً ساده هستند و عملیات محاسباتی مورد نیاز برای هر هویت و تصویر به صورت مستقل و بدون وابستگی به سایر داده‌ها انجام می‌شود. همچنین، با توجه به اندازه کوچک مجموعه‌داده، تعداد این عملیات‌ها نیز محدود است که باعث افزایش سرعت اجرای کد می‌شود.

با این حال، در صورت افزایش اندازه مجموعه‌داده یا پیچیدگی الگوریتم‌ها، ممکن است سرعت اجرای کد کاهش یابد. به عبارت دیگر، با افزایش تعداد هویت‌ها یا تصاویر و افزایش پیچیدگی الگوریتم‌ها، هزینه محاسباتی و زمان اجرای کد افزایش می‌یابد. از این رو، اگر مجموعه‌داده اندازه بزرگتری داشته باشد یا الگوریتم‌ها پیچیده‌تر شوند، ممکن است سرعت اجرای کد کاهش یابد و هزینه محاسباتی به طور محسوسی افزایش یابد.

فصل ۴

نتایج و آزمایش‌ها

۴ـ۱ آماده سازی مجموعه‌داده IUST

این مجموعه‌داده شامل تصاویری است که از دوربین‌های حراست درب ۱ قسمت بانوان دانشگاه علم و صنعت به دست آمده است. دو دوربین به نزدیکی یکدیگر نصب شده‌اند و فعالیت‌های حرکت و ورود و خروج را کنترل می‌کنند. مجموعه‌داده اصلی ما شامل ۳۵ پوشه [1] است، هر کدام شامل تصاویری از یک شخص مشخص می‌باشد که توسط این دو دوربین، ثبت شده‌اند. برای سهولت کار، مجموعه‌داده را به سه بخش "آموزش"، "گالری" و "پرس‌وجو" تقسیم کرده‌ایم. سپس از دوربین ۱ برای بخش‌های آموزش و گالری و از دوربین ۳ برای بخش پرس‌وجو استفاده کرده‌ایم. در ادامه، با استفاده از یک الگوریتم، برای هر شخص، یک تصویر از دوربین ۳ به عنوان نمونه پرس‌وجو و دو تصویر متفاوت از دوربین ۱ به عنوان نمونه آموزش و گالری انتخاب کرده‌ایم. این الگوریتم به گونه‌ای عمل می‌کند که تصاویر وسط برای هر شخص را انتخاب می‌کند تا تصویر وضوح بیشتری داشته باشد؛ چرا که در قاب‌های ابتدایی شخص در حال وارد شدن است و تصاویر اشخاص کامل نیستند. برای نمونه در قاب‌های اولیه بعضی اشخاص، فقط صورت، پا یا بقیه اجزای بدن وجود دارند. بنابراین، به‌منظور انتخاب تصاویر بهتر برای مدل شناسایی افراد، از تصاویر وسط استفاده می‌کنیم که تقریبا کامل هستند.

علاوه بر این، برای انجام آزمایش‌های بیشتر و استفاده از روش Few-Shot Learning، یک مجموعه‌داده

[1] Folder

دیگر نیز تهیه شده است. تفاوت این مجموعه‌داده در تعداد تصاویر گالری و آموزش می‌باشد، به این صورت که برای هر شخص، چهار تصویر مختلف به عنوان آموزش و گالری انتخاب می‌شود.

روش انتخاب داده برای مدل زمانی نیز، به طور کامل در بخش ۳ـ۳ـ۱ توضیح داده شد.

۴ـ۱ـ۱ نتایج مدل شناسایی مجدد انسان

نتایجی که در مقاله SOLIDER ذکر شده است، بیانگر عملکرد بهتر نسبت به مدل‌های به‌روز می‌باشد. این برتری به علت ترکیب روش‌های برتر شناسایی شده برای شناسایی مجدد می‌باشد. برای نمونه، این روش از خوشه‌بندی معنایی (حذف بخش پس‌زمینه به علت داده‌های کم‌ارزش)، استفاده از مدل‌سازی ماسک شده و یک کنترل‌گر برای تعیین میزان تاثیر اطلاعات معنایی و ظاهری می‌باشد. همانطور که در جدول ۴ـ۱ مشاهده می‌نمایید، عدد اول نشان‌دهنده دقت mAP و عدد دوم Rank 1 را نشان می‌دهد. نتایج نمایش داده شده نیز، بیانگر عملکرد بهتر SOLIDER نسبت به سایر روش‌های به‌روز شناسایی مجدد است.

جدول ۴ـ۱: نتایج به دست آمده برای مدل شناسایی مجدد

Dataset	SCSN	ABDNet	TranReID	UP-ReID	PASS	SOLIDER
Market1501	88.5/95.7	88.3/95.6	89.5/95.2	91.1/97.1	93.3/96.9	**93.9/96.9**
MSMT17	58.5/83.8	60.8/82.3	69.4/86.2	63.3/84.3	74.3/89.7	**77.1/90.7**

حال، اگر بخواهیم نتایج این مدل بر روی مجموعه‌داده IUST را به دست بیاوریم، متوجه می‌شویم که عملکرد این مدل به شدت کاهش پیدا می‌کند (جدول ۴ـ۲). از دلایل این کاهش، می‌توان به تعداد بسیار کم هویت‌های مستقل IUST (حدودا ۳۴ نفر) نسبت به مجموعه‌داده Market (۱۵۰۱ هویت مستقل) و همچنین داده‌های سخت اشاره کرد. برای نمونه در این مجموعه‌داده، اکثر افراد مقنعه و لباس‌های تیره بر تن دارند که این موضوع، باعث شباهت داده‌ها به یکدیگر و سخت شدن قدرت تشخیص برای مدل می‌شد، در صورتی که در مجموعه‌داده Market1501، این مشکل وجود ندارد و اشخاص از لحاظ ظاهری متفاوت هستند. به دلیل مشکلات گفته شده، مدل SOLIDER را روی مجموعه‌داده Market1501 در ۱۲۰ دوره آموزش دادیم. سپس، مدل را بر روی مجموعه‌داده IUST آزمایش کردیم. برای آزمایش داده‌ها، از دو روش One-Shot Learning و Few-Shot Learning استفاده کردیم.

در روش Few-Shot Learning، ما می‌توانیم دو رویه را در پیش بگیریم. به این صورت که زمانی که ۴

تصویر گالری به ازای یک پرس‌وجو داریم، بین امتیازهای گالری‌های متناظر با پرس‌وجو میانگین بگیریم (۱)، یا اینکه بیشترین امتیاز را انتخاب کنیم (۲).

جدول ۴ـ۲: نتایج به دست آمده برای مدل شناسایی مجدد بر روی مجموعه‌داده IUST

Method	# Query	# Gallery	mAP	Rank 1	Rank 5	Rank 10
One-Shot Learning	32	34	29.8%	12.9%	45.2%	71.0%
Few-Shot Learning (1)	32	136	27.9%	19.4%	54.8%	74.2%
Few-Shot Learning (2)	32	136	30.1%	21.6%	57.3%	75.6%

همانطور که در جدول ۴ـ۲ نیز نشان داده شده است، دقت در روش‌های Few-Shot بهتر شده است. این بهبود ناشی از افزایش تعداد نمونه‌ها در تصاویر گالری است. بنابراین، اطلاعات بیشتری را در اختیار دارد و این باعث می‌شود که احتمالاً شباهت‌های بیشتری بین تصاویر پرس‌وجو و گالری وجود داشته باشد. در روش دوم که از روش انتخاب بالاترین امتیاز به عنوان امتیاز شباهت نهایی تصویر پرس‌وجو و گالری استفاده شده است، دقت بالاتر رفته است. این موضوع، می‌تواند به این علت باشد که در این روش، بهترین شباهت بین تصویر پرس‌وجو و تصاویر گالری در نظر گرفته می‌شود، بنابراین این انتخاب می‌تواند باعث بهبود در دقت شود، زیرا تصویری با شباهت بیشتر از دیگر تصاویر، به عنوان شباهت نهایی انتخاب می‌شود؛ اما زمانی که میانگین می‌گیریم، ممکن است یکی از تصاویر گالری مربوط به یک شخص، بسیار با تصویر پرس‌وجو متناظرش متفاوت باشد و باعث امتیاز پایین شود. در نتیجه، اگر در اینجا میانگین بگیریم، این داده با امتیاز پایین، باعث پایین آمدن امتیاز نهایی بشود.

۴ـ۱ـ۲ نتایج مدل زمانی

همانطور که در بخش‌های قبل گفته شد، در مدل زمانی نیز، ما دو حالت در نظر گرفتیم. روش اول مدل زمانی را با تمامی تصاویر گالری و پرس‌وجو به دست آوردیم. در روش دوم، از روش Leave-One-Out استفاده کردیم. به این صورت که، به ازای هر پرس‌وجو، آن پرس‌وجو را حذف کرده و با استفاده از سایر پرس‌وجوها و گالری متناظرشان، یک مدل زمانی جدید به دست آوردیم. سپس، اختلاف قاب پرس‌وجو حذف شده با تک تک تصاویر گالری را به دست آورده و امتیازشان را با استفاده از مدل زمانی که به دست آوردیم، محاسبه می‌کنیم. این کار را برای تمامی پرس‌وجوها انجام می‌دهیم.

در جدول ۴ـ۳ نیز نتایج حاصل شده، نمایش داده شده است. همانطور که مشاهده می‌شود، میزان Rank 1 در روش اول، 90% می‌باشد و Rank 5 و Rank 10 نیز، برابر با 100% هستند. علت کم شدن دقت در Rank 1 نیز، همان دو نمونه استثنا (شخص 0008 و شخص 0031) می‌باشد. اگر این دو نمونه نیز رفتار طبیعی داشتند، دقت Rank 1 به 100% می‌رسید.

جدول ۴ـ۳: نتایج به دست آمده برای مدل زمانی

Method	# Query	# Gallery	Rank 1	Rank 5	Rank 10
Simple Temporal Model	32	32	90.3%	100%	100%
Leave-One-Out Temporal Model	32	32	83.9%	93.5%	100%

روش Leave-One-Out باعث می‌شود که مدل‌ها بیشتر به نمونه‌های دیده‌نشده و کمتر به نمونه‌های مشابه دیده‌شده توجه کنند، که ممکن است باعث افزایش دقت بر روی داده‌های جدید شود. اما به دلیل حذف نمونه‌ها از داده‌های آموزشی، این روش ممکن است باعث افت دقت شود، زیرا اطلاعات مهمی که ممکن است در نمونه‌های حذف شده وجود داشته باشد، دیگر در آموزش مدل در نظر گرفته نمی‌شود. به عبارت دیگر، این روش ممکن است باعث از دست رفتن اطلاعات مفید و ارزشمندی شود که مدل برای تمیزدهی بهتر و دقیق‌تر نیاز دارد، که این باعث افت دقت مدل در ارزیابی‌های عملکرد می‌شود. بنابراین، اگر از این مدل برای آزمون داده‌های جدید استفاده کنیم، احتمالا دقت بالاتری به‌دست می‌آوریم. چرا که با استفاده از این روش، قابلیت تعمیم‌دهی [۲] آن افزایش یافته است.

۴ـ۱ـ۳ نتایج ترکیب مدل زمانی و شناسایی مجدد

ترکیب مدل شناسایی مجدد و مدل زمانی می‌تواند منجر به افزایش دقت نسبت به هر مدل به تنهایی شود زیرا هر کدام از این دو روش قابلیت‌ها و مزایای منحصربه‌فرد خود را دارند که می‌توانند با ترکیب آنها بهبود عملکرد مدل نهایی را فراهم کنند.

مدل شناسایی مجدد معمولاً برای استخراج ویژگی‌های ظاهری و معنایی از تصاویر و تشخیص افراد استفاده می‌شود که می‌تواند باعث افزایش دقت در تشخیص افراد با شباهت بالا شود. از طرف دیگر، مدل زمانی برای مدل‌سازی روابط زمانی بین داده‌ها و پیش‌بینی زمان ظاهر شدن فرد در دوربین دیگر استفاده می‌شود،

[2] Generalization

که می‌تواند باعث افزایش دقت در شناسایی افراد شود.

از ترکیب این دو مدل، اطلاعات و ویژگی‌های مختلف از دو جنبه مختلف (فضای ویژگی و زمان) در نظر گرفته می‌شوند که این امکان را فراهم می‌کند که مدل نهایی از تمامی این اطلاعات بهره‌مند شده و بتواند به صورت جامع‌تر و دقیق‌تر از داده‌ها استفاده کند. به عبارت دیگر، ترکیب این دو مدل باعث افزایش چندپارامتری و پیچیدگی مدل می‌شود که باعث افزایش دقت و کارایی مدل در تشخیص و پیش‌بینی افراد و رخدادها می‌شود.

همانطور که در جدول ۴ـ۴ نیز مشخص است، با ترکیب امتیاز دو مدل، دقت حاصل شده از دقت به دست آمده از هر دو مدل بهتر است. این نشان می‌دهد با ترکیب اطلاعات معنایی و زمانی، دقت مدل شناسایی مجدد بر روی مجموعه‌داده IUST به طرز چشمگیری افزایش یافت. همچنین، اطلاعات معنایی نیز باعث بهتر شدن نتایج مدل زمانی شدند.

جدول ۴ـ۴: نتایج به دست آمده برای ترکیب مدل شناسایی مجدد و مدل زمانی

Method	Rank 1	Rank 5	Rank 10
Leave-One-Out Temporal + One-Shot Learning ReID	87.1%	93.5%	93.5%
Leave-One-Out Temporal + Few-Shot Learning ReID (1)	87.1%	93.5%	96.8%
Leave-One-Out Temporal + Few-Shot Learning ReID (2)	87.1%	93.5%	96.8%

فصل ۵

جمع‌بندی و پیشنهادها

۵ـ۱ جمع‌بندی

در این پروژه، از روش و الگوریتم SOLIDER به عنوان یک روش خودنظارتی مبتنی بر کنترل‌گر معنایی استفاده کردیم تا تصاویر افراد را شناسایی کنیم. آزمایش‌های انجام شده نشان دادند که الگوریتم ارائه شده عملکرد بسیار بهتری نسبت به الگوریتم‌های نظارت‌شده و سایر الگوریتم‌های قبلی دارد. این آزمایش‌ها بر روی مجموعه‌داده‌های Market1501 و MSMT در حوزه شناسایی مجدد انسان صورت گرفت و در تمامی آزمایش‌ها، نتایج به دست آمده از روش‌های پیشین بهتر بود. اما با توجه به مشکل کم بودن تعداد هویت‌ها و سختی داده‌ها در مجموعه‌داده، نتوانستیم نتایج دلخواه را به دست بیاوریم. بنابراین، تصمیم گرفتیم از یک مدل زمانی استفاده کنیم تا این مدل را بر روی مجموعه‌داده IUST نیز بهبود بخشیم و عملکرد بهتری حاصل کنیم.

در ادامه، از تجربیات به دست آمده در این پروژه، می‌توان بهبودهایی را در زمینه‌های مختلف مدل‌های شناسایی مجدد و مدل‌های زمانی پیشنهاد داد. همچنین، مطالعات بیشتر در مورد تأثیرات مدل‌های زمانی بر عملکرد مدل‌های شناسایی مجدد و برعکس، می‌تواند به روش‌های بهتری برای ترکیب این دو نوع مدل و بهبود عملکرد کلی سیستم منجر شود. در ادامه، به کارهایی که می‌تواند در آینده برای بهبود بیشتر این مدل ارائه شود، می‌پردازیم.

۵ـ۲ پیشنهادها و کارهای آینده

۵ـ۲ـ۱ استفاده از سن اشخاص

یکی از کارهای آینده که می‌تواند بهبود عملکرد مدل‌های زمانی را تضمین کند، استفاده از ویژگی‌های جانبی افراد مانند سن، جنسیت، وضعیت فیزیکی و... است. به عنوان مثال، افراد سالمند عموماً حرکت کندتری دارند نسبت به افراد جوان، در حالی که افراد جوان ممکن است سریع‌تر و با حرکت‌های پویا عمل کنند. اگر بتوانیم افراد را بر اساس سن یا ویژگی‌های دیگر دسته‌بندی کنیم، می‌توانیم مدل‌های زمانی را به گونه‌ای طراحی کنیم که این اطلاعات جانبی را در نظر بگیرند و بر اساس آن‌ها پیش‌بینی دقیق‌تری انجام دهند. این اقدام می‌تواند بهبود قابل توجهی در عملکرد مدل‌های زمانی و بهبود دقت در تخمین زمان ظاهر شدن افراد در تصاویر داشته باشد.

۵ـ۲ـ۲ استفاده از ساعت روز

در این پروژه، تاثیر سرعت حرکت افراد بر روی دقت مدل ارزیابی نشد. بنابراین، یکی از کارهایی که در آینده می‌توان انجام داد، بهره‌گیری از اطلاعات زمانی به شکل ساعت روز نیز می‌تواند باشد. به عنوان مثال، در زمان‌های مختلف روز، افراد ممکن است رفتارها و عادات مختلفی داشته باشند. صبح‌ها افراد بیشتر به دلیل شروع یک روز کاری، با انرژی بیشتری عمل می‌کنند و ممکن است در حال عجله و شتاب باشند. در مقابل، بعد از ظهرها که افراد از کار یا تحصیل خود تعطیل می‌شوند، انرژی آن‌ها کمتر می‌شود و ممکن است حرکت‌هایشان کندتر باشد.

با در نظر گرفتن این تفاوت‌ها در رفتارهای زمانی، می‌توانیم برای هر بازه زمانی مدل‌های جداگانه‌ای را طراحی کنیم. به عنوان مثال، مدل زمانی صبحگاهی و مدل زمانی شبانه را جداگانه در نظر بگیریم. این اقدام می‌تواند به دقت بیشتر و بهبود کارایی مدل‌ها در تخمین زمان ظاهر شدن افراد کمک کند. به این ترتیب، می‌توانیم با تنظیم پارامترها و ویژگی‌های مدل بر اساس زمان روز، عملکرد مدل‌ها را بهبود بخشیم و نتایج دقیق‌تری در شناسایی مجدد انسان به دست آوریم.

همچنین، با در نظر گرفتن ساعت تردد افراد، می‌توانیم الگوهایی را در رفتارهای زمانی آن‌ها شناسایی کنیم. به عنوان مثال، فرض کنید یک شخص معمولاً در ساعت ۸ صبح به محل کارش می‌رود. با در نظر

گرفتن این موضوع، می‌توانیم تصاویری که در این ساعت گرفته می‌شوند را با امتیاز بیشتری در مدل زمانی ما در نظر بگیریم. این به ما اجازه می‌دهد که از داده‌هایی که بازه زمانی مشخصی را پوشش می‌دهند، بهره‌مند شویم و در نتیجه، دقت و عملکرد مدل‌هایمان در شناسایی مجدد انسان را بهبود بخشیم.

۵ـ۲ـ۳ استفاده از شرایط آب و هوا

با استفاده از پس‌زمینه تصویر نیز می‌توانیم از شرایط آب و هوا بهره‌بری کنیم. به عنوان مثال، در روزهای بارانی، پس‌زمینه‌های تصاویر معمولاً شامل مناظری مرتبط با باران می‌شوند، مانند خیابان‌های خیس و قطره‌های باران. با در نظر گرفتن این پس‌زمینه‌ها، می‌توانیم تصاویر را به دسته‌بندی‌های مختلفی تقسیم کنیم و به هر دسته امتیاز مناسبی اختصاص دهیم. به عنوان مثال، تصاویر با پس‌زمینه بارانی ممکن است از اهمیت بیشتری برخوردار باشند، زیرا در شرایطی مشابه با شرایط تصویربرداری، مدل می‌تواند بهترین عملکرد را ارائه دهد. این اقدام می‌تواند به بهبود دقت و کارایی مدل زمانی در شناسایی مجدد انسان کمک کند، زیرا مدل می‌تواند با دقت بیشتری به شرایط محیطی تصاویر واکنش نشان دهد و عملکرد بهتری ارائه کند.

۵ـ۲ـ۴ استفاده از روش یادگیری افزایشی

استفاده از یادگیری افزایشی [1] می‌تواند بهبود عملکرد مدل زمانی را تضمین کند. در این روش، مدل‌ها به طور پیوسته با داده‌های جدید آموزش داده می‌شوند، به گونه‌ای که با افزایش تعداد داده‌ها، عملکرد آن‌ها بهبود می‌یابد. این رویکرد می‌تواند برای بهبود مدل زمانی بسیار موثر باشد، زیرا اجازه می‌دهد تا مدل با تجربه بیشتری از محیط و شرایط مختلف برخورد کند و عملکرد بهتری ارائه دهد.

به عنوان مثال، اگر ما از یک مدل یادگیری افزایشی استفاده کنیم، می‌توانیم مدل را به طور پیوسته با داده‌های جدیدی که در طول زمان جمع‌آوری می‌شوند، آموزش دهیم. این امر به ما این امکان را می‌دهد که مدل بهبود یابد و با توجه به تغییرات در محیط، به دقت و عملکرد بهتری برسد. به عنوان مثال، اگر شرایط آب و هوایی یا الگوهای رفتاری افراد با گذر زمان تغییر کنند، مدل با استفاده از یادگیری افزایشی می‌تواند با این تغییرات همگام شود و عملکرد مناسبی ارائه دهد. به این ترتیب، می‌توانیم با استفاده از این رویکرد، مدل زمانی را بهبود بخشیم و به دقت و کارایی بیشتری در شناسایی مجدد انسان دست پیدا کنیم.

[1] Incremental Learning

مراجع

[1] Wang, Xiao, Zheng, Shaofei, Yang, Rui, Zheng, Aihua, Chen, Zhe, Tang, Jin, and Luo, Bin. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022.

[2] Huang, Qingqiu, Liu, Wentao, and Lin, Dahua. Person search in videos with one portrait through visual and temporal links. in *Proceedings of the European conference on computer vision (ECCV)*, pp. 425–441, 2018.

[3] Zhao, Jian, Li, Jianshu, Nie, Xuecheng, Zhao, Fang, Chen, Yunpeng, Wang, Zhecan, Feng, Jiashi, and Yan, Shuicheng. Self-supervised neural aggregation networks for human parsing. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7–15, 2017.

[4] Jaiswal, Ashish, Babu, Ashwin Ramesh, Zadeh, Mohammad Zaki, Banerjee, Debapriya, and Makedon, Fillia. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[5] Caron, Mathilde, Touvron, Hugo, Misra, Ishan, Jégou, Hervé, Mairal, Julien, Bojanowski, Piotr, and Joulin, Armand. Emerging properties in self-supervised vision transformers. in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

[6] Bao, Hangbo, Dong, Li, Piao, Songhao, and Wei, Furu. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[7] Xie, Zhenda, Zhang, Zheng, Cao, Yue, Lin, Yutong, Bao, Jianmin, Yao, Zhuliang, Dai, Qi, and Hu, Han. Simmim: A simple framework for masked image modeling. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

[8] Chen, Weihua, Xu, Xianzhe, Jia, Jian, Luo, Hao, Wang, Yaohua, Wang, Fan, Jin, Rong, and Sun, Xiuyu. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15050–15061, 2023.

[9] Saber, Shimaa, Meshoul, Souham, Amin, Khalid, Pławiak, Paweł, and Hammad, Mohamed. A multi-attention approach for person re-identification using deep learning. *Sensors*, 23(7):3678, 2023.

[10] Zheng, Zhedong, Wang, Xiaohan, Zheng, Nenggan, and Yang, Yi. Parameter-efficient person re-identification in the 3d space. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[11] Zheng, Zhedong, Yang, Xiaodong, Yu, Zhiding, Zheng, Liang, Yang, Yi, and Kautz, Jan. Joint discriminative and generative learning for person re-identification. in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2138–2147, 2019.

[12] Li, Minxian, Zhu, Xiatian, and Gong, Shaogang. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1770–1782, 2019.

[13] Fu, Dengpan, Chen, Dongdong, Bao, Jianmin, Yang, Hao, Yuan, Lu, Zhang, Lei, Li, Houqiang, and Chen, Dong. Unsupervised pre-training for person re-identification. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14750–14759, 2021.

[14] Tang, Chufeng, Sheng, Lu, Zhang, Zhaoxiang, and Hu, Xiaolin. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4997–5006, 2019.

[15] Xia, Jiangyue, Rao, Anyi, Huang, Qingqiu, Xu, Linning, Wen, Jiangtao, and Lin, Dahua. Online multi-modal person search in videos. in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 174–190. Springer, 2020.

[16] Khan, Abdul Hannan, Nawaz, Mohammed Shariq, and Dengel, Andreas. Localized semantic feature mixers for efficient pedestrian detection in autonomous driving. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5476–5485, 2023.

[17] Lin, Zebin, Pei, Wenjie, Chen, Fanglin, Zhang, David, and Lu, Guangming. Pedestrian detection by exemplar-guided contrastive learning. *IEEE transactions on image processing*, 32:2003–2016, 2022.

[18] Li, Rui and Zu, Yaxin. Research on pedestrian detection based on the multi-scale and feature-enhancement model. *Information*, 14(2):123, 2023.

[19] Zhang, Xiaomei, Zhu, Xiangyu, Tang, Ming, and Lei, Zhen. Deep learning for human parsing: A survey. *arXiv preprint arXiv:2301.12416*, 2023.

[20] Zhang, Zihao, Hu, Lei, Deng, Xiaoming, and Xia, Shihong. Sequential 3d human pose estimation using adaptive point cloud sampling strategy. in *IJCAI*, pp. 1330–1337, 2021.

[21] Chen, Xinlei, Fan, Haoqi, Girshick, Ross, and He, Kaiming. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[22] Fang, Hongchao and Xie, Pengtao. An end-to-end contrastive self-supervised learning framework for language understanding. *Transactions of the Association for Computational Linguistics*, 10:1324–1340, 2022.

[23] Chen, Ting, Kornblith, Simon, Norouzi, Mohammad, and Hinton, Geoffrey. A simple framework for contrastive learning of visual representations. in *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[24] Yokoo, Shuhei. Contrastive learning with large memory bank and negative embedding subtraction for accurate copy detection. *arXiv preprint arXiv:2112.04323*, 2021.

[25] He, Kaiming, Fan, Haoqi, Wu, Yuxin, Xie, Saining, and Girshick, Ross. Momentum contrast for unsupervised visual representation learning. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[26] Caron, Mathilde, Misra, Ishan, Mairal, Julien, Goyal, Priya, Bojanowski, Piotr, and Joulin, Armand. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[27] Grill, Jean-Bastien, Strub, Florian, Altché, Florent, Tallec, Corentin, Richemond, Pierre, Buchatskaya, Elena, Doersch, Carl, Avila Pires, Bernardo, Guo, Zhaohan, Gheshlaghi Azar, Mohammad, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[28] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Zhu, Kuan, Guo, Haiyun, Liu, Zhiwei, Tang, Ming, and Wang, Jinqiao. Identity-guided human semantic parsing for person re-identification. in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 346–363. Springer, 2020.

[30] Mauthner, Thomas, Possegger, Horst, Waltner, Georg, and Bischof, Horst. Encoding based saliency detection for videos and images. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2494–2502, 2015.

[31] Hartigan, John A and Wong, Manchek A. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[32] Fu, Dengpan, Chen, Dongdong, Yang, Hao, Bao, Jianmin, Yuan, Lu, Zhang, Lei, Li, Houqiang, Wen, Fang, and Chen, Dong. Large-scale pre-training for person re-identification with noisy labels. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2476–2486, 2022.

Abstract:

Person re-identification is a critical issue in computer vision, where the goal is to locate a specific person in images captured from multiple camera views. With the continuous advancement of deep learning techniques, particularly convolutional neural networks, this field has become vital in surveillance and security domains. The primary motivation behind this topic is the need for effective tracking of individuals in crowded and complex environments such as airports and train stations. These studies aim to optimize human resource utilization and locate specific individuals, like lost children at airports, from thousands of candidate images. However, challenges such as intra-class variations in images and the inaccuracies of current models persist.

Overall, person re-identification is highly challenging due to significant intra-class variations in images. Most current methods, especially deep learning models, rely on supervised learning, requiring a large number of labeled images of each person from different cameras. Therefore, the recent focus has shifted towards unsupervised learning for person re-identification. However, unsupervised learning models often exhibit lower accuracy compared to supervised ones. The main objective of this research is to propose a temporal model that matches individuals between two cameras based on the temporal difference in their appearances. This approach enhances identification accuracy and considers additional features beyond visual and semantic aspects, leading to reduced errors and improved overall performance of person re-identification systems.

Keywords: Person re-identification, Computer vision, Deep learning, Convolutional neural networks, Tracking, Unsupervised learning, Temporal model

Iran University of Science and Technology
Computer Engineering Department

Developing an Intelligent Image Labeling Tool for Human Recognition Applications

Bachelor of Science Thesis in Computer Engineering

By:

Elnaz Rezaee

Supervisor:

Dr. Mohammadreza Mohammadi

February 2024