دانشگاه تهران
پردیس فارابی
دانشکده مهندسی
گروه مهندسی کامپیوتر

ارائه‌ی مدلی تعمیم‌پذیر برای بازشناسایی شخص با رویکرد یادگیری عمیق

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشتهٔ مهندسی فناوری اطلاعات
گرایش فناوری اطلاعات

صبا سادات فقیه‌ایمانی

استاد راهنما

دکتر کاظم فولادی قلعه

استاد مشاور

دکتر حسین آقابابا

شهریور ۱۳۹۹

بسم الله الرحمن الرحیم

دانشگاه تهران
پردیس فارابی
دانشکده مهندسی
گروه مهندسی کامپیوتر

ارائه‌ی مدلی تعمیم‌پذیر برای بازشناسایی شخص با رویکرد یادگیری عمیق

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشتهٔ مهندسی فناوری اطلاعات
گرایش فناوری اطلاعات

صبا سادات فقیه‌ایمانی

استاد راهنما

دکتر کاظم فولادی قلعه

استاد مشاور

دکتر حسین آقابابا

شهریور ۱۳۹۹

تعهدنامهٔ اصالت اثر

باسمه تعالی

اینجانب صبا سادات فقیه‌ایمانی تأیید می‌کنم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آن‌ها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتری ارائه نشده است.

نام و نام خانوادگی دانشجو: صبا سادات فقیه‌ایمانی

تاریخ و امضای دانشجو:

کلیهٔ حقوق مادی و معنوی این اثر
متعلّق به دانشگاه تهران است.

قدردانی

از خدای مهربانم متشکرم که همواره در مسیر زندگی دست مرا گرفته است.

از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای دکتر فولادی، صمیمانه تشکر و قدردانی می‌کنم که در طول انجام این پایان‌نامه همواره راهنما و مشوق من بودند و بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید. از جناب آقای دکتر آقابابا که زحمت مشاوره‌ی این پایان‌نامه را تقبل فرمودند متشکرم.

از خانواده‌ی عزیزم که همیشه مشوق من هستند و با نهایت صبوری در کنارم بودند کمال تشکر را دارم.

صبا سادات فقیه‌ایمانی

شهریور ۱۳۹۹

چکیده

مسئله‌ی بازشناسایی شخص شامل بازیابی تصاویر یک فرد در میان تصاویر جمع‌آوری شده توسط مجموعه‌ای از دوربین‌های غیرهم‌پوشان می‌باشد. تصاویر مجموعه‌داده‌های بازشناسایی شخص توسط دوربین‌های امنیتی نظارتی جمع‌آوری شده‌اند و چالش‌های گوناگونی دارند. در سال‌های اخیر با استفاده از تکنیک‌های یادگیری عمیق، نتایج موفقیت‌آمیزی در حوزه‌ی بازشناسایی شخص به‌دست آمده است. باوجود نتایج موفق در این حوزه، هنگام آزمایش مدل روی مجموعه‌داده‌های بدون برچسب متفاوت با مجموعه‌داده‌های آموزشی برچسب‌گذاری شده، عملکرد مدل به شدت کاهش می‌یابد. برای حل این مشکل می‌توان از وفق‌دهی دامنه‌ی بدون نظارت استفاده کرد. در این پایان‌نامه، مدلی با تعمیم‌پذیری بالا برای وفق‌دهی دامنه‌ی بدون نظارت در مسئله‌ی بازشناسایی شخص ارائه شده است. در این مدل از مجموعه‌داده‌ی دارای برچسب منبع و مجموعه‌داده‌ی بدون برچسب هدف برای آموزش مدل استفاده شده است و مدل باید در هنگام آزمایش روی دامنه‌ی هدف عملکرد مناسبی داشته باشد. به‌منظور مقاوم‌سازی مدل نسبت به تغییرات در دامنه‌ی هدف، طی فرآیند آموزش سه ویژگی درون-دامنه‌ای در دامنه‌ی هدف بررسی می‌شوند. یادگیری سه ویژگی تغییرناپذیری نسبت به نمونه‌ها، تغییرناپذیری نسبت به دوربین‌ها و تغییرناپذیری نسبت به همسایه‌ها، تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف را تشکیل می‌دهد. برای شیوه‌ی انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها ، دو استراتژی مورد آزمایش قرار گرفته است. در استراتژی اول انتخاب همسایه‌ها، برای همه‌ی نمونه‌های دامنه‌ی هدف تعداد برابری همسایه انتخاب می‌شود. درحالیکه در استراتژی دوم انتخاب همسایه‌ها، یک مقدار آستانه تعیین می‌شود و تصاویری از دامنه‌ی هدف که میزان شباهت آن‌ها به یک نمونه‌ی دامنه‌ی هدف از مقدار آستانه بیشتر باشد به عنوان همسایه‌های آن نمونه انتخاب می‌شوند. بنابراین تعداد همسایه‌های تصاویر دامنه‌ی هدف لزوماً برابر نمی‌باشد. علاوه‌بر تابع اتلاف طبقه‌بندی داده‌های دامنه‌ی منبع و تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، یک تابع اتلاف سه‌گانه نیز در آموزش مدل به کار رفته است. این تابع اتلاف سه‌گانه علاوه‌بر تفاوت‌های درون-دامنه‌ای دامنه‌ی هدف، تفاوت‌های بین دامنه‌های منبع و هدف را نیز درنظر می‌گیرد. برای استخراج ویژگی، از شبکه‌ی ResNeXt-50 استفاده شده است که نسبت به شبکه‌ی ResNet-50 تعداد پارامترهای کمتر و عملکرد بهتری دارد. مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها توانسته است در تنظیمات duke→market در رتبه‌ی ۱ و ۵ معیار CMC مقدار ۷۷٫۴ درصد و ۸۹٫۱ درصد و مقدار mAP ۴۶ درصد را به‌دست آورد. همچنین در تنظیمات market → duke در رتبه‌ی ۱ و ۵ معیار CMC مقدار ۶۴ درصد و ۷۶٫۸ درصد و مقدار mAP

۴۱٫۱ درصد را به‌دست آورده است. مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها توانسته است در تنظیمات duke→market در رتبه‌ی ۱ و ۵ معیار CMC مقدار۸۴٫۵ درصد و ۹۳٫۱ درصد و مقدار mAP ۶۳ درصد را به‌دست آورد. همچنین در تنظیمات market → duke در رتبه‌ی ۱ و ۵ معیار CMC مقدار ۷۰٫۱ درصد و ۸۰٫۸ درصد و مقدار mAP ۴۹٫۱ درصد را به‌دست آورده است.

واژگان کلیدی بازشناسایی شخص، یادگیری عمیق، شبکه عصبی کانوولوشنال، وفق‌دهی دامنه، مدل تعمیم‌پذیر

فهرست مطالب

فهرست شکل‌ها ت

فهرست جدول‌ها ج

فهرست الگوریتم‌ها چ

فصل ۱: مقدمه ۱

۱.۱ پیش‌گفتار ۱

۲.۱ تعریف بازشناسایی شخص ۲

۳.۱ چالش‌های بازشناسایی شخص ۴

۴.۱ بازشناسایی شخص جهان-بسته و جهان-باز ۷

۵.۱ اهداف پایان‌نامه ۹

۶.۱ نوآوری‌های پایان‌نامه ۱۰

۷.۱ ساختار پایان‌نامه ۱۲

فصل ۲: مروری بر مطالعات انجام شده ۱۳

۱.۲ پیش‌گفتار ۱۳

۲.۲ یادگیری عمیق در بینایی ماشین ۱۳

۱.۲.۲ انواع مدل‌های یادگیری عمیق ۱۶

۱.۱.۲.۲ شبکه عصبی کانوولوشنال ۱۶

۲.۱.۲.۲ شبکه‌ی باور عمیق و ماشین بولتزمن عمیق ۱۷

۳.۱.۲.۲ خودکدگذار پشته‌گذاری شده ۱۹
۴.۱.۲.۲ شبکه مولد تخاصمی ۲۰
۲.۲.۲ انواع معماری‌های شبکه عصبی کانوولوشنال ۲۱
۳.۲ بازشناسایی شخص ۲۶
۱.۳.۲ تاریخچه‌ی بازشناسایی شخص ۲۶
۲.۳.۲ مدل‌های عمیق بازشناسایی شخص و کمبود داده‌های آموزشی ۲۹
۱.۲.۳.۲ استفاده ازمعماری شبکه‌ی عصبی سیامی ۳۰
۲.۲.۳.۲ روش‌های داده‌افزایی ۳۳
۳.۲.۳.۲ داده‌افزایی با استفاده از شبکه‌های مولد تخاصمی ۳۳
۳.۳.۲ تعمیم‌پذیری و وفق‌دهی دامنه در مدل‌های بازشناسایی شخص ۳۵
۴.۳.۲ توابع اتلاف در مسئله‌ی بازشناسایی شخص ۳۹
۵.۳.۲ معیارهای ارزیابی عملکرد مدل‌های بازشناسایی شخص ۴۳
۶.۳.۲ مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص ۴۵
۷.۳.۲ خلاصه و نتیجه‌گیری ۵۰

فصل ۳: روش پیشنهادی ۵۲
۱.۳ پیش‌گفتار ۵۲
۲.۳ مدل پایه ۵۳
۱.۲.۳ حافظه‌ی نمونه ۵۳
۲.۲.۳ یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف ۵۴
۳.۳ روش پیشنهادی ۵۸
۱.۳.۳ بررسی استراتژی انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها . ۶۳
۴.۳ خلاصه و نتیجه‌گیری ۶۴

فصل ۴: نتایج علمی ۶۸
۱.۴ پیش‌گفتار ۶۸
۲.۴ مجموعه‌های داده ۶۸

۱.۲.۴ آماده‌سازی داده‌ها ۷۰
۳.۴ تنظیمات آزمایش ۷۱
۱.۳.۴ مشخصات سخت‌افزاری و نرم‌افزاری ۷۳
۴.۴ نتایج آزمایش ۷۴
۱.۰.۴.۴ نتایج استفاده از معماری‌های مختلف CNN در عملکرد مدل . . ۷۵
۲.۰.۴.۴ بررسی میزان مصرف حافظه‌ی GPU ۷۶
۳.۰.۴.۴ بررسی تأثیر پارامتر کلیدی در دو استراتژی انتخاب همسایه . . . ۷۷
۵.۴ مقایسه‌ی روش پیشنهادی با سایر روش‌ها ۸۰
۶.۴ خلاصه و نتیجه‌گیری ۸۳

فصل ۵: خلاصه، بحث، نتیجه‌گیری و کارهای آینده ۸۴
۱.۵ خلاصه ۸۴
۲.۵ بحث ۸۵
۳.۵ نتیجه‌گیری ۸۷
۴.۵ کارهای آینده ۸۷

مراجع ۸۹

فهرست شکل‌ها

۱.۱ مثالی از دوربین‌های امنیتی نظارتی با محدوده‌ی نظارت غیرهم‌پوشان ۳

۲.۱ چالش‌های تصاویر مجموعه‌داده‌های بازشناسایی شخص ۴

۳.۱ پنج مرحله‌ی اصلی طرحی یک سیستم بازشناسایی شخص ۸

۱.۲ نمودار زمانی رویدادهای مؤثر در پیدایش و گسترش حوزه‌ی یادگیری عمیق ۱۵

۲.۲ ساختار ماشین بولتزمن عمیق (DBM) و شبکه‌ی باور عمیق (DBN) ۱۸

۳.۲ ساختار کلی خودکدگذار ۲۰

۴.۲ ساختار کلی شبکه‌ی مولد تخاصمی ۲۱

۵.۲ ساختار معماری LeNet5 . . . . . ۲۲

۶.۲ ساختار معماری AlexNet ۲۳

۷.۲ ساختار یک سلول Inception ۲۳

۸.۲ ساختار یک بلاک باقی‌مانده‌ی ساده ۲۵

۹.۲ سمت چپ: یک بلاک از ResNet. سمت راست: یک بلاک از ResNeXt. ۲۵

۱۰.۲ تعدادی از رویدادهای مهم در پیدایش و گسترش پژوهش‌های بازشناسایی شخص ۲۷

۱۱.۲ سمت چپ: شبکه‌ی سیامی با ورودی زوج‌هایی از داده‌ها. سمت راست: شبکه‌ی سیامی با ورودی سه‌تایی‌هایی از داده‌ها ۳۱

۱۲.۲ توابع اتلاف شناسایی، تصدیق و سه‌گانه ۴۰

۱۳.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی VIPeR ۴۵

۱۴.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی GRID ۴۶

۱۵.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی CAVIAR4ReID ۴۶

۱۶.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی WARD ۴۷
۱۷.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی CUHK03 . . . . . ۴۸
۱۸.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی RAID ۴۸
۱۹.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی MSMT17 . . . . . ۴۹

۱.۳ تغییرناپذیری نسبت به نمونه‌ها ۵۵
۲.۳ تغییرناپذیری نسبت به دوربین‌ها ۵۶
۳.۳ تغییرناپذیری نسبت به همسایه‌ها ۵۷
۴.۳ مدل ارائه شده ۵۹
۵.۳ مثال‌های از سه‌تایی‌های مناسب برای تابع اتلاف سه‌گانه L_{tri} ۶۲

۱.۴ نمونه‌هایی از تصاویر مجموعه‌داده‌های Market1501 و DukeMTMC-reID ۶۹
۲.۴ نمونه‌هایی از تصاویر مجموعه‌داده‌ی DukeMTMC-reID و تصاویر تولیدشده از آن‌ها . . ۷۱
۳.۴ دو نمونه تصویر پرس‌وجو و چند نمونه‌ی ابتدای لیست بازیابی شده ۷۳
۴.۴ نتایج اجرای آزمایش‌ها با مقادیر مختلف k ۷۸
۵.۴ نتایج اجرای آزمایش‌ها با مقادیر مختلف μ ۷۹

فهرست جدول‌ها

۱.۲ نتایج تعدادی از موفق‌ترین مقالات در حوزه‌ی بازشناسایی شخص ۲۹
۲.۲ مجموعه‌داده‌های بازشناسایی شخص ۵۰

۱.۴ نتایج چندین اجرای مدل پیشنهادی در تنظیمات duke $\rightarrow$ market ۷۴
۲.۴ نتایج چندین اجرای مدل پیشنهادی در تنظیمات market $\rightarrow$ duke ۷۵
۳.۴ نتایج آزمایش مدل با به کارگیری معماری‌های مختلف ۷۶
۴.۴ مقایسه‌ی عملکرد و میزان مصرف حافظه‌ی GPU ۷۷
۵.۴ مقایسه‌ی روش پیشنهادی با سایر روش‌ها ۸۲

فهرست الگوریتم‌ها

۱.۳ استراتژی اول انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها ۶۶

۲.۳ استراتژی دوم انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها ۶۷

فصل ۱

مقدمه

۱.۱ پیش‌گفتار

مسئله‌ی بازشناسایی شخص یکی از مسائل پیچیده و پرکاربرد در حوزه‌ی بینایی ماشین است. این مسئله شامل بازیابی تصاویر فرد مورد جستجو[1] از میان مجموعه‌ی تصاویر ثبت شده توسط دوربین‌های غیرهم‌پوشان می‌باشد. باتوجه‌به اهمیت حفظ امنیت مکان‌های عمومی و افزایش کاربردهای دوربین‌های امنیتی نظارتی، حجم زیادی از داده‌های تصویری تولید می‌شوند که معمولاً دارای کیفیت و وضوح پایینی می‌باشند. پردازش و تحلیل و بررسی این داده‌ها در هنگام نیاز توسط نیروهای انسانی کار بسیار پیچیده و تقریباً غیرممکنی است. به همین دلیل وجود سیستم‌های اتوماتیک و هوشمند برای مسئله‌ی بازشناسایی شخص لازم می‌باشد.

به‌دلیل شرایط خاص ثبت تصاویر مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص توسط دوربین‌های امنیتی نظارتی، اغلب این تصاویر چالش‌هایی از قبیل تنوع میزان روشنایی، کیفیت پایین تصاویر، تنوع زاویه‌ی دید دوربین‌ها و ... را دارند. ازطرفی در تصاویر ثبت شده توسط دوربین‌های امنیتی نظارتی، به‌دلیل کیفیت پایین، چهره‌ی افراد با وضوح قابل مشاهده نیست و همچنین ممکن است افراد چهره‌ی خود را از دوربین‌ها مخفی نگه دارند. بنابراین نمی‌توان از ویژگی‌های چهره‌ی افراد برای بازشناسایی آن‌ها استفاده کرد و معمولاً از ویژگی‌های ظاهری افراد مثل لباس، وضعیت بدن و ... می‌توان بهره برد.

چالشی بودن مجموعه‌داده‌های بازشناسایی شخص، این مسئله را به یک مسئله‌ی پیچیده در حوزه‌ی بینایی

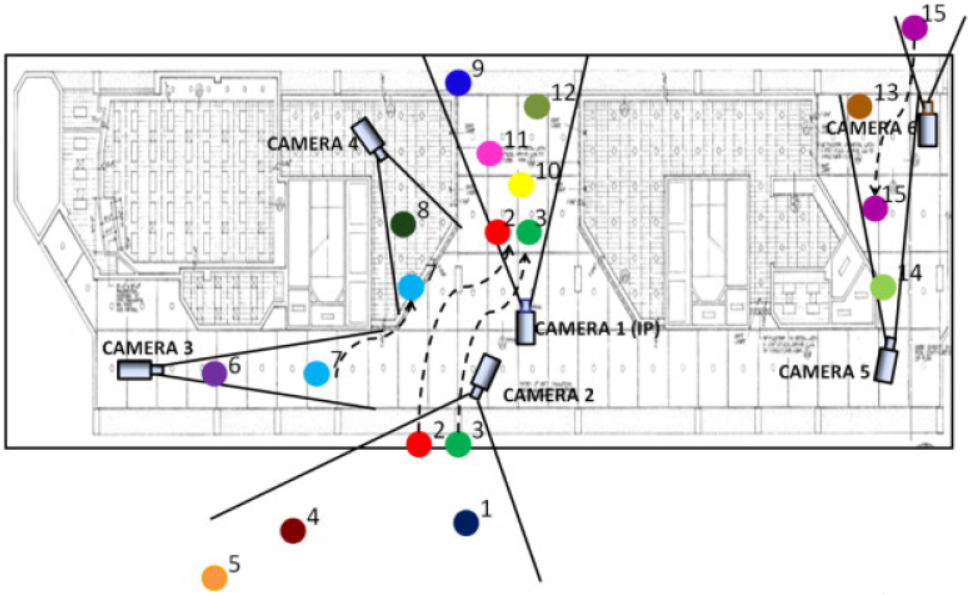
[1] query

ماشین تبدیل کرده است. در سال‌های اخیر با به کارگیری تکنیک‌های یادگیری عمیق، پیشرفت‌های الگوریتمی در یادگیری عمیق، پیشرفت سخت‌افزارهای پردازش موازی و گسترش مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص، نتایج موفقیت‌آمیزی در این حوزه حاصل شده است. باتوجه‌به پرهزینه‌بودن فرآیند برچسب‌گذاری داده‌ها و همچنین افت عملکرد مدل آموزش دیده هنگام آزمایش روی مجموعه‌داده‌ی بدون برچسب متفاوت با مجموعه‌داده‌ی آموزشی برچسب‌گذاری شده، مسئله‌ی وفق‌دهی دامنه[۲] در بازشناسایی شخص بسیار مهم می‌باشد. در این پژوهش سعی شده است که به این چالش پرداخته شود.

۲.۱ تعریف بازشناسایی شخص

امروزه استفاده از دوربین‌های امنیتی نظارتی با اهداف مختلف بسیار رایج است. از ویدیوها و تصاویر تهیه شده توسط این دوربین‌ها در کاربردهای مختلفی مانند حفظ امنیت محیط‌های عمومی ازجمله فرودگاه‌ها، دانشگاه‌ها و فروشگاه‌ها ، ردیابی حرکت عابرین پیاده، تحلیل آماری ترافیک، تحلیل رفتار بلند مدت انسان و ... استفاده می‌شود. مسئله‌ی بازشناسایی شخص یکی از مسائل مهم و پرکاربرد در حوزه‌ی بینایی ماشین است. بازشناسایی شخص شامل تشخیص وجود تصویر یک فرد در میان تصاویر جمع‌آوری شده توسط شبکه‌ای از دوربین‌های غیرهم‌پوشان می‌باشد. در مسئله‌ی بازشناسایی شخص تصویر یک فرد به‌عنوان پرس‌وجو به سیستم داده می‌شود و سیستم باید تصویر آن شخص را از میان تصاویر جمع‌آوری شده توسط دوربین‌های مختلف بازیابی کند. درواقع در این مسئله تعیین می‌شود که آیا دو تصویر از عابرپیاده که توسط دوربین‌های مختلف ثبت شده‌اند و یا توسط یک دوربین در زمان‌های متفاوتی ثبت شده‌اند متعلق به یک هویت می‌باشند و یا هویت‌های متفاوتی دارند. همچنین بازشناسایی شخص می‌تواند تعیین کند که تصویر فرد مدنظر توسط کدام دوربین‌ها ثبت شده است و بدین‌ترتیب مسیر حرکت فرد قابل ردیابی خواهد بود. در هنگام عدم وجود هم‌پوشانی در محدوده‌های تحت نظارت دوربین‌های مختلف، فرآیند بازیابی شخص به‌دلیل نبود دنباله‌ای از اطلاعات، دشوار خواهد بود. در شکل ۱.۱ مثالی از دوربین‌های امنیتی نظارتی با محدوده‌ی نظارت غیرهم‌پوشان نمایش داده شده است.

[2] domain adaptation



شکل ۱.۱: مثالی از دوربین‌های امنیتی نظارتی با محدوده‌ی نظارت غیرهم‌پوشان [۴۳]

باتوجه‌به اهمیت حفظ امنیت مکان‌های عمومی و افزایش کاربرد دوربین‌های امنیتی نظارتی، حجم عظیمی از داده‌های تصویری تولید می‌شوند که پردازش آن‌ها در هنگام نیاز به صورت غیراتوماتیک و توسط اپراتورهای انسانی کاری بسیار پرهزینه و تقریباً غیرممکن است. ازطرفی باتوجه‌به پیشرفت‌هایی که در سخت‌افزارهای پردازش موازی و الگوریتم‌های یادگیری ماشین به وجود آمده است، بازشناسایی شخص اتوماتیک و هوشمند به مسئله‌ی پرطرفداری تبدیل شده است.

تا کنون مطالعات زیادی در حوزه‌ی تشخیص چهره انجام شده است اما در کاربردهای عملی، دوربین‌ها همیشه قادر به تهیه‌ی تصویر واضح از صورت افراد نیستند. درنتیجه وجود سیستم‌هایی که قادر به بازشناسایی شخص با استفاده از ویژگی‌های تمام بدن فرد باشند ضروری است. ازطرفی دوربین‌ها اغلب محدوده‌های غیرهم‌پوشانی را نظارت می‌کنند. بنابراین بازشناسایی شخص با استفاده از ویژگی‌های کلی بدن افراد و ردیابی فرد در میان دوربین‌های مختلف می‌تواند در کنار تکنولوژی شناسایی چهره در سناریوهای دنیای واقعی بسیار کمک‌کننده باشد.

مسائل آشکارسازی فرد[۳] ، بازشناسی فرد[۴] ، شناسایی شخص[۵] و جستجوی شخص[۶] مسائلی هستند که

[3]person detection
[4]person recognition
[5]person identification
[6]person search

مفهومی نزدیک به مسئله‌ی بازشناسایی شخص دارند اما مسائل کاملاً متفاوتی می‌باشند. در مسئله‌ی آشکارسازی فرد وجود و یا عدم وجود انسان در یک تصویر مشخص می‌شود. در مسائل بازشناسی، نوع مدنظر(مثلا حیوان، انسان و ...) در تصویر تشخیص داده می‌شود. در مسئله‌ی شناسایی شخص یک فرد خاص از میان افراد مختلف شناسایی می‌شود. اما در مسئله‌ی بازشناسایی شخص سیستم باید تمامی تصاویر یک شخص خاص که توسط چندین دوربین ثبت شده‌اند را بازیابی کند. مسئله‌ی جستجوی شخص مقداری پیچیده‌تر از بازشناسایی شخص است. در بازشناسایی شخص فرض می‌شود که تصاویر دراختیار، محدود شده هستند وهویت افراد در آن‌ها مشخص است اما در جستجوی شخص چنین فرض محدودکننده‌ای وجود ندارد.

۳.۱ چالش‌های بازشناسایی شخص

مسئله‌ی بازشناسایی شخص یکی از مسائل پیچیده و چالشی در حوزه‌ی بینایی ماشین محسوب می‌شود. یکی از دلایل پیچیده بودن این مسئله، وجود چالش‌های مختلف درتصاویر مجموعه‌داده‌های این حوزه است.

شکل ۲.۱: a) کیفیت پایین تصاویر b) تنوع میزان روشنایی c) تنوع زاویه دید دوربین/حالت قرارگیری

- **تنوع میزان روشنایی**[7]: از آنجایی که تصاویر مجموعه‌داده‌های بازشناسایی شخص درساعت‌های مختلف شبانه‌روز و در محیط‌های سرپوشیده و سرباز ثبت شده‌اند، از لحاظ میزان روشنایی باهم تفاوت دارند. تفاوت شرایط روشنایی روی رنگ تصویرهای مختلف تأثیرگذار خواهد بود و باعث افزایش تفاوت‌های درون-کلاسی می‌شود.

- **تنوع زاویه دید دوربین**[8]: تصاویر مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص اغلب توسط دوربین‌های مختلف ثبت شده‌اند. باتوجه‌به زاویه‌ی دید دوربین‌ها و فاصله‌ی افراد از دوربین‌ها تصاویر ثبت شده معمولاً شامل بخشی از بدن افراد می‌باشند و در هر زاویه‌ی دید بخشی از ظاهر افراد قابل مشاهده نیست. این مسئله چالش بسیار مهمی در حوزه‌ی بازشناسایی شخص به‌شمار می‌آید به‌نحوی‌که در سال‌های اخیر تعدادی از مقالات سعی کرده‌اند با تمرکز روی چالش تنوع زاویه دید و سبک دوربین‌های مختلف، مدل‌های بازشناسایی شخص موفقی را ارائه کنند [۱۱۴].

- **تنوع حالت قرارگیری افراد**[9]: در تصاویر مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص تصاویری در حالت‌های مختلف از افراد وجود دارد. دلیل این تنوع حالات این است که دوربین‌ها ممکن است تصاویری از افراد در حالت راه رفتن، در حالت نشستن و ... ثبت کنند. تنوع حالت می‌تواند چالش محسوب شود. در سال‌های اخیر تعدادی از مقالات سعی کردند که با تمرکز روی این چالش، مدل‌هایی را ارائه کنند که نسبت به حالات قرارگیری افراد مقاوم هستند [۲۳، ۵۸].

- **رزولوشن و وضوح پایین تصاویر**: تصاویر مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص اغلب توسط دوربین‌های امنیتی نظارتی در محیط‌های عمومی ثبت شده‌اند. برخی از این دوربین‌ها، قدیمی و با کیفیت پایین می‌باشند. ازطرفی دوربین‌های امنیتی نظارتی اغلب هزینه‌های زیادی دارند و برای نظارت کامل یک محیط تعداد قابل توجهی دوربین نیاز است. اگر این دوربین‌ها در فواصل نزدیک به زمین نصب شوند، از محدوده‌ی کمتری پشتیبانی می‌کنند ولی تصاویر با وضوح بیشتری از افراد را در اختیار قرار می‌دهند. برای اینکه دوربین‌ها از محدوده‌های بیشتری پشتیبانی کنند معمولاً آن‌ها را در ارتفاع نصب می‌کنند بنابراین درصورت استفاده از دوربین‌هایی با وضوح بالا نیز، تصاویر افراد کیفیت زیادی ندارند.

[7] illumination variation
[8] camera viewpoint variation
[9] pose variation

- **وجود افراد مختلف با لباس‌های مشابه:** از آنجایی که در تصاویر مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص، چهره‌ی افراد با جزئیات قابل مشاهده نیست، نمی‌توان از ویژگی‌های چهره برای بازشناسایی آن‌ها بهره‌ی زیادی برد. بنابراین به نظر می‌رسد که ویژگی‌های ظاهری مثل لباس و رنگ آن برای بازشناسایی کمک‌کننده می‌باشد. در این حالت وجود افراد مختلف با لباس‌های مشابه می‌تواند چالش محسوب شود.

- **انسداد**[10]: در یک تصویر ممکن است تصاویر چندین فرد وجود داشته باشد و بخش‌هایی از تصاویر این افراد باهم هم‌پوشانی داشته باشند. معمولاً هنگامی که افراد در یک محیط شلوغ در حال راه رفتن هستند، انسداد در تصاویرشان وجود دارد. انسداد باعث می‌شود که تمامی بخش‌های کلیدی تصویر یک فرد قابل مشاهده نباشد و می‌تواند باعث بروز مشکل در فرآیند بازشناسایی گردد.

- **کمبود داده‌های آموزشی:** در مجموعه‌داده‌های بازشناسایی شخص معمولاً به ازای هر هویت تعداد تصاویر کمی وجود دارد. برای مثال در مجموعه‌داده‌ی VIPeR از هر هویت دو تصویر موجود است. کمبود نمونه‌های آموزشی از هر کلاس، می‌تواند باعث بیش‌برازش شده و سبب می‌شود که مدل قادر به یادگیری ویژگی‌های تمییزدهنده نباشد. تکنیک‌های مختلفی برای مقابله با چالش کمبود نمونه‌های آموزشی وجود دارد که در بخش ۲.۳.۲ به این مسئله پرداخته می‌شود.

- **تعمیم‌پذیری مدل:** یکی از چالش‌های مهم مسئله‌ی بازشناسایی شخص تعمیم‌پذیری مدل هنگام آزمایش روی یک مجموعه‌داده‌ی متفاوت از مجموعه‌داده‌ی آموزشی برچسب‌گذاری شده است. در این حالت مشاهده شده که عملکرد مدل به شدت کاهش می‌یابد. این کاهش عملکرد به‌دلیل تفاوت دامنه‌های مجموعه‌داده‌های آموزشی و آزمایشی می‌باشد. در سال‌های اخیر مقالات متعددی سعی کرده‌اند که به مسئله‌ی بازشناسایی شخص کنار-دامنه‌ای[11] و ارائه‌ی مدل‌های تعمیم‌پذیر در دامنه‌های مختلف بپردازند. در بخش ۳.۳.۲ به این چالش مهم در حوزه‌ی بازشناسایی شخص پرداخته شده است.

- **برچسب‌گذاری داده‌ها:** در مسئله‌ی پیچیده‌ای مانند بازشناسایی شخص، حجم قابل توجهی از داده‌های آموزشی دارای برچسب نیاز است که بتوان یک مدل را در این حوزه به‌شکل کارآمد و با قابلیت بازشناسایی مناسب آموزش داد. جمع‌آوری تصاویر از افراد از شبکه‌ای از دوربین‌های مختلف و برچسب‌گذاری این تصاویر جمع‌آوری شده کار بسیار پرهزینه‌ای است.

[10] occlusion
[11] cross-domain person reidentification

- **سرعت پاسخگویی**: باتوجه‌به کاربرد عمده‌ی بازشناسایی شخص در حوزه‌ی امنیت، در شرایط اضطراری نیاز است که تصویر فرد موردنظر به سرعت بازشناسایی شود و سیستم در زمان کوتاه عملکرد مناسبی را ارائه دهد.

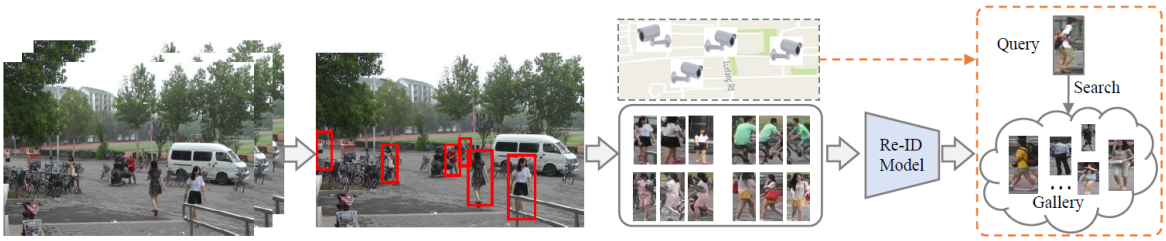
- **پیچیدگی مدل و نیاز سخت‌افزاری**: باتوجه‌به چالشی بودن وظیفه‌ی بازشناسایی شخص، و حجیم بودن مجموعه‌داده‌های موجود برای این مسئله، اغلب مدل‌های پیچیده‌ای با لایه‌های متعدد برای این مسئله ارائه می‌شود. پیچیده بودن مدل‌ها باعث می‌شود که برای فرآیند آموزش آن‌ها سخت‌افزارهای پیشرفته‌ای نیاز باشد.

۴.۱ بازشناسایی شخص جهان-بسته و جهان-باز

مسائل بازشناسایی شخص را می‌توان به دو دسته‌ی مسائل بازشناسایی شخص جهان-بسته[۱۲] و مسائل بازشناسایی شخص جهان-باز[۱۳] تقسیم کرد. در مسائل جهان-باز فرض‌هایی محدودکننده وجود دارد که ممکن است در مسائل دنیای واقعی مدل را ناکارآمد سازند. در سال‌های اخیر، به‌ویژه با به کارگیری تکنیک‌های یادگیری عمیق، در مسئله‌ی جهان-بسته‌ی بازشناسایی شخص، نتایج بسیار خوبی به‌دست آمده است، به‌طوری‌که بهبود عملکرد مدل‌ها روی مجموعه‌داده‌های این حوزه کار دشواری است و تقریباً به حالت اشباع نزدیک شده است. در سال‌های اخیر بسیاری از پژوهش‌ها سعی می‌کنند که به مسئله‌ی بازشناسایی شخص با فرض‌های محدودکننده‌ی کمتری بپردازند و به تنظیمات جهان-باز نزدیک‌تر شوند.

در مقاله‌ی [۹۴] مسئله‌ی بازشناسایی شخص به پنج مرحله تقسیم می‌شود. در شکل ۳.۱ این پنج مرحله نمایش داده شده است. بسیاری از مقالات بازشناسایی شخص را با بازیابی شخص[۱۴] یکسان درنظر می‌گیرند، درحالی‌که بازیابی شخص یکی از مراحل مسئله‌ی بازشناسایی شخص می‌باشد. پنج مرحله‌ی بازشناسایی شخص در ادامه آورده شده است.

[12] closed-world
[13] open-world
[14] person retrieval



شکل ۳.۱: پنج مرحله‌ی اصلی طرحی یک سیستم بازشناسایی شخص [۹۴]

- **جمع‌آوری داده** : داده‌های مسئله‌ی بازشناسایی شخص اغلب از دوربین‌های امنیتی نظارتی که در محیط‌های مختلف مثل فروشگاه‌ها، دانشگاه‌ها، فرودگاه‌ها و ... وجود دارند، جمع‌آوری می‌شوند. این دوربین‌ها در شرایط مختلف ویدیوها یا تصاویر را ثبت کرده‌اند. در تنظیمات جهان-بسته داده‌های خام اغلب با کیفیت‌های تقریباً مشابهی و از دامنه‌های تقریباً یکسانی ثبت شده‌اند. درحالی‌که در تنظیمات جهان-باز، داده‌های جمع‌آوری شده می‌توانند ناهمگون باشند.

- **قرار دادن چهارگوش محیطی روی تصویر شخص**[۱۵] : داده‌های خام جمع‌آوری شده اغلب شامل تصاویر اشیای مختلفی هستند. باید در داده‌های خام اولیه، تصاویر هویت‌ها مشخص شود و اطراف تصویر هر فرد یک کادر تعیین گردد. از آنجایی که حجم داده‌های جمع‌آوری شده معمولاً زیاد است و به‌صورت دستی امکان تعیین تصاویر افراد و برش آن‌ها وجود ندارد، اغلب از الگوریتم‌های ردیابی فرد[۱۶] و آشکارسازی فرد استفاده می‌شود. در تنظیمات جهان-بسته فرض می‌شود که تصاویری که دراختیار داریم محدود شده هستند و تصاویر افراد و هویت‌ها در آن‌ها مشخص شده است. درحالی‌که در تنظیمات جهان-باز ممکن است که مسئله‌ی بازشناسایی شخص روی داده‌های خام و بدون مشخص شدن هویت‌ها صورت گیرد که در این صورت به مسئله‌ی جستجوی شخص[۱۷] تبدیل می‌شود.

- **برچسب‌گذاری تصاویر آموزشی** : برای آموزش یک مدل بازشناسایی شخص با قدرت تشخیص مناسب روی مجموعه‌داده‌ی جمع‌آوری شده از دوربین‌های مختلف، باید برچسب داده‌های آموزشی در دسترس باشد. در شرایطی که تغییر دامنه‌ی زیادی وجود داشته باشد، لازم است که داده‌های آموزشی در هر سناریوی جدید، برچسب‌گذاری شوند. در تنظیمات جهان-بسته فرض می‌شود که برای آموزش مدل

[15] bounding box
[16] person tracking
[17] person search

بانظارت، به اندازه‌ی کافی داده‌ی آموزشی برچسب‌گذاری شده موجود است. درحالی‌که در تنظیمات جهان-باز داده‌های دارای برچسب می‌توانند محدود باشند و یا هیچ اطلاعاتی از برچسب‌های داده‌های آموزشی وجود نداشته باشد. این مسئله به مدل‌های بازشناسایی شخص نیمه نظارتی یا بدون نظارت منجر می‌شود.

- **آموزش مدل بازشناسایی شخص** : در این مرحله با استفاده از داده‌های برچسب‌گذاری شده در مرحله‌ی قبل، یک مدل مقاوم بازشناسایی شخص آموزش داده می‌شود. در تنظیمات جهان-بسته فرض می‌شود که تمامی برچسب‌گذاری‌ها در مرحله‌ی قبل صحیح انجام شده است و دارای هیچ نویزی نیست. درحالی‌که در عمل وجود نویز اجتناب ناپذیر می‌باشد.

- **بازیابی عابرپیاده**[۱۸] : در فاز آزمایش باید تصویر یک عابرپیاده بازیابی شود. روش کار معمولاً به این صورت است که یک تصویر به‌عنوان پرس‌وجو و یک مجموعه‌ای از تصاویر وجود دارند. پس از استخراج بازنمایی‌های ویژگی با استفاده از مدل یاد گرفته شده در مرحله‌ی قبل، یک لیست رتبه‌بندی شده از طریق مرتب‌سازی شباهت‌های محاسبه شده‌ی تصویر پرس‌وجو با مجموعه‌ی تصاویر، بازیابی می‌شود. در این مرحله تمرکز بر روی بهینه‌سازی فرآیند رتبه‌بندی می‌تواند به بهبود عملکرد کمک زیادی کند. در تنظیمات جهان-بسته فرض می‌شود که تصویر پرس‌وجو حتما در مجموعه‌ی تصاویر موجود است. درحالی‌که در شرایط جهان-باز ممکن است تصویر فرد موردنظر در مجموعه‌ی تصاویر موجود نباشد. در این حالت به‌جای بازیابی باید از تصدیق[۱۹] استفاده کرد.

۵.۱ اهداف پایان‌نامه

هدف این پایان‌نامه، ارائه‌ی مدلی تعمیم‌پذیر برای مسئله‌ی بازشناسایی شخص با استفاده از رویکرد یادگیری عمیق می‌باشد. منظور از تعمیم‌پذیری مدل، عملکرد مناسب آن هنگام آزمایش روی مجموعه‌داده‌ی بدون برچسب می‌باشد. فرآیند برچسب‌گذاری تصاویر، بسیار پرهزینه و زمان‌بر است. بنابراین برچسب‌گذاری تصاویر در هر سناریوی واقعی غیرممکن می‌باشد. موضوع وفق‌دهی دامنه درمسئله‌ی بازشناسایی شخص از اهمیت زیادی

[18] pedestrian retrieval
[19] verification

برخوردار است. در این پایان‌نامه سعی شده که به مسئله‌ی وفق‌دهی دامنه‌ی بدون نظارت در بازشناسایی شخص پرداخته شود و مدلی ارائه شود که در فرآیند آموزش از یک مجموعه‌داده‌ی دارای برچسب و یک مجموعه‌داده‌ی بدون برچسب بهره می‌برد و باید در هنگام آزمایش عملکرد مناسبی روی مجموعه‌داده‌ی بدون برچسب داشته باشد. مجموعه‌داده‌ی دارای برچسب، مجموعه‌داده‌ی منبع[۲۰] و مجموعه‌داده‌ی بدون برچسب، مجموعه‌داده‌ی هدف[۲۱] نام دارند. در مدل ارائه شده علاوه بر پرداختن به ویژگی‌های دامنه‌ی منبع و تفاوت‌های ویژگی‌های دامنه‌ی منبع و هدف، تفاوت‌های درون-دامنه‌ای در دامنه‌ی هدف نیز درنظر گرفته می‌شوند. بنابراین درهنگام آزمایش مدل روی مجموعه‌داده‌ی هدف، عملکرد مدل بهبود می‌یابد.

در مدل پیشنهادی علاوه‌بر تابع اتلاف طبقه‌بندی داده‌های منبع، از تابع اتلاف مربوط به یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف و همچنین از تابع اتلاف سه‌گانه مربوط به یادگیری تفاوت‌های بین دامنه‌ای دامنه‌ی منبع و دامنه‌ی هدف و تفاوت‌های درون-دامنه‌ای دامنه‌ی هدف، نیز استفاده می‌شود.

به این ترتیب این تحقیق با اهداف زیر انجام می‌شود:

- آیا می‌توان مدلی برای بازشناسایی شخص ارائه کرد به‌نحوی‌که هنگام آزمایش روی یک مجموعه‌داده‌ی بدون برچسب، عملکرد قابل قبولی داشته باشد؟
- آیا درنظرگرفتن ویژگی‌های داده‌های دامنه‌ی هدف در فرآیند آموزش، می‌تواند موجب بهبود عملکرد مدل در هنگام آزمایش روی داده‌های بدون برچسب دامنه‌ی هدف شود؟
- آیا استفاده از روش‌های داده‌افزایی، به‌ویژه داده‌افزایی با استفاده از شبکه‌های مولد تخاصمی، تأثیری بر عملکرد مدل بازشناسایی شخص دارد؟
- آیا اضافه کردن تابع اتلاف سه‌گانه می‌تواند موجب بهبود عملکرد مدل بازشناسایی شخص شود؟

۶.۱ نوآوری‌های پایان‌نامه

این پایان‌نامه به مسئله‌ی بازشناسایی شخص و به‌طور ویژه به موضوع وفق‌دهی دامنه در بازشناسایی شخص پرداخته است. مرور ادبیات موضوع به‌طور مفصل انجام شده است. در این پایان‌نامه مدلی تعمیم‌پذیر برای مسئله‌ی

[20] source
[21] target

وفق‌دهی دامنه‌ی بدون نظارت در بازشناسایی شخص ارائه شده است. مدل پیشنهادی در فرآیند آموزش از یک مجموعه‌داده‌ی برچسب‌گذاری شده و یک مجموعه‌داده‌ی بدون برچسب یاد می‌گیرد. مدل پیشنهادی قادر است که در هنگام آزمایش روی مجموعه‌داده‌ی بدون برچسب، عملکرد خوبی داشته باشد. درحین آموزش، علاوه بر درنظرگرفتن تفاوت‌های بین دامنه‌ای میان دو دامنه، سه ویژگی درون-دامنه‌ای در مجموعه‌داده‌ی بدون برچسب نیز درنظر گرفته شده است. توجه به تغییرات درون-دامنه‌ای در مجموعه‌داده‌ی بدون برچسب باعث می‌شود که مدل نسبت به تغییرات در این دامنه مقاوم‌تر شود و هنگام آزمایش روی دامنه‌ی بدون برچسب هدف عملکرد بهتری داشته باشد. سه ویژگی درنظر گرفته شده در دامنه‌ی هدف، ویژگی‌های تغییرناپذیری نسبت به نمونه‌ها، تغییرناپذیری نسبت به دوربین‌ها و تغییرناپذیری نسبت به همسایه‌ها می‌باشند. در یادگیری تغییرناپذیری نسبت به همسایه‌ها، دو استراتژی انتخاب همسایه‌ها مورد آزمایش قرار گرفته‌اند.

در مدل پیشنهادی از یک تابع اتلاف سه‌گانه نیز استفاده شده است. در این تابع اتلاف سه‌گانه علاوه‌بر یادگیری تفاوت‌های درون-دامنه‌ای در دامنه‌ی هدف، تفاوت‌های بین دو دامنه نیز بررسی می‌شوند. بهره بردن از تابع اتلاف سه‌گانه‌ی یاد شده، باعث بهبود عملکرد مدل نسبت به مدل پایه [۱۱۲] شده است. تابع اتلاف نهایی شبکه از مجموع تابع اتلاف طبقه‌بندی داده‌ها در مجموعه‌داده‌ی برچسب‌گذاری شده، تابع اتلاف یادگیری تغییرناپذیری‌ها در مجموعه‌داده‌ی بدون برچسب و تابع اتلاف سه‌گانه تشکیل می‌شود.

به‌منظور استخراج ویژگی از تصاویر دو دامنه، از شبکه‌ی ResNeXt-50 [۹۱] استفاده شده است که نسبت به شبکه‌ی ResNet-50 [۲۹] تعداد پارامترهای قابل آموزش کمتری داشته و در عین حال عملکرد مناسب‌تری در مدل پیشنهادی دارد.

مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها توانسته است روی duke→market در رتبه‌ی ۱ و ۵ معیار CMC مقدار ۷۷/۴ درصد و ۸۹/۱ درصد و مقدار mAP ۴۶ درصد را به‌دست آورد. همچنین در تنظیمات market → duke در رتبه‌ی ۱ و ۵ معیار CMC مقدار ۶۴ درصد و ۷۶/۸ درصد و مقدار mAP ۴۱/۱ درصد را به‌دست آورده است. مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها توانسته است روی duke→market در رتبه‌ی ۱ و ۵ معیار CMC مقدار۸۴/۵ درصد و ۹۳/۱ درصد و مقدار mAP ۶۳ درصد را به‌دست آورد. همچنین در تنظیمات market → duke در رتبه‌ی ۱ و ۵ معیار CMC مقدار ۷۰/۱ درصد و ۸۰/۸ درصد و مقدار mAP ۴۹/۱ درصد را به‌دست آورده است.

۷.۱ ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل می‌باشد. فصل اول، فصل مقدمه است . در این فصل (مقدمه) در رابطه با تعریف مسئله‌ی بازشناسایی شخص و چالش‌های آن توضیحاتی ارائه شده است. همچنین مراحل مسئله‌ی بازشناسایی شخص جهان-بسته و جهان-باز شرح داده شده‌اند.

فصل دوم که مربوط به مروری بر مطالعات پیشین و کارهای دیگران است، شامل دو بخش کلی است. بخش اول آن فصل مربوط به یادگیری عمیق و تأثیر یادگیری عمیق در بینایی ماشین می‌باشد. در آن بخش انواع مدل‌های یادگیری عمیق و معماری‌های متداول شبکه‌های عصبی کانوولوشنال مختصراً معرفی می‌شوند. در بخش دوم فصل دوم، به مسئله‌ی بازشناسایی شخص پرداخته شده است. تاریخچه‌ی بازشناسایی شخص، توابع اتلاف رایج در بازشناسایی شخص، معیارهای ارزیابی عملکرد مدل‌های بازشناسایی شخص و مجموعه‌داده‌های این حوزه مورد بحث قرار می‌گیرند. همچنین دو مورد از مهم‌ترین چالش‌های حوزه‌ی بازشناسایی شخص یعنی کمبود داده‌های آموزشی و مسئله‌ی تعمیم‌پذیری و وفق‌دهی دامنه در بازشناسایی شخص بررسی می‌شوند.

در فصل سوم روش پیشنهادی توضیح داده شده است. در بخش اول آن فصل، ساختار مدل پایه و نکات مربوط به یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف بیان می‌شوند. در بخش بعدی فصل سوم، ساختار مدل پیشنهادی و توابع اتلاف استفاده شده در مدل پیشنهادی توضیح داده می‌شوند. همچنین دو استراتژی مختلف انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها بیان می‌شوند.

در فصل چهارم تنظیمات آزمایش و نحوه‌ی آماده‌سازی داده‌ها توضیح داده شده و نتایج اجراهای مختلف مقایسه و تحلیل می‌شوند. فصل پنجم مربوط به نتیجه‌گیری و بیان ایده‌هایی برای کارهای بعدی می‌باشد.

فصل ۲

مروری بر مطالعات انجام شده

۱.۲ پیش‌گفتار

بخش اول این فصل درباره‌ی تاریخچه‌ی یادگیری عمیق، انواع مدل‌های عمیق و اهمیت یادگیری عمیق در بینایی ماشین می‌باشد. در بخش دوم این فصل به مسئله‌ی بازشناسایی شخص پرداخته می‌شود. ابتدا تاریخچه‌ی مسئله‌ی بازشناسایی شخص بیان شده و سپس به دو چالش مهم در مسئله‌ی بازشناسایی شخص، یعنی کمبود نمونه‌های آموزشی و تعمیم‌پذیری و وفق‌دهی دامنه در این حوزه و راهکارهای ارائه شده برای این چالش‌ها، پرداخته شده است. توابع اتلاف رایج در مسئله‌ی بازشناسایی شخص، معیارهای ارزیابی عملکرد مدل و مجموعه‌داده‌های بازشناسایی شخص در این فصل بررسی شده‌اند.

۲.۲ یادگیری عمیق در بینایی ماشین

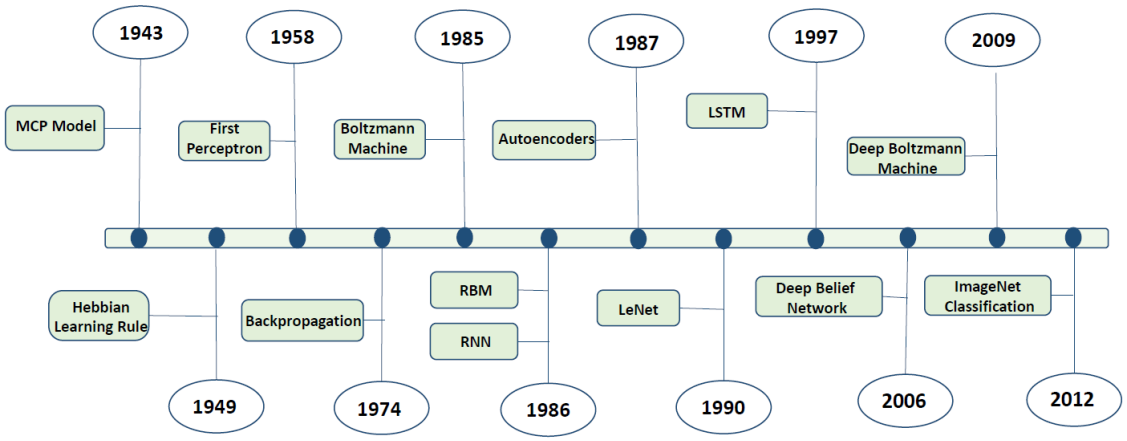
در سال‌های اخیر الگوریتم‌های یادگیری عمیق توانسته‌اند در بسیاری از حوزه‌ها به‌ویژه مسائل بینایی ماشین، عملکرد بهتری از الگوریتم‌های یادگیری ماشین کلاسیک داشته باشند. در یادگیری عمیق مدل می‌تواند لایه‌های متعددی از انتزاع را یاد بگیرد و با استفاده از سلسله مراتبی از چندین لایه، از داده‌ها معنا استخراج کند. عملکرد مدل‌های عمیق از نحوه‌ی عملکرد مغز در چگونگی دریافت و فهم اطلاعات الگوبرداری شده‌است و بدین‌ترتیب

می‌توانند ساختارهای پیچیده در داده‌ها را استخراج کنند. البته عملکرد مغز بسیار پیچیده‌تر از مدل‌های عمیق امروزی می‌باشد.

در الگوریتم‌های یادگیری کلاسیک مرحله‌ی استخراج ویژگی و مهندسی ویژگی‌ها، که مرحله‌ای پرهزینه می‌باشد، برعهده‌ی انسان است و به‌صورت غیراتوماتیک انجام می‌شود. درحالی‌که در الگوریتم‌های یادگیری عمیق، داده‌های ورودی به مدل داده می‌شوند و مدل مرحله‌ی استخراج ویژگی‌ها را به‌شکل اتوماتیک انجام می‌دهد.

در سال ۱۹۴۳ McCulloch و Pitt [۶۰] سعی کردند که چگونگی عملکرد مغز در فهم و تولید الگوهای پیچیده را با استفاده از سلول‌های پایه‌ی به‌هم متصل به نام نورون، متوجه شوند. آن‌ها مدل MCP را ارائه کردند که سهم بزرگی در تولید شبکه‌های عصبی مصنوعی داشت. پس از آن تلاش‌های زیادی برای گسترش حوزه‌ی یادگیری ماشین و شبکه‌های عصبی مصنوعی صورت گرفت. در سال ۱۹۴۹ قاعده‌ی یادگیری هبی [۳۰] معرفی شد. در سال ۱۹۵۸ نخستین پرسپترون[1] توسط Rosenblatt ایجاد شد [۶۹] و در سال ۱۹۷۴ Werbos الگوریتم پس‌انتشار[2] را معرفی کرد [۹۰]. در دهه‌ی ۱۹۸۰ ماشین بولتزمن[3][۱]، ماشین بولتزمن محدود[4][۷۵]، شبکه‌های عصبی بازگشتی[5][۴۱] و خودکدگذارها[6][۲] معرفی شدند. در سال ۱۹۹۰ شبکه‌ی LeNet توسط LeCun معرفی شد [۴۴] و باعث شروع حوزه‌ی شبکه‌های عصبی کانوولوشنال گردید. شبکه‌ی LSTM در سال ۱۹۹۷ ایجاد شد [۳۶]. یکی از عوامل اساسی شکل‌گیری یادگیری عمیق شبکه‌های باور عمیق[7] می‌باشند که در سال ۲۰۰۶ توسط Hinton معرفی شدند [۳۲]. شبکه‌ی باور عمیق شامل چندین لایه از ماشین بولتزمن محدود است. در سال ۲۰۰۹ ماشین بولتزمن عمیق معرفی گردید [۷۰] و از سال ۲۰۱۲ با معرفی AlexNet [۴۲] استفاده از شبکه‌های عصبی کانوولوشنال برای وظیفه‌ی طبقه‌بندی روی مجموعه‌داده‌ی ImageNet [۱۶] آغاز شد. در شکل ۱.۲ نمودار زمانی رویدادهای مهم در حوزه‌ی یادگیری ماشین که منجر به پیدایش و گسترش حوزه‌ی یادگیری عمیق شده‌اند، نمایش داده شده است [۸۵].

[1] perceptron
[2] backpropagation
[3] boltzmann machine
[4] restricted boltzmann machine
[5] recurrent neural network
[6] autoencoders
[7] deep belief network



شکل ۱.۲: نمودار زمانی رویدادهای مؤثر در پیدایش و گسترش حوزه‌ی یادگیری عمیق

کاربرد الگوریتم‌های یادگیری عمیق در مسائل مختلفی در بینایی ماشین به اثبات رسیده است. مدل‌های عمیق متعددی در مسائل مختلف حوزه‌ی بینایی ماشین ازجمله تشخیص چهره، بخش‌بندی تصویر، ترجمه‌ی تصویر به تصویر، بهبود کیفیت تصاویر، بازشناسایی شخص و ... ارائه شده‌اند. عوامل مختلفی در پیشرفت کاربردها و موفقیت‌های یادگیری عمیق مؤثر بوده‌اند. مدل‌های عمیق نیاز به حجم زیادی از داده دارند، با فراهم شدن مجموعه‌داده‌های متعدد و به اندازه‌ی کافی بزرگ در حوزه‌های مختلف امکان استفاده از مدل‌های عمیق برای مسائل گوناگون فراهم شد.

ازطرفی مدل‌های عمیق مدل‌های پیچیده‌ای هستند و نیاز به پردازش‌های پیچیده دارند و برای استفاده‌ی بهینه از آن‌ها سخت‌افزارهایی با قدرت پردازشی بالا لازم است. با فراهم شدن امکان پردازش های موازی توسط GPU های قدرتمند و امکان انتقال پردازش‌ها از CPU به GPU فرآیند آموزش در مدل‌های عمیق با سرعت بسیار بالاتری قابل انجام خواهد بود.

با معرفی ابزارهای پیاده‌سازی نسبتاً ساده برای الگوریتم‌های یادگیری عمیق مثل Tensorflow، Pytorch، Keras، Theano، Caffe و ... استفاده از یادگیری عمیق برای کاربردهای پژوهشی و صنعتی بسیار آسان‌تر شد. پیش از معرفی چنین کتابخانه‌هایی، پیاده‌سازی مدل‌های عمیق نیاز به تخصص بالا در زبان برنامه‌نویسی C++ و CUDA داشت.

یکی دیگر از عوامل رشد سریع یادگیری عمیق، پیشرفت‌های الگوریتمی در این حوزه بود. در ابتدا آموزش شبکه‌های بسیار عمیق با لایه‌های متعدد به‌دلیل مشکل انتشار گرادیان امکان‌پذیر نبود. سیگنال فیدبک مورداستفاده برای آموزش شبکه‌ی عصبی با افزایش تعداد لایه‌ها به‌تدریج محو می‌شد. به‌تدریج بهبودهای الگوریتمی ساده

امکان انتشار بهتر گرادیان‌ها را فراهم کردند. یکی از این بهبودها ارائه‌ی توابع فعالیت بهتری مانند تابع فعالیت ReLU برای لایه‌های شبکه‌ی عصبی بود. روش‌های مقداردهی اولیه‌ی وزن‌ها و روش‌های بهینه‌سازی مانند RMSProp و یا Adam نیز در عملکرد الگوریتم‌های یادگیری عمیق تأثیر زیادی دارند. همچنین روش‌های مختلف منظم‌سازی مانند استفاده از برون‌اندازی[۸]، نرمال‌سازی دسته‌ای[۹] و داده‌افزایی[۱۰] نیز موجب بهبود عملکرد مدل‌های عمیق می‌شود [۸۵].

۱.۲.۲ انواع مدل‌های یادگیری عمیق

۱.۱.۲.۲ شبکه عصبی کانوولوشنال

شبکه‌های عصبی کانوولوشنال از ساختار سیستم بصری، به‌ویژه مدل‌های ارائه شده در [۳۹] الهام گرفته شده‌اند. در سال ۱۹۹۰ LeCun و همکارانش شبکه‌ی عصبی کانوولوشنالی طراحی کردند که ازخطای گرادیانی استفاده می‌کرد و نتایج بسیار خوبی در وظایف مختلف بازشناسی الگو داشت [۴۴] . شبکه‌ی عصبی کانوولوشنال در وظایف مختلف بینایی ماشین نیز نتایج بسیار موفقی داشته است.

شبکه‌ی عصبی کانوولوشنال سه نوع لایه‌ی اصلی دارد. لایه‌ی کانوولوشنال[۱۱] یک آشکارساز ویژگی است و به‌صورت خودکار یاد می‌گیرد و با استفاده از کرنل کانوولوشنال اطلاعات غیرلازم از یک ورودی را فیلتر می‌کند. با استفاده از کرنل‌های مختلف در لایه‌ی کانوولوشنال، می‌توان نقشه‌های ویژگی مختلفی از تصویر ورودی به‌دست آورد.

لایه‌ی تلفیق[۱۲] مقدار بیشینه و یا مقدار متوسط ویژگی‌های خاص بر روی یک ناحیه از داده‌های ورودی را محاسبه می‌کند. درواقع وظیفه‌ی لایه‌ی تلفیق، کاهش ابعاد (طول×عرض) حجم ورودی برای لایه‌ی کانوولوشنال بعدی است. لایه‌ی تلفیق عمق را تغییر نمی‌دهد. اعمال لایه‌ی تلفیق منجر به کاهش اندازه و درنتیجه، ازدست رفتن میزانی از اطلاعات می‌شود. البته این ازدست دادن اطلاعات برای شبکه مفید است چراکه کاهش اندازه باعث کاهش سربار محاسباتی می‌شود و با بیش‌برازش[۱۳] نیز مقابله می‌کند.

[8] dropout
[9] batch normalization
[10] data augmentation
[11] convolutional layer
[12] pooling layer
[13] overfitting

در شبکه‌ی عصبی کانوولوشنال پس از قرارگیری چندین لایه‌ی کانوولوشنال و چندین لایه‌ی تلفیق، نتیجه‌گیری نهایی شبکه توسط یک لایه‌ی تماماً متصل[۱۴] انجام می‌گیرد. نورون‌ها در یک لایه‌ی تماماً متصل، اتصالات کاملی با تمام فعالیت‌ها در لایه‌ی قبلی دارند. لایه‌های تماماً متصل نقشه‌ی ویژگی دوبعدی را به بردار ویژگی یک‌بعدی تبدیل می‌کنند. بردار به‌دست آمده می‌تواند به‌عنوان بردار ویژگی برای پردازش‌های بعدی استفاده شود و یا وظیفه‌ی طبقه‌بندی را انجام دهد. در هر شبکه‌ی عصبی کانوولوشنال دو مرحله برای آموزش وجود دارد. مرحله‌ی حرکت رو به جلو و مرحله‌ی پس‌انتشار. در مرحله‌ی اول تصویر ورودی به شبکه تغذیه می‌شود و این عمل چیزی جز ضرب نقطه‌ای بین ورودی و پارامترهای هر نورون و نهایتاً اعمال عملیات کانولوشن در هر لایه نیست. سپس خروجی شبکه محاسبه می‌شود. در هنگام ساخت نقشه‌ی ویژگی، کل تصویر با یک واحدی که حالت‌های آن در مکان متناظر در نقشه‌ی ویژگی ذخیره می‌شوند، اسکن می‌شود. این ساخت نقشه‌ی ویژگی معادل عملیات کانوولوشن است که توسط یک بایاس جمع‌پذیر و یک تابع سیگموییدِ[۱۵] دنبال می‌شود.

$$y^{(d)} = \sigma(Wy^{(d-1)} + b), \tag{۱.۲}$$

d عمق لایه‌ی کانوولوشنال است. W ماتریس وزن و b بایاس است. برای شبکه‌های عصبی تماماً متصل، مولفه‌های ماتریس وزن، غیرصفر هستند. اما برای شبکه‌های عصبی کانوولوشنال ماتریس وزن بسیار تنک[۱۶] است. به کارگیری ماتریس وزن تنک، پارامترهای قابل تنظیم شبکه را کاهش می‌دهد و درنتیجه قابلیت تعمیم‌پذیری شبکه افزایش می‌یابد.

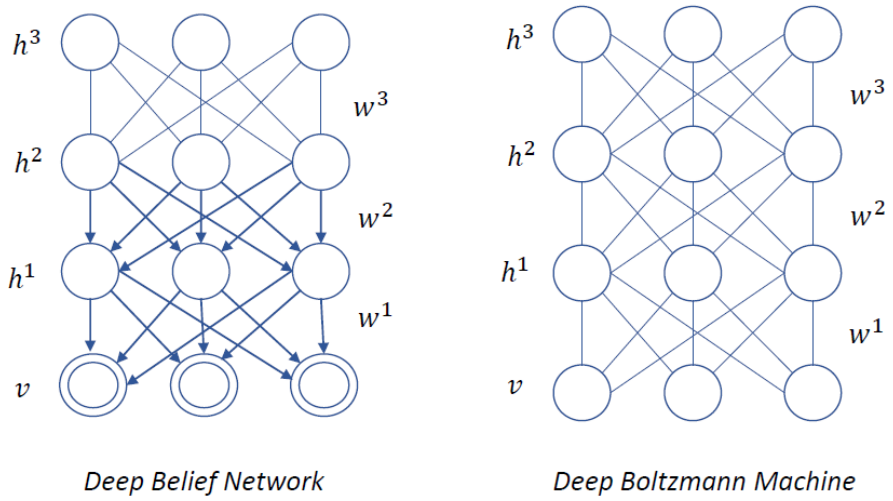
۲.۱.۲.۲ شبکه‌ی باور عمیق و ماشین بولتزمن عمیق

شبکه‌ی باور عمیق و ماشین بولتزمن عمیق مدل‌هایی از یادگیری عمیق می‌باشند که از ماشین بولتزمن محدود (RBM) به‌عنوان ماژول یادگیری استفاده می‌کنند. ماشین بولتزمن محدود یک شبکه‌ی عصبی تصادفی مولد[۱۷] است که می‌تواند توزیع احتمالاتی را روی مجموعه ورودی‌های خود یاد بگیرد. RBM یک شبکه‌ی عصبی کم‌عمق است که دارای دو لایه می‌باشد. در این شبکه هر گره به تمام گره‌های لایه‌ی مجاور متصل می‌شود. منظور

[14] fully connected layer
[15] sigmoid
[16] sparse
[17] generative stochastic neural network

از محدودیت در عنوان این شبکه این است که هیچ دو گره‌ای از یک لایه به‌هم متصل نیستند. این محدودیت باعث الگوریتم‌های آموزشی بهینه‌تر به‌ویژه الگوریتم واگرایی متقابل برپایه‌ی گرادیان[۱۸] [۶] می‌شود.

در شبکه‌های باور عمیق دو لایه‌ی بالایی یک گراف بدون جهت را شکل می‌دهند اما سایر لایه‌ها یک شبکه‌ی باور با اتصالات جهت‌دار و بالا به پایین را می‌سازد. ماشین‌های بولتزمن عمیق اتصالات بدون جهت بین تمام لایه‌های شبکه دارند. در شکل ۲.۲ ساختار DBM و DBN نمایش داده شده است.



شکل ۲.۲: ساختار ماشین بولتزمن عمیق (DBM) و شبکه‌ی باور عمیق (DBN)

- **شبکه‌ی باور عمیق**

شبکه‌ی باور عمیق مدل مولد احتمالاتی است که یک توزیع احتمالاتی مشترک را روی داده‌های قابل مشاهده و برچسب‌ها فراهم می‌کند. DBN از پشته کردن تعدادی RBM ایجاد می‌شود و آموزش آن‌ها به روش حریصانه انجام می‌گیرد [۳۳] . شبکه‌ی باور عمیق درابتدا یک استراتژی یادگیری بهینه‌ی لایه‌به‌لایه‌ی حریصانه را به کار می‌گیرد و در ادامه تمام وزن‌ها را با خروجی‌های متناظر تنظیم می‌کند. شبکه‌های باور عمیق مدل‌های گرافیکی هستند که یاد می‌گیرند که یک بازنمایی عمیق سلسله مراتبی از داده‌های آموزشی استخراج کنند.

- **ماشین بولتزمن عمیق**

ماشین بولتزمن عمیق یکی دیگر از مدل‌های یادگیری عمیق است که از ماشین بولتزمن محدود به‌عنوان

[18] gradient-based contrastive divergence algorithm

ماژول سازنده‌ی آن استفاده می‌شود. تفاوت ماشین بولتزمن عمیق با شبکه‌ی باور عمیق در این است که در DBN دو لایه‌ی بالایی گراف بدون جهت هستند درحالی‌که باقی لایه‌ها گراف جهت‌دار می‌باشند، اما در DBM تمامی اتصالات بدون جهت است. ماشین‌های بولتزمن عمیق چندین لایه از واحدهای پنهان دارند به‌نحوی‌که واحدها در لایه‌های با شماره‌ی فرد از واحدها در لایه‌های با شماره‌ی زوج مستقل شرطی[۱۹] می‌باشند و برعکس. درنتیجه فرآیند استنتاج در ماشین بولتزمن عمیق دشوار است. در فرآیند آموزش شبکه، یک ماشین بولتزمن عمیق مشترکاً تمامی لایه‌های یک مدل بدون نظارت خاص را آموزش می‌دهد و به‌جای بیشینه‌کردن درست‌نمایی[۲۰] به‌طور مستقیم، از الگوریتم برپایه‌ی درست‌نمایی بیشینه تصادفی (SML)[۹۷] به‌منظور بیشینه کردن مرز پایینی درست‌نمایی استفاده می‌کند. از آنجایی که این فرآیند امکان گرفتار شدن در کمینه‌ی محلی را دارد، یک استراتژی آموزش لایه به لایه‌ی حریصانه در [۷۱] ارائه شده است که در آن پیش‌آموزش لایه‌های DBM شبیه به DBN انجام می‌شود.

ماشین‌های بولتزمن عمیق توانایی استخراج لایه‌های متعددی از بازنمایی‌های پیچیده از داده‌های ورودی را دارند و به این دلیل که توانایی آموزش روی داده‌های بدون برچسب را دارند، برای یادگیری بدون نظارت بسیار مناسب می‌باشند. اما امکان استفاده از ماشین بولتزمن عمیق برای یادگیری بانظارت نیز وجود دارد. ماشین بولتزمن عمیق برای یادگیری مدل روی داده‌های ناهمگون که از ماهیت‌های مختلف آمده‌اند مناسب است. یکی از مهم‌ترین نقاط ضعف ماشین‌های بولتزمن عمیق هزینه‌ی محاسباتی بسیار بالای استنتاج در آن است به‌ویژه هنگامی که اندازه‌ی مجموعه داده بزرگ باشد. روش‌های مختلفی برای بهبود کارآمدی DBM ارائه شده است [۷۱، ۳۴، ۶۲، ۱۲].

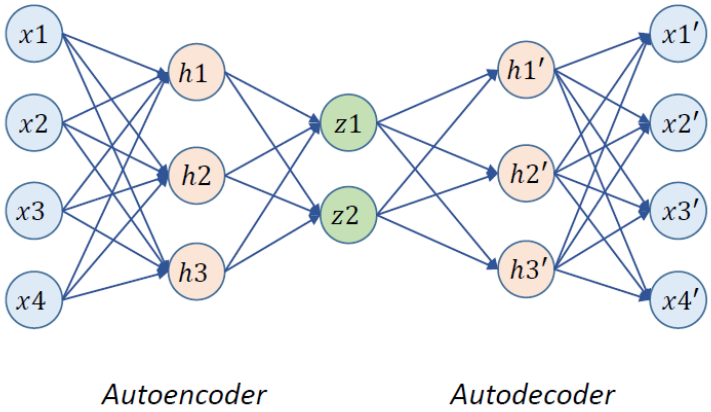
۳.۱.۲.۲ خودکدگذار پشته‌گذاری شده

خودکدگذارهای پشته‌گذاری شده[۲۱] از خودکدگذارها به‌عنوان بلاک‌های سازنده استفاده می‌کنند، شبیه به روشی که شبکه‌های باور عمیق از ماشین بولتزمن محدود به‌عنوان بلاک تشکیل‌دهنده بهره می‌برد. یک خودکدگذار با هدف کدگذاری ورودی x به بازنمایی $r(x)$ به‌نحوی‌که ورودی از $r(x)$ قابل بازسازی باشد، آموزش می‌بیند [۴]. درنتیجه خروجی هدف یک خودکدگذار همان ورودی خودکدگذار می‌باشد. بنابراین ابعاد ورودی و خروجی در

[19] conditionaly independent
[20] likelihood
[21] stacked autoencoder

خودکدگذار یکسان است.

خودکدگذار یک لایه‌ی درونی دارد که کد استفاده شده برای بازنمایی داده‌ی ورودی را توصیف می‌کند. این لایه از دو بخش اصلی تشکیل شده است: یک کدگذار برای نگاشت داده‌ی ورودی به کد متناظر و یک کدگشا به‌منظور نگاشت کد به یک بازسازی از داده‌ی ورودی اصلی. طی فرآیند آموزش، خطای بازسازی[۲۲] کمینه می‌شود و پارامترهای مدل با کمینه شدن خطای بازسازی متوسط، بهینه می‌شوند. در شکل ۳.۲ ساختار کلی خودکدگذار نمایش داده شده است. خودکدگذارها در کاهش ابعاد و یا یادگیری ویژگی‌ها کاربرد دارند. در سال‌های اخیر



شکل ۳.۲: ساختار کلی خودکدگذار

استفاده از خودکدگذارها به‌منظور یادگیری مدل‌های مولد از داده‌ها، رنگی‌کردن تصاویر و بهبود رزولوشن تصاویر نیز متداول است.

۴.۱.۲.۲ شبکه مولد تخاصمی

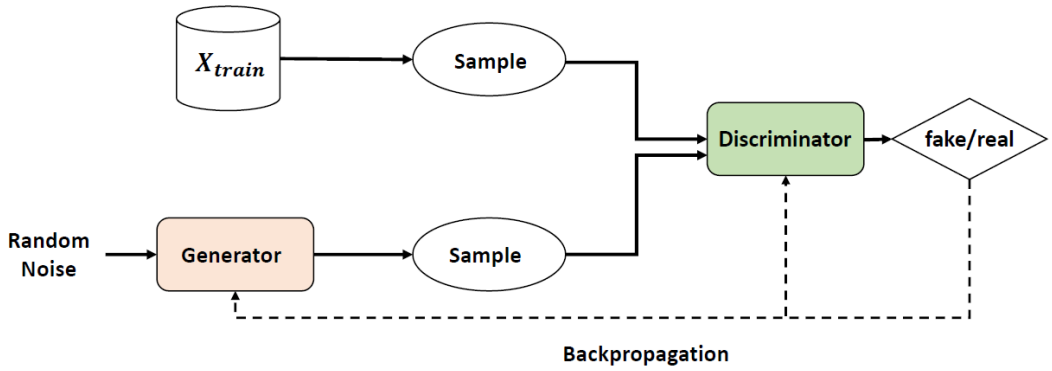
شبکه‌های مولد تخاصمی (GAN) در سال ۲۰۱۴ توسط Goodfellow معرفی شدند [۲۵]. در سال ۲۰۱۶ Yann LeCun از شبکه‌های مولد تخاصمی به‌عنوان «هیجان‌انگیزترین ایده‌ی مطرح شده در ۲۰ سال اخیر یادگیری ماشین» یاد می‌کند.

مدل‌سازی مولد[۲۳] یک وظیفه‌ی بدون نظارت در یادگیری ماشین است که شامل یادگیری الگوهای داده‌ی ورودی است به‌نحوی‌که مدل بتواند خروجی‌های قابل‌قبولی از مجموعه‌داده‌ی اصلی تولید کند. به عبارت دیگر

[22] reconstruction error
[23] generative modeling

یک مدل مولد، یک مدل از احتمال مشروط مشاهده‌ی X است با داشتن هدف Y. قبل ازمعرفی شبکه‌های مولد تخاصمی، مدل‌های مولد متعددی وجود داشتند؛ برای مثال مدل پنهان مارکوف، خانواده‌ی ماشین‌های بولتزمن مثل ماشین بولتزمن عمیق (DBM) ، ماشین بولتزمن محدود (RBM) و شبکه‌ی باور عمیق (DBN) و خودکدگذارها نمونه‌هایی از مدل‌های مولد می‌باشند.

شبکه‌های مولد تخاصمی یک راه هوشمندانه برای آموزش یک مدل مولد می‌باشد. این مدل شامل دو زیرشبکه است: یک شبکه‌ی مولد[۲۴] و یک شبکه‌ی ممیز[۲۵] . فرآیند آموزش در شبکه‌ی مولد تخاصمی به‌صورت یک بازی minimax بین شبکه‌ی مولد و شبکه‌ی ممیز پیاده‌سازی می‌شود. هردو شبکه سعی دارند که سود خود را بیشینه کنند درحالی‌که افزایش سود هرکدام منجر به کاهش سود دیگری می‌شود. شبکه‌ی مولد سعی می‌کند با یادگیری توزیع آماری داده‌های آموزشی، داده‌های جدیدی تولید کند که به‌شکل قابل قبولی شبیه به داده‌های آموزشی اصلی باشد. شبکه‌ی ممیز سعی می‌کند که داده‌های واقعی را از داده‌های تولید شده توسط شبکه‌ی مولد تمییز دهد. ساختار کلی شبکه‌ی مولد تخاصمی در شکل ۴.۲ نمایش داده شده است.



شکل ۴.۲: ساختار کلی شبکه‌ی مولد تخاصمی

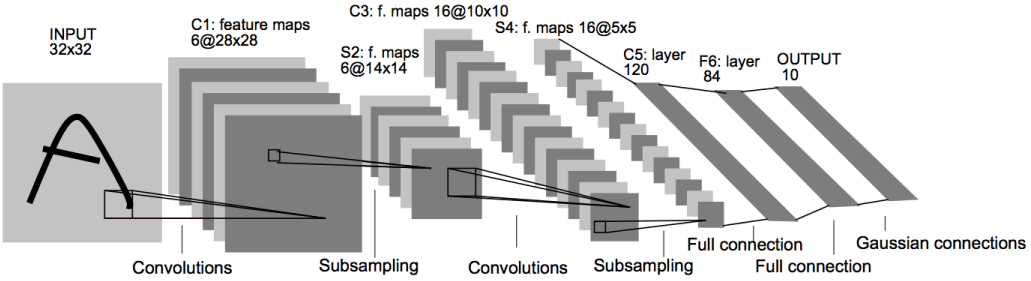
۲.۲.۲ انواع معماری‌های شبکه عصبی کانوولوشنال

تاکنون معماری‌های متفاوتی برای شبکه‌های عصبی کانوولوشنال ارائه شده است. اکثر این معماری‌ها اصول طراحی کلی ثابتی دارند، به‌نحوی‌که با اعمال لایه‌ی کانوولوشنال و لایه‌ی تلفیق روی داده‌ی ورودی از ابعاد ورودی می‌کاهند و نقشه‌ی ویژگی تولید می‌کنند. معماری‌های کلاسیک شبکه‌های عصبی کانوولوشنال اغلب از قرار

[24] generatore
[25] discriminator

گرفتن متوالی لایه‌های کانوولوشنال تشکیل می‌شود، درحالی‌که معماری‌های جدیدتر به دنبال راه‌های جدید و خلاقانه برای ایجاد یک شبکه با قابلیت یادگیری بهتر می‌باشند. اغلب این معماری‌ها از یک واحد تکرارشونده در شبکه استفاده می‌کنند.

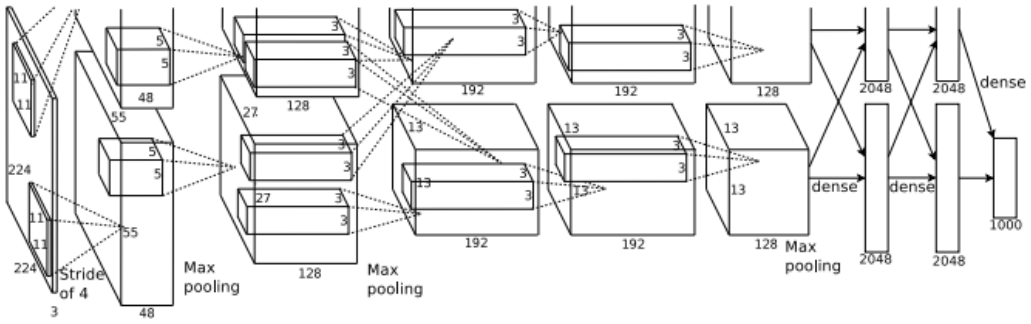
LeNet-5: این مدل در سال ۱۹۹۸ توسط Yann Lecun به‌منظور تشخیص ارقام دست‌نویس کد پستی معرفی شد. معماری ارائه شده در این مدل، معرفی‌کننده‌ی شبکه‌های عصبی کانوولوشنال می‌باشد. در این شبکه یک لایه‌ی کانوولوشنال جهت عمل کانوولوشن روی تصویر ورودی اعمال می‌شود. به‌تدریج ابعاد ورودی کاهش پیدا می‌کند. این کاهش ابعاد توسط لایه‌ی تلفیق میانگین[۲۶] انجام می‌شود. در نهایت لایه‌های تماماً متصلی وجود دارند که وظیفه‌ی طبقه‌بندی را انجام می‌دهند. این مدل ۶۰۰۰۰ پارامتر دارد [۴۵]. در شکل ۵.۲ ساختار معماری این شبکه نمایش داده شده است.



شکل ۵.۲: ساختار معماری LeNet5 [۴۵]

AlexNet: شبکه‌ی AlexNet در سال ۲۰۱۲ برای رقابت ImageNet توسط Alex Krizhevsky طراحی شد. معماری این شبکه شباهت زیادی به شبکه‌ی LeNet دارد با این تفاوت که بسیار بزرگ‌تر است و پارامترهای بیشتری دارد. شبکه‌ی AlexNet نسبت به LeNet بسیار عمیق‌تر است و در هر لایه فیلترهای بیشتری دارد. همچنین از لایه‌های کانوولوشنال پشته‌گذاری شده[۲۷] در آن استفاده شده است. این شبکه شامل کانوولوشن‌های ۱۱ × ۱۱، ۵ × ۵ و ۳ × ۳ می‌باشد. همچنین لایه‌ی تلفیق بیشینه[۲۸]، برون‌اندازی، تابع فعالسازی ReLU و SGD با مومنتوم در این معماری استفاده شده است. AlexNet ۶۰ میلیون پارامتر دارد [۴۲]. در شکل ۶.۲ ساختار معماری این شبکه نمایش داده شده است.

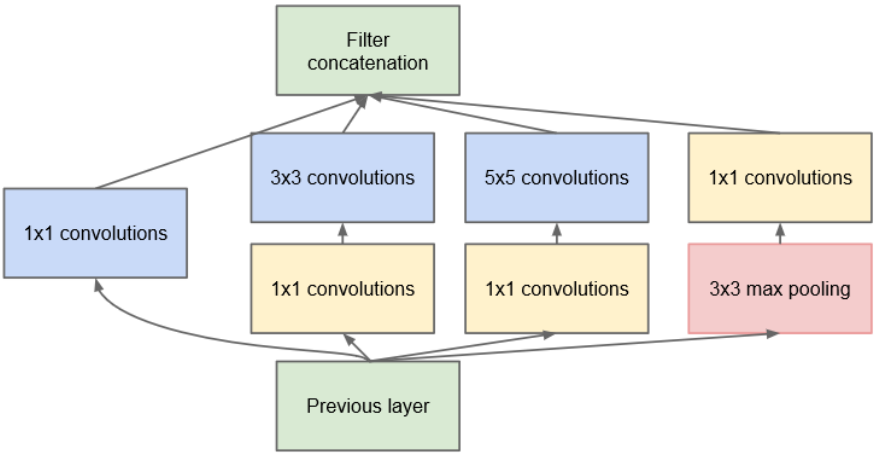
[26] average pooling
[27] stacked convolutional layers
[28] max pooling



شکل ۶.۲: ساختار معماری AlexNet [۴۲]

VGG-16: شبکه‌ی VGG-16 در سال ۲۰۱۴ معرفی شد. تعداد پارامترهای این شبکه ۱۳۸ میلیون است. شبکه‌ی VGG-16 بهبودیافته‌ی شبکه‌ی AlexNet می‌باشد. در VGG-16 فیلترهای با هسته‌ی بزرگ (لایه‌ی اول ۱۱ × ۱۱ و لایه‌ی دوم ۵ × ۵) با چند فیلتر با اندازه‌ی هسته‌ی ۳ × ۳ جایگزین می‌شوند [۷۴].

Inception(GoogLeNet): شبکه‌ی Inception [۸۰] توسط پژوهشگران گوگل در سال ۲۰۱۴ ارائه شد و توانست برنده‌ی رقابت ImageNet در سال ۲۰۱۴ شود. این مدل از واحد پایه‌ای با عنوان سلول Inception تشکیل شده است. در این ماژول دنباله‌ای از کانوولوشن‌ها در مقیاس‌های متفاوت اعمال می‌شود و متعاقباً نتایج آن‌ها تجمیع می‌شوند. در شکل ۷.۲ نمونه ای از یک سلول Inception نمایش داده شده است.



شکل ۷.۲: ساختار یک سلول Inception [۸۰]

به‌منظور صرفه جویی در محاسبات، کانوولوشن‌های ۱ × ۱ برای کاهش عرض کانال ورودی استفاده می‌شوند. در هر سلول، مجموعه‌ای از فیلترهای ۱ × ۱، ۳ × ۳ و ۵ × ۵ اعمال می‌شوند تا ویژگی‌هایی در مقیاس‌های

مختلف را از داده‌ی ورودی یاد بگیرند.

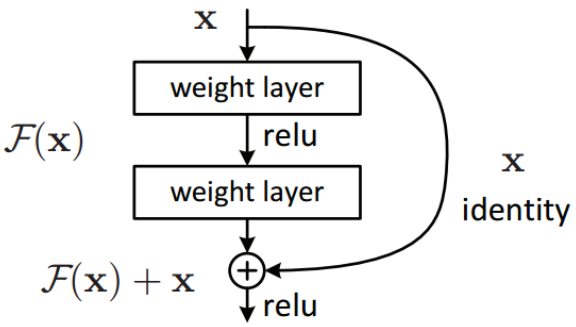
در سال ۲۰۱۶ محققان مقاله‌ی دیگری را ارائه کردند که جایگزین کارآمدتری نسبت به سلول Inception اولیه در آن معرفی شده بود [۸۱] . لایه‌های کانوولوشن با اندازه‌ی فیلترهای بزرگ مثل ۵ × ۵ و ۷ × ۷ از لحاظ قدرت استخراج ویژگی در ابعاد بزرگ‌تر بسیار سودمند می‌باشند اما هزینه‌ی محاسباتی بالایی دارند. پژوهشگران به‌جای استفاده از فیلتر با اندازه‌ی ۵ × ۵، استفاده از دو فیلتر پشته‌گذاری شده‌ی ۳ × ۳ را پیشنهاد کردند. همچنین کانوولوشن ۳ × ۳ می‌تواند به کانوولوشن‌های متوالی ۳ × ۱ و ۱ × ۳ تجزیه شود. به‌منظور بهبود عملکرد شبکه، دو خروجی اضافی در شبکه درنظر گرفته می‌شود. اضافه کردن این خروجی‌های اضافی همانند یک تکنیک منظم‌سازی عمل می‌کند و باعث همگرا شدن بهتر مدل می‌شود.

ResNet: تا قبل از معرفی معماری ResNet [۲۹] این باور وجود داشت که با افزایش تعداد لایه‌های یک شبکه می‌توان به دقت و عملکرد بهتری دست‌یافت. اما با افزایش بیش از حد تعداد لایه‌ها، مشکل ناپدیدشدن گرادیان[۲۹] به وجودآمده و دقت مدل به حالت اشباع می‌رسد و دقت افزایش نمی‌یابد. پژوهشگران در ارائه‌ی مدل ResNet با معرفی اتصالات جهشی[۳۰] به این چالش پرداخته‌اند. در این شبکه ارتباطاتی خارج از ساختار کانوولوشنال بین لایه ها درنظر گرفته شده تا ورودی های لایه قبلی را بدون واسطه به لایه‌ی بعدی منتقل کند و در مرحله انتشار پس‌رو[۳۱] یا اصلاح شبکه، خطای هر لایه را به لایه‌ی قبلی انتقال دهد تا بتوان شبکه را عمیق‌تر کرد و آن را سریع‌تر آموزش داد. به این ارتباطات، اتصالات جهشی و به ساختار حاصل از آن بلاک باقی‌مانده[۳۲] می‌گویند. با انتقال مقادیر از لایه‌های قبل به لایه‌های بالایی، مشکل گرادیان ناپدید شده حل می‌شود و می‌توان شبکه‌های عمیق‌تری را آموزش داد. همچنین اتصالات جهشی به مدل امکان یادگیری یک تابع هویتی[۳۳] را می‌دهند که تضمین می‌کند لایه‌های بالایی حداقل به اندازه‌ی لایه‌های پایینی خوب عمل می‌کنند.

اگر ورودی یک بلاک شبکه‌ی عصبی x درنظر گرفته شود، هدف یادگیری توزیع صحیح $H(x)$ می‌باشد. تفاضل بین ورودی و خروجی را می‌توان به‌صورت $F(x) = H(x) - x$ تعریف کرد. بنابراین می‌توان گفت $H(x) = F(x)+x$. به‌طور خلاصه می‌توان گفت که در شبکه‌های سنتی، لایه‌ها خروجی صحیح $(H(x))$ را یاد می‌گیرند درحالی‌که در شبکه‌ی باقی‌مانده، لایه‌ها باقی‌مانده $(F(x))$ را یاد می‌گیرند. تصویر یک بلاک باقیمانده‌ی ساده در شکل ۸.۲ نمایش داده شده است. شبکه‌ی ResNet جزء اولین معماری‌هایی است که از نرمال‌سازی دسته‌ای

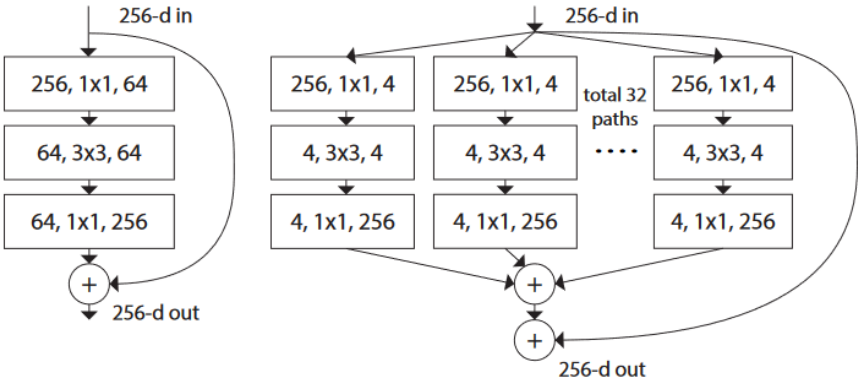
[29] vanishing gradient
[30] skip connections
[31] back propagation
[32] residual block
[33] identity function

استفاده کرده است. این معماری با ۱۵۲ لایه برنده‌ی مسابقه‌ی ImageNet در سال ۲۰۱۵ می‌باشد.



شکل ۸.۲: ساختار یک بلاک باقی‌مانده‌ی ساده [۲۹]

یکی از گسترش‌های ResNet شبکه‌ای با عنوان ResNeXt [۹۱] می‌باشد که در سال ۲۰۱۷ معرفی شده است. تفاوت ResNeXt با ResNet در وجود تعدادی انشعاب یا مسیر موازی داخل بلاک باقی‌مانده می‌باشد. درواقع در ResNeXt به‌جای اعمال کانوولوشنال به تمام نقشه‌ی ویژگی ورودی، ورودی یک بلاک به دنباله‌هایی با ابعاد (تعداد کانال) بازنمایی کمتر تبدیل شده و چند فیلتر کانوولوشنال قبل از ادغام نتیجه به‌طور جداگانه روی آن‌ها اعمال می‌شود. ResNeXt-50 دارای ۲۵ میلیون پارامتر است درحالی‌که ResNet-50 ۲۵٫۵ میلیون پارامتر دارد. در شکل ۹.۲ مقایسه‌ی بین یک بلاک از معماری ResNet و یک بلاک از معماری ResNeXt نمایش داده شده است.



شکل ۹.۲: سمت چپ: یک بلاک از ResNet، سمت راست: یک بلاک از ResNeXt [۹۱].

یکی دیگر از گسترش‌های ResNet ، معماری WideResNet [۹۹] می‌باشد. هدف معماری ResNet ارائه‌ی یک معماری عمیق است که موجب ناپدیدشدن گرادیان نشود و شبکه باوجود عمق زیاد، کارآیی خوبی

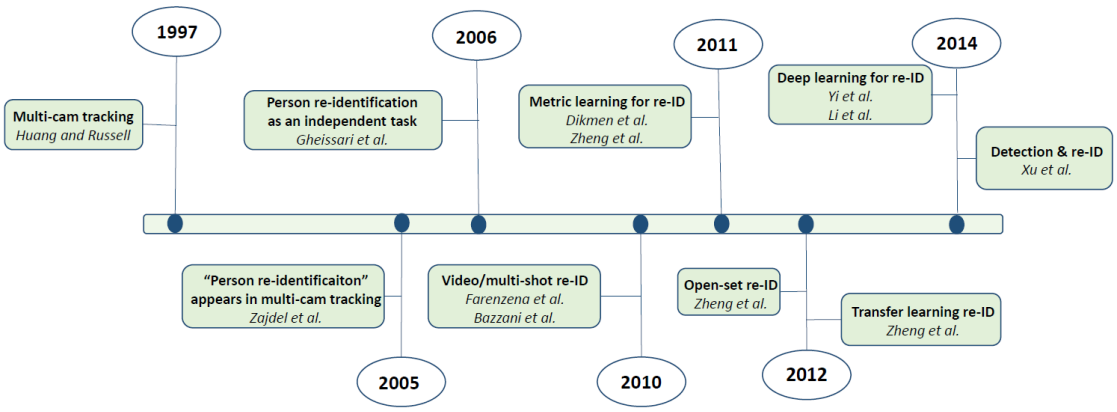
داشته باشد. در WideResNet پژوهشگران معتقدند که افزایش عرض شبکه (عمق کانال‌ها) روشی بهینه برای بالا رفتن ظرفیت شبکه می‌باشد.

EfficientNet: مدل EfficientNet [۸۲] در سال ۲۰۱۹ برنده‌ی رقابت ImageNet شد. درواقع EfficientNet گروهی از مدل‌های کانوولوشنال است که دارای نمونه‌های $B0$ تا $B7$ می‌باشد. در EfficientNet علاوه‌بر تمرکز بر افزایش دقت مدل، روی کارآمدی و بهینه بودن مدل نیز تمرکز شده است. به‌طور کلی سه بعد مقیاس‌بندی در CNN وجود دارد. بعد عمق به معنای تعداد لایه‌های شبکه است. بعد عرض به معنای میزان پهنای شبکه است. برای مثال یکی از معیارهای عریض بودن، تعداد کانال‌های یک لایه‌ی کانوولوشنال می‌باشد. بعد رزولوشن نیز به معنای رزولوشن تصویری است که به شبکه داده می‌شود. در معماری EfficientNet یک تکنیک مقیاس‌بندی بهینه ارائه شده است به‌نحوی‌که به‌صورت یکنواخت عرض، عمق و رزولوشن را مقیاس‌بندی می‌کند.

۳.۲ بازشناسایی شخص

۱.۳.۲ تاریخچه‌ی بازشناسایی شخص

امروزه مسئله‌ی بازشناسایی شخص، یکی از مسائل پرطرفدار در حوزه‌ی بینایی ماشین محسوب می‌شود. باتوجه‌به گسترش مجموعه‌داده‌های این حوزه و همچنین به کارگیری تکنیک‌های یادگیری عمیق، نتایج موفقیت‌آمیزی در حوزه‌ی بازشناسایی شخص به‌دست آمده است. در شکل ۱۰.۲ مهم‌ترین رویدادهایی که منجر به پیدایش و گسترش بازشناسایی شخص شده، نمایش داده شده‌اند [۴۶].



شکل ۱۰.۲: تعدادی از رویدادهای مهم در پیدایش و گسترش پژوهش‌های بازشناسایی شخص

پژوهش‌ها در حوزه‌ی بازشناسایی شخص با مسئله‌ی ردیابی چند-دوربینی[۳۴] شروع شد. در سال ۱۹۹۷، Huang و Russell [۳۷] یک فرمول بیزی را ارائه کردند که احتمال پسین مشاهده‌ی یک شی داده شده از یک دوربین، توسط دوربین‌های دیگر را محاسبه می‌کرد. در آن زمان مسئله‌ی بازشناسایی شخص به‌عنوان یک مسئله‌ی مستقل مطرح نبود. در سال ۲۰۰۵، Wojciech Zajdel و همکارانش برای اولین بار در مقاله‌ی خود [۱۰۰] از عبارت بازشناسایی شخص استفاده کرده و در مسئله‌ی ردیابی چند-دوربینی به بازشناسایی شخص پرداختند. در روش ارائه شده در آن مقاله، یک برچسب منحصر بفرد برای هر شخص درنظر گرفته می‌شد و یک شبکه‌ی بیزی پویا برای رمزگذاری رابطه‌ی احتمالاتی بین برچسب‌ها و ویژگی‌ها در مسیر ردیابی طراحی شده بود. هویت شخص ورودی با استفاده از الگوریتم استنتاج بیزی تقریبی، تعیین می‌شد. در سال ۲۰۰۶، Gheissari در مقاله‌ی خود [۲۴] روی مسئله‌ی بازشناسایی شخص به‌صورت مستقل کار کرد. در سال ۲۰۱۰، دو مقاله‌ی [۳] و [۲۲] برای بازشناسایی شخص چند-شاتی ارائه شدند. هر دو کار از ویژگی رنگ استفاده می‌کردند. در [۲۲] یک مدل بخش‌بندی برای تشخیص تصویر شخص، به کار رفته است. در مقاله‌ی [۳] از فاصله‌ی بهاتاچاریا[۳۵] برای محاسبه‌ی فاصله‌ی تصاویر استفاده می‌شود. پژوهشگران دریافتند که استفاده از چندین فریم از هر شخص، نسبت به روش‌های تک فریمی عملکرد بهتری دارد.

در سال‌های ۲۰۱۰ و ۲۰۱۱، مقالات [۱۸] و [۱۰۵] برای مسئله‌ی بازشناسایی شخص از یادگیری معیار استفاده کردند به‌نحوی‌که با بیشینه کردن احتمال کمتر بودن فاصله‌ی زوج‌های منطبق صحیح نسبت به فاصله‌ی زوج‌های منطبق ناصحیح، معیار فاصله‌ی بهینه را یاد بگیرند. در سال ۲۰۱۲، اولین مقاله [۱۰۶] در حوزه‌ی بازشناسایی

[34] multi-camera tracking
[35] Bhattacharyya distance

شخص مجموعه-باز[۳۶] ارائه شد. در بازشناسایی شخص مجموعه-باز، هویت تصویر ورودی امکان دارد در میان هویت‌های تصاویر موجود در گالری نباشد.

مسئله‌ی تشخیص عابرین پیاده در تصاویر موضوع مهمی می‌باشد. در اکثر کارها از تصاویری که به‌صورت دستی محدود شده و برش داده شده‌اند استفاده می‌شود. در سال ۲۰۱۴، مقاله‌ی [۹۲] با ترکیب آشکارسازی و بازشناسایی به این موضوع پرداخته است. در سال ۲۰۱۴ مقالات [۹۵] و [۴۹] برای اولین بار از یادگیری عمیق برای حل مسئله‌ی بازشناسایی شخص استفاده کردند. این دو مقاله از شبکه‌ی عصبی سیامی برای تعیین احتمال تعلق دو زوج تصویر ورودی به یک هویت واحد، بهره بردند. یکی از دلایل استفاده از شبکه‌ی سیامی برای این هدف، این است که تعداد نمونه‌های آموزشی به ازای هر هویت محدود می‌باشد. این دو مقاله تفاوت‌هایی در پارامترها باهم دارند، همچنین مقاله‌ی [۹۵] یک تابع هزینه‌ی اضافی را در شبکه استفاده کرده درحالی‌که مقاله‌ی [۴۹] بخش‌بندی بدن بهتری را به کار گرفته است. از آنجایی که در مقالات [۹۵] و [۴۹] مجموعه‌داده‌های مورداستفاده مشابه نیستند، امکان مقایسه‌ی مستقیم عملکرد دو مقاله وجود ندارد.

در سیستم‌های سنتی بازشناسایی شخص اغلب دو گام اصلی وجود دارد. درمرحله‌ی اول، یک تصویر به‌عنوان ورودی دریافت می‌شود و با استفاده از روش‌های استخراج ویژگی دستی[۳۷] مثل هیستوگرام رنگی، سعی می‌شود که ویژگی‌های متمایزدهنده‌ی تصویر استخراج شود و یک بردار ویژگی از آن تصویر به‌دست آید. پس از به‌دست آمدن بردار ویژگی، هدف سیستم بازشناسایی این است که تصویر شخص مورد نظر را از میان تمام تصاویر موجود در گالری تصاویر که توسط دوربین‌های مختلف ثبت شده‌اند، پیدا کند. این هدف با محاسبه‌ی فاصله‌ی بین تصویر ورودی و تمامی تصاویر موجود در گالری محقق می‌شود. سپس تصاویری از گالری که کمترین فاصله را با تصویر ورودی دارند، در خروجی برگردانده می‌شوند.

با گسترش یادگیری عمیق و استفاده از تکنیک‌های یادگیری عمیق در مسئله‌ی بازشناسایی شخص، پیشرفت‌های بسیار زیادی در این حوزه رخ داده است. در جدول ۱.۲ نتایج تعدادی از موفق‌ترین مقالات روی دو مجموعه‌داده‌ی معروف حوزه‌ی بازشناسایی شخص آورده شده است.

[36] open-set
[37] handcrafted

جدول ۱.۲: نتایج تعدادی از موفق‌ترین مقالات در حوزه‌ی بازشناسایی شخص روی مجموعه‌داده‌های Market1501 و DukeMTMC-reID

DukeMTMC-reID		Market1501			
mAP(%)	Rank-1(%)	mAP(%)	Rank-1(%)	Year	Method
47.13	67.68	66.07	83.97	2017	GAN [109]
53.50	72.44	69.14	84.92	2017	TriNet[31]
57.61	78.32	71.55	89.49	2018	CamStyle [114]
64.5	80.0	77.7	90.5	2018	FD-GAN[23]
69.2	83.3	81.6	93.8	2018	PCB[79]
73.5	88.6	84.9	94.8	2019	OSNet [116]
74.8	86.6	86	94.8	2019	DG-Net [107]
89.2	91.4	94.2	95.4	2019	Auto-ReID [67]
89.1	90.2	94.24	95.43	2019	BoT Baseline [55]
92.7	94.4	95.5	98.1	2019	st-ReID [86]
90.7	92.2	94.4	96	2020	AdaptiveReID [63]

۲.۳.۲ مدل‌های عمیق بازشناسایی شخص و کمبود داده‌های آموزشی

بازشناسایی شخص یکی از مسائل چالشی و پیچیده در حوزه‌ی بینایی ماشین می‌باشد. تصاویر مجموعه‌داده‌های این حوزه اغلب توسط دوربین‌هایی با کیفیت پایین و تحت شرایط انسداد و در شرایط روشنایی مختلف گرفته می‌شود. ازطرفی از آنجایی که در این تصاویر معمولاً چهره‌ی افراد با وضوح بالا مشخص نیست، نمی‌توان از ویژگی‌های چهره‌ی افراد برای بازشناسایی آن‌ها بهره‌ی کافی را برد. بنابراین استفاده از ویژگی‌های ظاهری افراد مثل جنس و رنگ لباس و ساختار بدنی افراد می‌تواند کمک‌کننده باشد. اما ویژگی‌های ظاهری نیز همیشه کارساز نمی‌باشند برای مثال ممکن است تعدادی از افراد لباس‌های مشابهی داشته باشند.

باوجود پیچیده بودن مسئله‌ی بازشناسایی شخص، با استفاده از تکنیک‌های یادگیری عمیق نتایج قابل قبولی در این حوزه به‌دست آمده است. اما همچنان چالش‌های مختلفی در مسئله‌ی بازشناسایی شخص وجود دارد. یکی از این چالش‌ها، کمبود داده‌های آموزشی می‌باشد. مدل‌های عمیق اغلب مدل‌های بسیار پیچیده با لایه‌های متعدد می‌باشند. یکی از ملزومات استفاده از مدل‌های عمیق و کارآمدی این مدل‌ها، در دسترس داشتن میزان

کافی داده‌های آموزشی است تا مدل دچار بیش‌برازش نشود.

اکثر مجموعه‌داده‌های بازشناسایی شخص تعداد تصاویر کمی از هر هویت دارند، برای مثال مجموعه‌داده‌ی VIPeR به ازای هر هویت، دو تصویر را داراست. در مجموعه‌داده‌های بزرگ مقیاس این حوزه مانند CUHK03، Market1501 و DukeMTMC-reID به‌طور متوسط به ازای هر هویت تعداد ۹/۶، ۱۷/۲ و ۲۳/۵ تصویر وجود دارد. این مسئله می‌تواند باعث شود که مدل عمیق بازشناسایی شخص در مرحله‌ی آزمایش دچار بیش‌برازش شود. برای مقابله با مشکل کمبود داده‌های آموزشی در مسئله‌ی بازشناسایی شخص تاکنون راه‌حل‌های متفاوتی ارائه شده است.

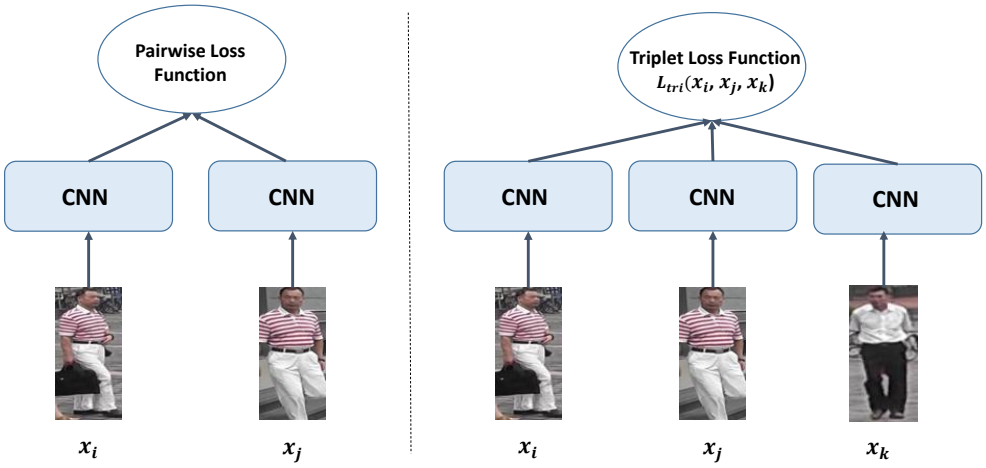
۱.۲.۳.۲ استفاده ازمعماری شبکه‌ی عصبی سیامی

یکی از این روش‌های مقابله با کمبود داده‌های آموزشی در مسئله‌ی بازشناسایی شخص استفاده از معماری شبکه‌های سیامی[۳۸] می‌باشد. شبکه‌های سیامی نخستین بار در دهه‌ی ۱۹۹۰ برای حل مسئله‌ی تصدیق امضا [۵]، معرفی شدند. شبکه‌های سیامی، شبکه‌های عصبی هستند که از دو و یا بیشتر زیرشبکه تشکیل شده‌اند. زیرشبکه‌ها کاملاً مشابه هستند و علاوه‌بر معماری آن‌ها، وزن‌ها و پارامترهایی که زیرشبکه‌ها به اشتراک می‌گذارند نیز یکسان می‌باشد. روش کار شبکه‌های سیامی به این صورت می‌باشد که زیرشبکه‌ها ورودی‌های متفاوتی را می‌پذیرند و در خروجی میزان شباهت ورودی‌های متفاوت را یاد می‌گیرند. اگر فرض شود که تعداد ۱۰ دسته‌ی مختلف در مسئله وجود دارد، روش کار شبکه‌ی سیامی به این صورت نیست که مستقیما میزان احتمال تعلق تصویر ورودی به تمامی ۱۰ کلاس خروجی را یاد بگیرد. در عوض شبکه‌ی سیامی تابع فاصله‌ای را یاد می‌گیرد که دو تصویر را به‌عنوان ورودی دریافت کرده و میزان شباهت آن تصاویر را بیان می‌کند.

معمولاً شبکه‌های سیامی از دو زیرشبکه و یا سه زیرشبکه تشکیل می‌شوند. در مدل‌های مبتنی بر زوج، یک زوج تصویر به‌عنوان ورودی به دو زیرشبکه داده می‌شود و میزان شباهت آن‌ها تعیین می‌گردد اما در مدل‌های سه‌گانه، سه تصویر به‌عنوان ورودی به سه زیرشبکه داده می‌شود. در مدل سیامی سه‌گانه یک تابع هدف برای آموزش مدل‌های شبکه استفاده می‌شود که یک حاشیه‌ای[۳۹] بین معیار فاصله‌ی زوج‌های مثبت و معیار فاصله‌ی زوج‌های منفی ایجاد می‌کند. ورودی‌های شبکه‌ی سیامی سه‌گانه، یک تصویر لنگر[۴۰]، یک تصویر مثبت (دارای

[38] siamese network
[39] margin
[40] anchor

هویت یکسان با تصویر لنگر) و یک تصویر منفی (دارای هویت متفاوت با تصویر لنگر) می‌باشد. نحوه‌ی انتخاب نمونه‌های مثبت و منفی در عملکرد شبکه بسیار مؤثر است. اگر نمونه‌های مثبت و منفی آسانی برای شبکه‌ی سیامی سه‌گانه انتخاب شود، عملکرد آن قابل قبول نخواهد بود. تابع اتلاف سه‌گانه در فضای ویژگی‌های یاد گرفته شده، فاصله‌ی بین نمونه‌های مثبت را کمینه و فاصله‌ی بین نمونه‌های منفی را بیشینه می‌کند و بدین‌ترتیب مدل آموزش می‌یابد. برای تعیین خروجی این مدل از شبکه‌های سیامی، یک لایه‌ی Softmax در بالای شبکه و روی هر دو خروجی فاصله‌ای، قرار داده می‌شود. در شکل ۱۱.۲ مثالی از شبکه‌های سیامی که به‌عنوان ورودی، یک زوج از داده‌ها و یک سه‌تایی از داده‌ها را می‌پذیرند، نمایش داده شده است.



شکل ۱۱.۲: سمت چپ: شبکه‌ی سیامی با ورودی زوج‌هایی از داده‌ها. سمت راست: شبکه‌ی سیامی با ورودی سه‌تایی‌هایی از داده‌ها

معماری شبکه‌ی عصبی سیامی نسبت به معماری استاندارد شبکه‌های عصبی دارای مزایایی است. یکی از این مزیت‌ها این است که شبکه‌ی عصبی سیامی می‌تواند به وسیله‌ی زوج‌هایی از تصاویر و یا سه‌تایی‌هایی از تصاویر آموزش داده شود و نیاز به برچسب کامل داده‌ها ندارد. مزیت بعدی این است که معمولاً شبکه‌های عصبی سیامی قابلیت تعمیم‌پذیری بالاتری در هنگام آزمایش دارند. اگر از هر کلاس تعداد نمونه‌های کمی موجود باشد عملکرد شبکه‌های سیامی مناسب می‌باشد. شبکه‌های عصبی سیامی به‌طور مداوم به‌روزرسانی می‌شوند

و زمانی‌که کلاس جدیدی به مجموعه‌ی آموزشی اضافه شود، شبکه‌ی سیامی می‌تواند مستقیما با تصویر کلاس جدید به‌روزرسانی شود. ولی این حالت برای مدل‌های آموزش داده شده با تابع خطای طبقه‌بندی امکان‌پذیر نمی‌باشد.

اکثر مجموعه‌داده‌های بازشناسایی شخص به ازای هر هویت تعداد نمونه‌های محدودی دارند و استفاده از شبکه‌های سیامی در مسئله‌ی بازشناسایی شخص رایج می‌باشد. در مقاله‌ی [۹۵] نویسندگان شبکه‌ی سیامی عمیق را برای یادگیری معیار طراحی کرده‌اند. معماری آن شبکه از سه شبکه‌ی کانوولوشنال مستقل تشکیل شده است که پارامترها را با یکدیگر به اشتراک گذاشته و روی سه بخش غیرهم‌پوشان از دو تصویر کار می‌کنند. در نهایت میزان شباهت دو تصویر توسط تابع شباهت کسینوسی محاسبه می‌گردد.

در مقاله‌ی [۸۳] یک شبکه‌ی کانوولوشنال سیامی به همراه یک تابع دریچه تطبیقی[۴۱] قابل یادگیری، که می‌تواند رفتار شبکه در فاز آموزش و فاز آزمایش را از هم جدا کند، طراحی شده است. این شبکه به وسیله‌ی تابع اتلاف هم‌سنجی[۴۲] بهینه‌سازی می‌شود.

در مقاله‌ی [۱۴] یک شبکه‌ی عصبی کانوولوشنال دو جریانه ارائه شده است که هر جریان آن یک شبکه‌ی سیامی می‌باشد. این معماری می‌تواند اطلاعات فضایی[۴۳] و زمانی[۴۴] را به‌صورت مجزا یاد بگیرد. درنهایت هم یک تابع هدف وجود دارد که نتایج دو شبکه‌ی سیامی را ترکیب کرده و هویت فرد را پیش‌بینی می‌کند.

مقاله‌ی [۱۰۴] یک مدل جدید عمیق سیامی مبتنی بر توجه[۴۵] ارائه کرده است. در این مقاله با استفاده از مکانیزم توجه، می‌توان دلیل تصمیم‌گیری شبکه‌ی سیامی برای پیش‌بینی را تا حدودی درک کرد.

در مقاله‌ی [۲۳] به چالش تغییر حالت قرارگیری افراد در مسئله‌ی بازشناسایی شخص پرداخته شده است. در این مقاله یک شبکه‌ی مولد تخاصمی برای یادگیری بازنمایی‌های مرتبط با هویت و غیروابسته به طرز قرارگیری، ارائه شده است. چارچوب ارائه شده در این مقاله برمبنای شبکه‌ی سیامی می‌باشد.

مقاله‌ی [۱۰] یک مدل سه‌گانه‌ی شبکه‌ی عصبی کانوولوشنال مبتنی بر بخش[۴۶] چند کاناله معرفی کرده است. در این مدل چندین کانال وجود دارند که ویژگی‌های سراسری تمام بدن و ویژگی‌های محلی بخش‌های بدن را یاد بگیرند. این مدل به‌منظور یادگیری از تابع اتلاف سه‌گانه استفاده می‌کند تا در فضای ویژگی‌های یاد گرفته شده،

[41] matching gate
[42] contrastive loss
[43] spatial
[44] temporal
[45] attention-driven
[46] parts-based

نمونه‌های متعلق به یک فرد را به یکدیگر نزدیک کرده و نمونه‌های متعلق به افراد متفاوت را از یکدیگر دور سازد.

۲.۲.۳.۲ روش‌های داده‌افزایی

استفاده از روش‌های داده‌افزایی نیز برای مقابله با مشکل کمبود داده‌های آموزشی می‌تواند به کار گرفته شود. روش‌های داده افزایی متفاوتی در مقالات برای مقابله با بیش‌برازش در اثر کمبود داده‌های آموزشی در مسئله‌ی بازشناسایی شخص ارائه شده است. روش‌های داده‌افزایی متداول مثل چرخاندن تصویر، جابه‌جایی تصویر و ... در اکثر پژوهش‌ها استفاده می‌شوند. روش داده‌افزایی پاک کردن تصادفی[۴۷] [۱۱۰] یکی از روش‌هایی است که در بسیاری از مقالات ازجمله [۵۵] و [۱۱۴] استفاده شده است. این روش یک روش داده‌افزایی برای شبکه‌های عصبی کانوولوشنال می‌باشد. در فاز آموزش، نواحی مستطیلی از تصویر به‌صورت تصادفی انتخاب شده و پیکسل‌های آن ناحیه با مقادیر تصادفی پاک می‌شود. در این فرآیند تصاویر آموزشی با سطوح مختلف انسداد تولید می‌شوند و ریسک بیش‌برازش کاهش می‌یابد. روش داده‌افزایی پاک کردن تصادفی باوجود ساده بودن می‌تواند در مسائل بازشناسایی شخص مؤثر باشد. در مقاله‌ی [۶۱] روش داده‌افزایی با تغییر پس‌زمینه‌ی تصاویر معرفی شده و نشان داده شده است که تولید تصاویر بیشتر از طریق تغییر پس‌زمینه‌ی تصویر در مسئله‌ی بازشناسایی شخص باعث بهبود عملکرد مدل می‌شود. یکی از موضوعات بسیار به‌روز در حوزه‌ی بازشناسایی شخص استفاده از داده‌های تولید شده توسط شبکه‌های مولد تخاصمی (GAN) می‌باشد که می‌تواند به عملکرد مدل‌های بازشناسایی شخص کمک زیادی کند.

۳.۲.۳.۲ داده‌افزایی با استفاده از شبکه‌های مولد تخاصمی

در سال‌های اخیر داده‌افزایی با استفاده از شبکه‌های مولد تخاصمی (GAN) به یکی از روش‌های قدرمند داده‌افزایی در مسائل مختلف تبدیل شده است. شبکه‌ی مولد تخاصمی که در سال ۲۰۱۴ توسط Goodfellow معرفی شده است یک موضوع بسیار پرطرفدار در حوزه‌ی یادگیری ماشینی می‌باشد. فرآیند آموزش در GAN به‌صورت یک بازی minimax بین دو شبکه‌ی مولد [۴۸] و ممیز [۴۹] انجام می‌شود. شبکه‌ی مولد سعی در تولید داده‌هایی مشابه با داده‌های آموزشی دارد به‌نحوی‌که شبکه‌ی ممیز قادر به تشخیص داده‌های واقعی از داده‌های تولید شده نباشد،

[47] random erasing data augmentation
[48] generator
[49] discriminator

ازطرفی شبکه‌ی ممیز سعی دارد بین داده‌های واقعی و داده‌های تولید شده تفاوت قائل شود. فرآیند آموزش تا جایی ادامه می‌یابد که شبکه‌ی مولد داده‌هایی تولید کند که به حد کافی به داده‌های آموزشی اصلی شبیه باشند و شبکه‌ی ممیز قادر به تشخیص داده‌های تولید شده از داده‌های اصلی نباشد.

از آنجایی که مدل‌های عمیق بازشناسایی شخص معمولاً مدل‌هایی بسیار پیچیده و با لایه‌های متعدد می‌باشند، درصورتی‌که تعداد نمونه‌های آموزشی به اندازه‌ی کافی زیاد نباشد، مدل در هنگام آزمایش دچار بیش‌برازش می‌شود. در اختیار داشتن تعداد بیشتری نمونه‌های آموزشی می‌تواند عملکرد مدل را بهبود دهد. متداول‌ترین روش‌هایی که برای داده‌افزایی در مسائل بینایی ماشین استفاده می‌شود، شامل چرخاندن تصویر، جابه‌جایی تصویر، تغییر رنگ تصویر و ... می‌باشد. داده‌افزایی با استفاده از شبکه‌های مولد تخاصمی موضوعی است که در سال‌های اخیر با اقبال زیادی مواجه شده است. البته این مسئله چالش‌های خاصی نیز دارد و در همه‌ی مسائل، داده‌افزایی با استفاده از شبکه‌های مولد تخاصمی بهترین روش داده‌افزایی نمی‌باشد [۶۵] .

استفاده از شبکه‌های مولد تخاصمی به‌منظور داده‌افزایی در مسائل مختلفی ازجمله ترجمه‌ی تصویر به تصویر [۴۰، ۵۳، ۱۱۸، ۹۶، ۱۳] ، انتقال سبک [۲۰] و ... به کار گرفته شده است. استفاده از شبکه‌های مولد تخاصمی در مسئله‌ی بازشناسایی شخص نیز حوزه‌ی به‌روز و فعالی می‌باشد.

در [۱۰۹] با استفاده از DCGAN [۶۸] به تولید تصاویر آموزشی جدید پرداخته شده است. از آنجایی‌که تصاویر تولید شده دارای برچسب نمی‌باشند، روش منظم‌سازی هموارسازی برچسب‌ها برای داده‌های پرت[۵۰] (LSRO) در این مقاله ارائه شده است. این روش یک توزیع یکنواخت از برچسب‌ها را به تصاویر بدون برچسب اختصاص می‌دهد. تصاویر تولید شده توسط شبکه‌ی مولد تخاصمی به همراه تصاویر آموزشی اولیه برای آموزش مدل استفاده می‌شوند. در [۳۸] به‌جای DCGAN از WGAN-GP [۲۷] برای تولید تصاویر آموزشی بیشتر استفاده شده است.

همان‌طور که پیش از این گفته شد یکی از چالش‌های مجموعه داده‌های حوزه‌ی بازشناسایی شخص تفاوت حالات قرارگیری افراد در تصاویر است. در تعدادی از مقالات سعی شده است که از GAN به‌منظور تولید تصاویر جدید از افراد در حالات قرارگیری مختلف استفاده شود. بدین‌ترتیب مدل آموزش دیده در برابر تغییر حالت قرارگیری افراد مقاوم تر خواهد بود [۲۳، ۵۸].

در [۱۱۴] نویسنده به مسئله‌ی تفاوت سبک تصاویر دوربین‌های مختلف اشاره کرده و با استفاده از تصاویر تولید شده توسط CycleGAN [۱۱۸] ، مدل آموزش داده شده را نسبت به تغییر سبک دوربین‌ها مقاوم‌تر می‌کند.

[50] outlier

در [۱۰۷] با جابه‌جا کردن کدهای ظاهری و کدهای ساختاری هر دو تصویر آموزشی، تصاویر جدید ایجاد می‌شود. نکته‌ی ویژه‌ی این مقاله این است که فاز تولید تصاویر آموزشی جدید از فاز آموزش مدل بازشناسایی شخص کاملاً مجزا نیست.

۳.۳.۲ تعمیم‌پذیری و وفق‌دهی دامنه در مدل‌های بازشناسایی شخص

باتوجه‌به اهمیت مسئله‌ی بازشناسایی شخص در حوزه‌های مختلفی ازجمله حفظ امنیت، توجه زیادی به این مسئله در حوزه‌ی پژوهش و صنعت وجود دارد. همچنین به‌دلیل وجود دیتاست‌های متعدد و به کارگیری تکنیک‌های پیشرفته ازجمله یادگیری عمیق در این حوزه، پیشرفت‌های زیادی رخ داده است به‌نحوی‌که مدل‌های ارائه شده برای این مسئله توانسته‌اند به دقت‌های بالایی برسند. اما باوجود عملکرد قابل قبول مدل‌های ارائه شده روی مجموعه‌داده‌های بازشناسایی شخص، چالش‌هایی برای این مسئله وجود دارد. یکی از این چالش‌ها این است که مجموعه‌داده‌های جمع‌آوری شده در این حوزه با داده‌هایی که در یک مسئله‌ی دنیای واقعی جمع‌آوری می‌شوند متفاوت هستند. برای مثال اغلب مجموعه‌داده‌های ارائه شده تصاویر تعداد محدودی از افراد را دارند و تحت شرایط روشنایی خاصی هستند درحالی‌که در شرایط واقعی مسئله می‌تواند پیچیده‌تر باشد. بنابراین مدل‌های ارائه شده برای بازشناسایی شخص باید بتوانند با چالش‌های مختلفی ازجمله تعداد افراد زیاد، شرایط روشنایی متفاوت و پیچیده و ... مواجه شوند.

چالش دیگری که در مسئله‌ی بازشناسایی شخص وجود دارد این است که اگر یک مدل روی یک مجموعه‌داده‌ی حوزه‌ی بازشناسایی شخص آموزش داده شود و روی مجموعه‌داده‌ی دیگری از این حوزه آزمایش شود، عملکرد مدل به شدت کاهش می‌یابد. این تفاوت عملکرد به‌دلیل فاصله‌ی دامنه‌های مجموعه‌های داده ناشی از تفاوت رزولوشن تصاویر، تفاوت میزان روشنایی، تفاوت سرعت حرکت افراد، تفاوت پس‌زمینه‌ها و ... می‌باشد. این چالش می‌تواند برای مسئله‌ی بازشناسایی شخص جدی باشد زیرا نمونه‌های آموزشی در دسترس نمی‌توانند برای دامنه‌های آزمایشی جدید کافی و مؤثر باشند. بنابراین توجه به موضوع وفق‌دهی دامنه در مسئله‌ی بازشناسایی شخص گسترش یافته است.

به‌دلیل هزینه‌بر بودن فرآیند برچسب‌گذاری تصاویر و همچنین کاهش قابل توجه کارآیی مدل‌های بازشناسایی شخص هنگام متفاوت بودن مجموعه داده‌های آموزشی و آزمایشی، مقالات زیادی سعی کرده‌اند که روی موضوع وفق‌دهی دامنه در مسئله‌ی بازشناسایی شخص کار کنند. به نظر می‌آید که روش‌های بانظارت تک‌دامنه‌ای در

مسئله‌ی بازشناسایی شخص، برای سناریوهای واقعی دارای محدودیت باشند. یک روش متداول برای حل این مشکل وفق‌دهی دامنه‌ی بدون نظارت[۵۱] می‌باشد. وفق‌دهی دامنه بدون نظارت این امکان را می‌دهد که اطلاعات یاد گرفته شده از یک مجموعه‌داده‌ی دارای برچسب را بتوان برای پیش‌بینی روی یک مجموعه‌داده‌ی بدون برچسب مرتبط، استفاده کرد. در حالت کلی وفق‌دهی دامنه از داده‌های دارای برچسب یک و یا بیشتر دامنه‌ی منبع به‌منظور حل وظیفه‌ی دامنه‌ی هدف مرتبط استفاده می‌کند. هرچقدر سطح شباهت و مرتبط بودن بین دامنه‌های منبع و هدف بیشتر باشد، این عملیات موفقیت‌آمیزتر خواهد بود. درواقع در وفق‌دهی دامنه‌ی بدون نظارت، دانش از دامنه‌ی دارای برچسب منبع به دامنه‌ی بدون برچسب هدف منتقل می‌شود.

وفق‌دهی دامنه می‌تواند به سه شکل بانظارت، نیمه نظارتی و بدون نظارت انجام شود. در حالت بانظارت، تصاویر دامنه‌ی هدف دارای برچسب می‌باشند ولی مقدار این داده‌ها برای آموزش یک مدل کامل، کافی نیست. در حالت نیمه نظارتی هم داده‌های دارای برچسب و هم داده‌های بدون برچسب در دامنه‌ی هدف موجود می‌باشد. در حالت بدون نظارت تصاویر دامنه‌ی هدف بدون برچسب می‌باشند. در وفق‌دهی دامنه‌ی بدون نظارت، شبکه روی داده‌های دارای برچسب دامنه‌ی منبع و داده‌های بدون برچسب دامنه‌ی هدف متفاوت ولی مرتبط به دامنه‌ی منبع، آموزش می‌بیند با این هدف که در هنگام آزمایش روی دامنه‌ی هدف عملکرد قابل قبولی داشته باشد.

در [۷۷] مسئله‌ی وفق‌دهی دامنه‌ی بدون نظارت در مسئله‌ی بازشناسایی شخص بررسی شده است. در این مقاله تعدادی فرض روی فضای ویژگی استخراج شده معرفی می‌شود و از این فرض‌ها چندین تابع اتلاف مشتق می‌شود. برای بهینه‌سازی آن‌ها یک شمای جدید خود-آموزشی[۵۲] برای وظیفه‌ی وفق‌دهی دامنه‌ی بدون نظارت در بازشناسایی شخص معرفی می‌شود. این شِما مکرراً حدس‌هایی بر اساس کدگذار[۵۳] درباره‌ی داده‌های بدون برچسب هدف می‌زند و کدگذار را براساس برچسب‌های حدس‌زده شده آموزش می‌دهد.

در مقاله‌ی [۵۶] یک الگوریتم یادگیری افزایشی[۵۴] بدون نظارت با عنوان TFusion معرفی شده است. در این الگوریتم از یادگیری انتقالی الگوهای فضایی-زمانی[۵۵] عابرین پیاده در دامنه‌ی هدف استفاده می‌شود. در ابتدا الگوریتم، طبقه‌بندی‌کننده‌ی بصری آموزش دیده روی مجموعه‌داده‌ی دارای برچسب منبع را به‌منظور یادگیری الگوهای زمانی عابرین پیاده در دامنه‌ی هدف، به دامنه‌ی هدف منتقل می‌کند. سپس یک مدل ترکیب بیزی[۵۶]

[51] unsupervised domain adaptation
[52] self-training
[53] encoder
[54] incremental
[55] spatio-temporal
[56] Bayesian fusion model

به‌منظور ترکیب الگوهای فضایی-زمانی یاد گرفته شده با ویژگی‌های بصری، برای دستیابی به یک طبقه‌بندی‌کننده‌ی بهبودیافته معرفی می‌شود. در نهایت به‌منظور بهینه‌سازی افزایشی طبقه‌بندی‌کننده بر مبنای داده‌های بدون برچسب در دامنه‌ی هدف، یک روش ترویج متقابل[۵۷] بر مبنای یادگیری رتبه‌بندی معرفی می‌شود.

در مقاله‌ی [۸۹] برای مقابله با چالش افت عملکرد مدل آموزش دیده در هنگام آزمایش روی مجموعه‌داده‌ی هدف و کاهش فاصله‌ی دامنه‌های منبع و هدف، شبکه‌ی مولد تخاصمی انتقال شخص (PTGAN) معرفی شده است.

در مقاله‌ی [۵۷] برای وفق‌دهی دامنه‌ی بدون نظارت روشی ارائه شده است که طی آن، طبقه‌بندی‌کننده‌ی آموزش دیده بتواند روی داده‌های بدون برچسب هدف عملکرد خوبی داشته باشد. این کار با استفاده از شبکه‌ی مولد تخاصمی نگهدارنده‌ی شباهت SimPGAN[۵۸] معرفی شده در این مقاله، انجام می‌شود. SimPGAN از شبکه‌ی مولد تخاصمی به همراه اتلاف پایداری شباهت[۵۹]، که توسط یک شبکه‌ی عصبی کانوولوشنال عمیق سیامی محاسبه می‌شود، تشکیل شده است. این شبکه‌ی سیامی شباهت تصاویر منتقل شده از یک فرد یکسان را محاسبه می‌کند.

مقاله‌ی [۵۰] یک شبکه‌ی بازشناسایی و وفق‌دهی (ARN) را برای وفق‌دهی دامنه‌ی بدون نظارت ارائه کرده است به‌نحوی‌که این شبکه قادر است اطلاعات را در میان مجموعه‌داده‌ها به‌دست آورد و به ویژگی‌های یکسان بین دامنه‌ها با هدف بازشناسایی شخص پی ببرد.

در [۱۷] با استفاده از CycleGAN تصاویر دارای برچسب دامنه‌ی منبع به دامنه‌ی هدف منتقل می‌شوند به‌نحوی‌که تصاویر منتقل شده سبک مشابه با تصاویر دامنه‌ی هدف دارند. در گام بعدی تصاویر تغییر سبک داده شده به همراه برچسب‌های متناظر خود برای یادگیری بانظارت در دامنه‌ی هدف مورداستفاده قرار می‌گیرند. در هر تصویر، هویت آن باید پس از ترجمه ثابت بماند همچنین دو دامنه هویت‌های کاملاً متفاوتی دارند بنابراین هویت تصویر ترجمه شده باید با تمامی هویت‌های تصاویر دامنه‌ی هدف متفاوت باشد. برای رسیدن به این هدف دو نوع شباهت بدون نظارت تعریف می‌شود. مورد اول شباهت یک تصویر با خودش قبل و بعد از فرآیند ترجمه است و مورد دوم عدم شباهت تصویر ترجمه شده‌ی دامنه‌ی منبع و تصاویر دامنه‌ی هدف می‌باشد. هر دو قید توسط شبکه‌ای با عنوان SPGAN پیاده‌سازی می‌شوند که از یک شبکه‌ی سیامی و یک CycleGAN تشکیل شده است.

[57] mutual promotion procedure
[58] similarity preserved
[59] similarity consistency

در اکثر مقالات سعی شده است که فاصله‌ی بین دامنه‌ی منبع و دامنه‌ی هدف را از لحاظ سطح تصویر [۱۷، ۸۹] و یا سطح ویژگی [۶۴، ۸۷] کاهش دهند. باوجود مؤثر بودن این روش‌ها اغلب تفاوت‌های درون-دامنه‌ای در دامنه‌ی هدف درنظر گرفته نمی‌شود. اما در [۱۱۱] به‌طور صریح تفاوت درون-دامنه‌ای تنوع تصاویر ناشی از دوربین‌های مختلف در دامنه‌ی هدف درنظر گرفته می‌شود. در [۱۱۲] علاوه‌بر تنوع تصاویر ناشی از تنوع دوربین‌ها، دو تفاوت درون دامنه‌ای دیگر نیز برای دامنه‌ی هدف درنظر گرفته می‌شود.

در مقاله‌ی [۷۸] با استفاده از یادگیری معیار با عنوان تابع اتلاف AM-Softmax و چند تکنیک آموزشی دیگر یک مدل کارآمد و در عین حال بهینه ارائه شده است که قابلیت تعمیم‌پذیری مناسبی دارد. در این مقاله از معماری شبکه‌ی OSNet [۱۱۶] استفاده شده است.

در [۶۶] از منظر یادگیری بازنمایی‌ها به دو مسئله‌ی تفاوت توزیع‌های داده در دامنه‌های هدف و منبع و عدم وجود اطلاعات برچسب در دامنه‌ی هدف پرداخته شده است. برای مسئله‌ی اول، وفق‌دهی دامنه‌ی «آگاه از دوربین»[۶۰] به‌منظور کاهش تفاوت بین دامنه‌های هدف و منبع و همچنین درون زیردامنه‌ها، به کار گرفته شده است. برای مسئله‌ی دوم، پیوستگی زمانی[۶۱] در هر دوربین از دامنه‌ی هدف به‌منظور ایجاد اطلاعات تمایزدهنده استخراج می‌شود. این مسئله با تولید پویا و برخط سه‌تایی‌ها درهر دسته، به‌منظور استفاده‌ی بیشینه از بازنمایی ویژگی بهبودیافته در فرآیند آموزش، پیاده‌سازی می‌شود.

در [۱۰۱] برای وظیفه‌ی وفق‌دهی دامنه در مسئله‌ی بازشناسایی شخص، یک روش خود-آموزشی به همراه چارچوب افزایشی تدریجی[۶۲](PAST) به‌منظور بهبود تدریجی عملکرد مدل روی مجموعه‌داده‌ی هدف ارائه شده است. چارچوب PAST از دو مرحله تشکیل شده است. مرحله‌ی اول مرحله‌ی محافظ کار[۶۳] نام دارد و در آن ساختارهای محلی در داده‌های دامنه‌ی هدف، با استفاده از توابع اتلاف سه‌گانه به‌منظور بهبود بازنمایی ویژگی‌ها، به‌دست می‌آیند. مرحله‌ی دوم گسترش[۶۴] نام دارد و در آن مرحله، شبکه به‌طور پیوسته از طریق اضافه کردن یک لایه‌ی دسته‌بندی قابل تغییر به لایه‌ی آخر مدل، بهینه‌سازی می‌شود. این عمل، استفاده از اطلاعات سراسری توزیع داده را ممکن می‌سازد. علاوه‌براین به‌منظور افزایش قابلیت اعتماد سه‌تایی‌های انتخاب شده یک اتلاف سه‌گانه‌ی برپایه‌ی رتبه‌بندی در مرحله‌ی اول معرفی شده است.

[60] camera-aware
[61] temporal continuity
[62] progressive augmentation framework
[63] conservative
[64] promoting

در [۷۶] یک مدل عمیق بازشناسایی شخص با عنوان شبکه‌ی نگاشت نامتغیر دامنه‌ای[۶۵](DIMN) معرفی شده است. DIMN با هدف یادگیری نگاشت بین یک تصویر از شخص و طبقه‌بندی‌کننده‌ی هویت طراحی شده است. برای ایجاد یک مدل ثابت در دامنه، یک خط لوله فرایادگیری[۶۶] دنبال می‌شود و در هر مرحله‌ی آموزش، زیرمجموعه‌ای از دامنه‌ی منبع نمونه‌برداری می‌شود. مدل معرفی شده تفاوت چشمگیری با روش‌های فرایادگیری قبلی دارد در آن ۱) برای دامنه‌ی هدف نیازی به به‌روزرسانی مدل نیست.۲) زیرمجموعه‌های آموزشی مختلف به‌منظور دستیابی به مقیاس‌پذیری و قابلیت تمایز، یک حافظه را به اشتراک می‌گذارند. ۳) می‌تواند برای تطبیق تعداد دلخواهی از هویت‌ها در دامنه‌ی هدف استفاده شود.

در مقاله‌ی [۵۱] برای انجام وظیفه‌ی بازشناسایی شخص کنار-دامنه‌ای به‌طور مؤثر، یک شبکه‌ی انتقال وفقی[۶۷](ATNet) معرفی شده است. شبکه‌ی ATNet به دلایل اصلی ایجاد فاصله در دامنه‌های هدف و منبع نگاه می‌کند و از طریق اصل «تقسیم و غلبه» به این مسئله می‌پردازد. این شبکه انتقال کنار-دامنه‌ای را به مجموعه‌ای از زیرانتقال‌ها که هرکدام از آن‌ها به تغییر سبک با یک فاکتور خاص مثل روشنایی، رزولوشن، زاویه دید دوربین و ... تمرکز دارند، تجزیه می‌کند. سپس یک استراتژی ترکیبی وفق‌دهی ارائه می‌کند که در آن زیرانتقال‌ها باتوجه‌به اهمیت تأثیر فاکتورهای مختلف روی تصاویر، ادغام می‌شوند.

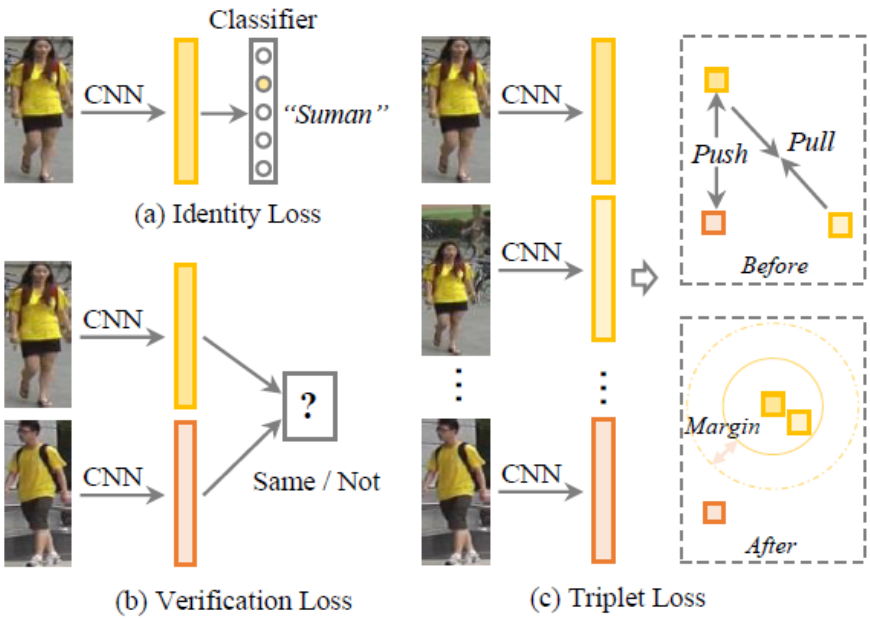
۴.۳.۲ توابع اتلاف در مسئله‌ی بازشناسایی شخص

در هر مدل عمیق بازشناسایی شخص، تعریف تابع اتلاف مناسب یکی از گام‌های مهم برای تضمین عملکرد مناسب مدل می‌باشد. تابع اتلاف می‌تواند انعطاف‌پذیری بسیار زیادی را به شبکه‌های عصبی بدهد و مشخص می‌کند که خروجی شبکه دقیقا چگونه به سایر قسمت‌های شبکه مربوط است. به‌طور کلی شبکه‌های عصبی وظایف متفاوتی را می‌توانند انجام دهند؛ برای مثال پیش‌بینی کردن مقادیر پیوسته مثل قیمت خانه، درآمد ماهیانه و ... و یا طبقه‌بندی کلاس‌های مجزا مانند تصاویر سگ‌ها و گربه‌ها می‌تواند از وظایف یک شبکه‌ی عصبی باشد. وظایف متفاوت نیاز به تعریف توابع اتلاف متفاوتی دارند زیرا قالب خروجی شبکه‌ی عصبی در آن‌ها مختلف می‌باشد.

برای وظایف پیچیده‌ای مانند بازشناسایی شخص باتوجه‌به نوع نگاه به مسئله می‌توان توابع اتلاف گوناگونی را

[65] domain-invariant mapping network
[66] meta learning
[67] adaptive transfer network

برای شبکه‌ی عصبی مربوطه انتخاب کرد. در مدل‌های عمیق بازشناسایی شخص اغلب از سه دسته توابع اتلاف و شکل‌های گوناگون آن‌ها و حتی ترکیب آن‌ها استفاده می‌شود. تابع اتلاف شناسایی[۶۸]، تابع اتلاف تصدیق[۶۹] و تابع اتلاف سه‌گانه[۷۰] پرکاربردترین توابع اتلاف استفاده شده در مدل‌های عمیق بازشناسایی شخص می‌باشند. در شکل ۱۲.۲ نحوه‌ی عملکرد توابع اتلاف شناسایی، تصدیق و سه‌گانه نمایش داده شده است. علاوه‌براین، تابع اتلاف چهارگانه[۷۱] در مقاله‌ی [۹] معرفی شده است که قدرت تعمیم‌پذیری بالاتری از تابع اتلاف سه‌گانه را ممکن می‌سازد. در ادامه توابع اتلاف ذکر شده مختصراً معرفی می‌شوند.



شکل ۱۲.۲: توابع اتلاف شناسایی، تصدیق و سه‌گانه [۹۴]

- اتلاف شناسایی: هنگامی که در مدل بازشناسایی شخص از تابع اتلاف شناسایی استفاده شود، به مسئله به‌عنوان یک وظیفه‌ی طبقه‌بندی تصاویر نگاه شده است. در این حالت، هر هویت یک کلاس جداگانه می‌باشد. بر اساس ویژگی‌هایی که از تصویر ورودی استخراج می‌شود، باید کلاس متناظر با آن تصویر توسط شبکه پیش‌بینی شود. اگر تصویر ورودی x_i برچسب y_i را داشته باشد، احتمال پیش‌بینی برچسب

[68] identity loss
[69] verification loss
[70] triplet loss
[71] quadruplet loss

y_i برای تصویر x_i با تابع Softmax کدگذاری می‌شود. سپس اتلاف شناسایی از طریق آنتروپی متقاطع[۷۲] محاسبه می‌شود:

$$L_{id} = -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i|x_i)) \tag{۲.۲}$$

که در آن n تعداد نمونه‌های آموزشی در هر دسته[۷۳] می‌باشد.

- اتلاف تصدیق: یک مدل تصدیق یک زوج از تصاویر را به‌عنوان ورودی دریافت می‌کند و میزان شباهت آن‌ها را به‌عنوان خروجی برمی‌گرداند، بدین‌ترتیب می‌توان نتیجه گرفت که آیا دو تصویر مربوط به یک هویت می‌باشند و یا مربوط به هویت‌های متفاوتی هستند. مدل‌های تصدیق به مسئله‌ی بازشناسایی شخص به‌عنوان یک مسئله‌ی طبقه‌بندی دودویی نگاه می‌کنند. تابع اتلاف تصدیق معمولاً به‌شکل اتلاف هم‌سنجی[۷۴] و یا اتلاف تصدیق دودویی[۷۵] است. اتلاف هم‌سنجی، مقایسه‌ی فاصله‌ی زوج‌ها را بهبود می‌دهد.

$$L_{con} = (1-\delta_{ij})\{max(0, \rho - d_{ij})\}^2 + \delta_{ij}d_{ij}^2 \tag{۳.۲}$$

که در آن d_{ij} فاصله‌ی اقلیدسی بین ویژگی‌های دو ورودی x_iو x_j می‌باشد. δ_{ij} برابر با ۱ است اگر x_i و x_j متعلق به یک هویت باشند، و برابر با ۰ است اگر x_i و x_j متعلق به هویت‌های متفاوتی باشند. ρ پارامتر حاشیه‌ای است.

اتلاف تصدیق دودویی مثبت یا منفی بودن زوج تصویر ورودی را تشخیص می‌دهد. ویژگی f_{ij} به‌صورت $f_{ij} = (f_j - f_i)^2$ تعریف می‌شود. f_i و f_j بردارهای ویژگی ورودی‌های x_i و x_j هستند. شبکه‌ی تصدیق ویژگی تعریف شده را به دو دسته‌ی مثبت یا منفی طبقه‌بندی می‌کند. $p(\delta_{ij}|f_{ij})$ بیانگر احتمال تشخص زوج ورودی x_i و x_j به‌عنوان δ_{ij}(۰ یا ۱) می‌باشد. اتلاف تصدیق با آنتروپی متقاطع به‌صورت

[72] cross-entropy
[73] batch
[74] contrastive loss
[75] binary verification loss

زیر تعریف می‌شود:

$$L_{veri}(i,j) = -\delta_{ij}\log(p(\delta_{ij}|f_{ij})) - (۱-\delta_{ij})\log(۱-p(\delta_{ij}|f_{ij})) \qquad (۴.۲)$$

به‌منظور عملکرد بهتر مدل،تابع اتلاف تصدیق اغلب با تابع اتلاف شناسایی ترکیب می‌شود [۷، ۸۴، ۱۷، ۱۰۸].

- اتلاف سه‌گانه: تابع اتلاف سه‌گانه برای اولین بار توسط گوگل [۷۲] برای وظیفه‌ی تشخیص چهره ارائه شد. هدف تابع اتلاف سه‌گانه بیشینه کردن تفاوت‌های بیرون-کلاسی و کمینه کردن تفاوت‌های درون-کلاسی می‌باشد. در تابع اتلاف سه‌گانه، یک سه‌تایی وجود دارد که شامل یک نمونه‌ی x_i، نمونه‌ی مثبت x_j که هویتی مشابه با نمونه‌ی x_i دارد و نمونه‌ی منفی x_k که هویتی متفاوت با نمونه‌ی x_i دارد، می‌باشد. ایده‌ی پایه‌ی تابع اتلاف سه‌گانه این است که فاصله‌ی بین زوج‌های مثبت کمتر از فاصله‌ی بین زوج‌های منفی باشد. اتلاف سه‌گانه به همراه پارامتر حاشیه‌ای این‌گونه تعریف می‌شود:

$$L_{tri}(i,j,k) = max(\rho + d_{ij} - d_{ik}, ۰) \qquad (۵.۲)$$

که در آن d فاصله‌ی اقلیدسی دو نمونه را نمایش می‌دهد.

در مدل سه‌گانه، انتخاب سه‌تایی‌ها اهمیت بسیار زیادی دارد. اگر اکثر سه‌تایی‌های انتخاب شده، نمونه‌های آسانی باشند، مدل آموزش دیده، قدرت تشخیص پایینی خواهد داشت. در سال ۲۰۱۷ در مقاله‌ی [۳۱] نشان داده شده است که استفاده از گونه‌ای از تابع اتلاف سه‌گانه در مسئله‌ی بازشناسایی شخص برای مدل‌هایی که از ابتدا آموزش داده می‌شوند و همچنین مدل‌های پیش‌آموزش داده شده، عملکرد بسیار مناسبی نسبت به سایر کارهای آن زمان دارد. تا قبل از آن زمان استفاده از توابع اتلاف شناسایی و تصدیق و ترکیب آن‌ها در مدل‌های عمیق بازشناسایی شخص، بسیار متداول بود. این مقاله با معرفی دو استراتژی برای انتخاب سه‌تایی‌ها، کارآیی تابع اتلاف سه‌گانه را برای مسئله‌ی بازشناسایی شخص بهبود داده است. این دو استراتژی، استراتژی دسته‌ی کامل[۷۶] و استراتژی دسته‌ی سخت[۷۷] می‌باشند. در استراتژی دسته‌ی کامل، تمامی سه‌تایی‌های معتبر و اتلاف میانگین آن‌ها انتخاب می‌شوند. در استراتژی دسته‌ی سخت در

[76] batch all
[77] batch hard

هنگام تشکیل سه‌تایی‌ها، دشوارترین نمونه‌های مثبت و منفی در دسته انتخاب می‌شوند.

به‌طور کلی اگر از تابع اتلاف سه‌گانه در مدل استفاده شود، یک بخش بسیار مهم فرآیند یادگیری، استخراج سه‌تایی‌های دشوار می‌باشد. اگر اکثر سه‌تایی‌ها آسان باشند، مدل آموزش دیده عملکرد مناسبی نخواهد داشت. ازطرفی استخراج سه‌تایی‌های دشوار کار زمان‌بری می‌باشد. همچنین اگر سه‌تایی‌های استخراج شده بیش از حد دشوار باشند فرآیند آموزش ناپایدار می‌شود.

در تعدادی از مقالات از ترکیب توابع اتلاف شناسایی و سه‌گانه برای آموزش مدل عمیق بازشناسایی شخص استفاده شده است [۱۱۷، ۸، ۹۳، ۲۸، ۷۶، ۵۲] .

- اتلاف چهارگانه: تابع اتلاف چهارگانه [۹] می‌تواند نسبت به تابع اتلاف سه‌گانه، مدلی با تفاوت‌های درون-کلاسی کمتر و تفاوت‌های بیرون-کلاسی بیشتر را سبب شود. این تابع اتلاف براساس تابع اتلاف سه‌گانه طراحی شده است. در تابع اتلاف چهارگانه، یک چهارتایی شامل یک نمونه‌ی اولیه، یک نمونه‌ی مثبت دارای هویت یکسان با نمونه‌ی اولیه و دو نمونه‌ی منفی استخراج شده دارای هویت متفاوت با نمونه‌ی اولیه می‌باشد. چهارتایی‌ها با استخراج برخط نمونه‌های منفی سخت مبتنی بر حاشیه، فرمول‌بندی شده‌اند.

۵.۳.۲ معیارهای ارزیابی عملکرد مدل‌های بازشناسایی شخص

در اکثر مقالات مربوط به بازشناسایی شخص، پژوهشگران معیار دقت و کارآیی مدل پیشنهادی را به‌وسیله‌ی معیارهای rank(k) و CMC بیان می‌کنند. rank(k) معیاری برای سنجش دقت است که در مسائل طبقه‌بندی کاربرد دارد. در صورتی‌که خروجی یک مسئله‌ی طبقه‌بندی به‌صورت یک لیست رتبه‌بندی شده باشد، به‌نحوی‌که عنصر اول لیست محتمل‌ترین پاسخ برای نمونه‌ی آزمایشی باشد و عنصر دوم لیست دومین پاسخ محتمل برای نمونه‌ی آزمایشی باشد و به همین ترتیب n-امین عنصر لیست خروجی n-امین پاسخ محتمل برای نمونه‌ی آزمایشی باشد، در آن صورت اگر پاسخ صحیح در k عنصر اول لیست خروجی باشد، خروجی دقت rank(k) را دارا می‌باشد. بنابراین، rank(1) سختگیرانه‌ترین معیار برای میزان دقت و خطا است، درحالی‌که rank(k), $k > 1$ امکان وجود میزانی از خطا را ممکن می‌سازد. در مسائل بازشناسایی شخص معمولاً rank(1) از اهمیت بسیار زیادی برخوردار است، چرا که سیستم باید بتواند از میان داده‌های محدود و دشوار، خروجی صحیح را در

نگاه اول تشخیص دهد. در بسیاری از مقالات معمولاً علاوه‌بر rank(1) ، میزان rank(5) و rank(10) نیز بیان می‌شود. معیار CMC درواقع احتمال تعلق پاسخ صحیح به k رتبه‌ی اول را بیان می‌کند.

در شرایطی که برای هر پرس‌وجو فقط یک حقیقت عینی[۷۸] وجود داشته باشد، معیار CMC دقیق می‌باشد. زیرا این معیار در فرآیند ارزیابی فقط تطابق اول[۷۹] را درنظر می‌گیرد. معیار دیگری که برای ارزیابی عملکرد الگوریتم‌های بازشناسایی شخص استفاده می‌شود معیار mAP است. دلیل استفاده از این معیار این است که یک سیستم بازشناسایی باید قادر باشد که تمام مطابقت‌های صحیح را به کاربر برگرداند. ممکن است دو سیستم مختلف در برگرداندن اولین خروجی صحیح، یکسان عمل کنند اما توانایی فراخوانی بازیابی[۸۰] متفاوتی داشته باشند. در چنین شرایطی معیار CMC قابلیت تمایز میان مقدار کارآیی دو سیستم را ندارد، درحالی‌که معیار mAP این توانایی را داراست.

درواقع در حوزه‌ی بازشناسایی شخص، معیار mAP می‌تواند به خوبی برای مقایسه‌ی سیستم‌هایی استفاده شود که در تطابق‌های اول بسیار شبیه به‌هم عمل می‌کنند اما توانایی بازیابی متفاوتی برای تطابق‌های بعدی، که معمولاً تطابق های دشوارتر می‌باشند، دارند. بنابراین در اکثر مقالات علاوه‌بر معیار rank(k) معیار mAP نیز بیان می‌شود.

در شرایطی که ابزارهای آموزش یا آزمایش مدل بازشناسایی شخص، منابع محدودی داشته باشند، معیارهای مرتبط با پیچیدگی مدل مثل اندازه‌ی پارامترهای شبکه و تعداد عملیات ممیز شناور در ثانیه (FLOPs) می‌توانند به‌عنوان معیارهای ارزیابی درنظر گرفته شوند.

۶.۳.۲ مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص

در این بخش معروف‌ترین مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص آورده شده است. در جدول ۲.۲ ویژگی‌های مختلف این مجموعه‌داده‌ها مانند تعداد دوربین‌ها، تعداد تصاویر و تعداد افراد ذکر شده است.

VIPeR: این مجموعه‌داده شامل تصاویری از دو دوربین متفاوت می‌باشد که هر کدام از دوربین‌ها یک تصویر از هر فرد گرفته‌اند. تصاویر دو دوربین از لحاظ روشنایی و حالت قرارگیری افراد متفاوت هستند. باوجود اینکه این مجموعه‌داده تاکنون در بسیاری از پژوهش‌ها مورداستفاده قرار گرفته است اما همچنان یکی از چالشی‌ترین

[78] ground truth
[79] first match
[80] retrieval recall

مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص می‌باشد. تمام تصاویر این دیتاست به اندازه‌ی ۴۸ × ۱۲۸ هستند [۲۶]. در شکل ۱۳.۲ نمونه‌هایی از تصاویر این مجموعه‌داده نمایش داده شده است.

شکل ۱۳.۲: نمونه‌هایی از تصاویر مجموعه‌داده‌ی VIPeR

ETHZ: برخلاف سایر مجموعه‌داده‌های اشاره شده که تصاویر همه‌ی آن‌ها توسط بیش از یک دوربین گرفته شده است، در این مجموعه‌داده دوربین متحرک وجود دارد. زاویه دید تصاویر تفاوت زیادی ندارند، اما تفاوت قابل ملاحظه‌ای در روشنایی و مقیاس تصاویر وجود دارد. اندازه‌ی تصاویر در این مجموعه‌داده متفاوت است که بسته به کاربرد می‌توان همگی را به یک اندازه‌ی مشخص تبدیل کرد. این مجموعه‌داده، سه دنباله از تصاویر را فراهم می‌کند که دنباله‌های ۱ و ۲ و ۳ به ترتیب شامل تصاویر ۸۵، ۳۵ و ۲۸ فرد می‌باشند [۷۳].

GRID: تصاویر این مجموعه‌داده از هشت دوربین مجزا در یک ایستگاه زیرزمینی شلوغ تهیه شده‌اند. این تصاویر کیفیت پایینی دارند و به همین دلیل مجموعه‌داده‌ی GRID یک مجموعه‌داده‌ی چالشی محسوب می‌شود. هر فرد دارای دو تصویر از زاویه‌های مختلف است. همچنین ۷۷۵ تصویر اضافی در این مجموعه وجود دارد که کاربرد آن در اعتبارسنجی و آزمایش مدل می‌باشد [۵۴]. در شکل ۱۴.۲ نمونه‌هایی از تصاویر این مجموعه‌داده نمایش داده شده است.

شکل ۱۴.۲: نمونه‌هایی از تصاویر مجموعه‌داده‌ی GRID

CAVIAR4ReID: در این مجموعه‌داده تصاویر ۷۲ نفر ثبت شده است. تصاویر ۵۰ نفر توسط دو دوربین ثبت شده است و ۲۲ نفر دیگر فقط از یک دوربین تصویر دارند. تصاویر هر فرد به گونه‌ای انتخاب شده است که تنوع رزولوشن را بیشینه کند. تفاوت زیادی در میزان روشنایی و طرز قرارگیری افراد بین تصاویر وجود دارد [۱۱]. در شکل ۱۵.۲ نمونه‌هایی از تصاویر این مجموعه‌داده نمایش داده شده است.

شکل ۱۵.۲: نمونه‌هایی از تصاویر مجموعه‌داده‌ی CAVIAR4ReID

PRID2011: تصاویر موجود در این مجموعه‌داده از دو دوربین A و B می‌باشند. تصاویر ۳۸۵ نفر توسط دوربین A و تصاویر ۷۴۹ نفر توسط دوربین B ثبت شده است که در این میان ۲۰۰ نفر بین تصاویر دو دوربین مشترک هستند. این مجموعه‌داده نسخه‌ی تک شات نیز دارد که از تصاویر لحظه‌ای تصادفی تشکیل شده است [۳۵].

WARD: این مجموعه‌داده شامل ۴۷۸۶ تصویر از ۷۰ نفر است که توسط سه دوربین بدون هم‌پوشانی ثبت شده‌اند. هر فرد تصاویر متعددی از هر دوربین دارد. تصاویر این مجموعه‌داده در میزان روشنایی، رزولوشن و طرز قرارگیری افراد با یکدیگر تفاوت بسیار زیادی دارند [۵۹]. در شکل ۱۶.۲ نمونه‌هایی از تصاویر این مجموعه‌داده نمایش داده شده است.

شکل ۱۶.۲: نمونه‌هایی از تصاویر مجموعه‌داده‌ی WARD

CUHK: این مجموعه‌داده توسط دانشگاه هنگ کنگ در چین، برای مسئله‌ی بازشناسایی شخص جمع‌آوری شده است. سه مجموعه‌داده‌ی مجزا با نام‌های CUHK01، CUHK02 و CUHK03 دارد. مجموعه‌داده‌ی CUHK01 [۴۸] شامل دو تصویر برای هر شخص از هر دوربین می‌باشد. تصاویر دوربین B بیشتر از روبه‌رو و از پشت افراد است اما تصاویر دوربین A زوایا و حالت‌های قرارگیری متفاوت‌تری را شامل می‌شود. کیفیت تصاویر این مجموعه نسبتاً خوب است. مجموعه داده‌ی CUHK02 [۴۷] گسترشی از مجموعه‌داده‌ی CUHK01 می‌باشد، به این صورت که علاوه‌بر جفت دوربین‌های CUHK01، چهار جفت دوربین دیگر نیز تصاویر را ثبت کرده‌اند. هر جفت دوربین به ترتیب شامل تصاویر ۹۷۱، ۳۰۶، ۱۹۳، ۱۰۷ و ۲۳۹ شخص می‌باشد. هر شخص دارای دو تصویر از دید هر دوربین است. مجموعه داده‌ی CUHK03 [۴۹] اولین مجموعه داده‌ی بازشناسایی شخص است که برای یادگیری عمیق به‌اندازه‌ی کافی بزرگ می‌باشد. تصاویر این مجموعه داده توسط پنج جفت دوربین نظارتی ثبت شده‌اند. در شکل ۱۷.۲ نمونه‌هایی از تصاویر مجموعه‌داده‌ی CUHK03 نمایش داده شده است.

شکل ۱۷.۲: نمونه‌هایی از تصاویر مجموعه‌داده‌ی CUHK03

RAID: تصاویر این مجموعه‌ی داده توسط چهار دوربین غیرهم‌پوشان ثبت شده‌اند. دو دوربین تصاویر فضای باز و دو دوربین تصاویر فضای بسته را ثبت کرده‌اند. به دلیل تفاوت موقعیت فضای باز و فضای بسته، تصاویر موجود در این مجموعه داده در میزان روشنایی تفاوت قابل ملاحظه‌ای دارند. در این مجموعه داده هر فرد تصاویری ثبت شده توسط چهار دوربین را داراست [۱۵]. در شکل ۱۸.۲ نمونه‌هایی از تصاویر این مجموعه‌داده نمایش داده شده است.

شکل ۱۸.۲: نمونه‌هایی از تصاویر مجموعه‌داده‌ی RAID

iLIDS-VID: بر اساس این فرض که در سیستم‌های واقعی بازشناسایی فرد، باید تراژکتوری‌های تصاویر افراد را داشته باشیم، این مجموعه داده تعداد ۶۰۰ تراژکتوری از ۳۰۰ نفر را از مجموعه‌داده‌ی iLIDS MCTS استخراج

کرده است [۸۸].

Market1501: تصاویر این مجموعه‌داده از مقابل یک سوپرمارکت در دانشگاه Tsinghua توسط شش دوربین مجزا ثبت شده‌اند. پنج دوربین با رزولوشن بالا و یک دوربین با رزولوشن پایین می‌باشد. تصاویر هر فرد حداکثر توسط شش دوربین و حداقل توسط دو دوربین گرفته شده‌است. کیفیت تصاویر این مجموعه‌داده از کیفیت تصاویر مجموعه‌داده‌ی CUHK03 کمتر است [۱۰۳].

MARS: این مجموعه‌داده، گسترشی از مجموعه‌داده‌ی Market1501 می‌باشد. MARS اولین مجموعه‌داده با مقیاس بزرگ برای بازشناسایی شخص مبتنی بر ویدیو است [۱۰۲].

DukeMTMC-reID: این مجموعه‌داده، یک مجموعه‌داده‌ی بزرگ-مقیاس، چند هدفی و چند دوربینی دارای برچسب می‌باشد. بیش از ۲۷۰۰ نفر با شناسه‌های منحصربه‌فرد برچسب‌گذاری شده‌اند. تصاویر این افراد توسط هشت دوربین در فضای باز گرفته شده است. این مجموعه‌داده امکان دسترسی به اطلاعات بیشتری مثل فریم‌های کامل، و اطلاعات درجه‌بندی را می‌دهد و به همین دلیل از پتانسیل‌های بالایی برخوردار است [۱۰۹].

MSMT17: تصاویر این مجموعه‌داده توسط ۱۲ دوربین در فضای باز و سه دوربین در فضای بسته ثبت شده‌اند. این تصاویر چهار روز با هوای متفاوت در یک ماه را پوشش می‌دهد. برای هر روز سه ویدئوی یک‌ساعته از صبح، ظهر و بعدازظهر انتخاب شده‌اند. این مجموعه‌داده، بزرگ‌ترین مجموعه‌داده برای بازشناسایی شخص است [۸۹].

شکل ۱۹.۲: نمونه‌هایی از تصاویر مجموعه‌داده‌ی MSMT17

جدول ۲.۲: مجموعه‌داده‌های بازشناسایی شخص

مجموعه داده	سال	تعداد افراد	تعداد دوربین‌ها	تعداد تصاویر	اندازه برش تصاویر	چندشات
VIPeR	۲۰۰۷	۶۳۲	۲	۱۲۶۴	۱۲۸ × ۴۸	
ETH1,2,3	۲۰۰۷	۸۵،۳۵،۲۸	۱	۸۵۸۰	متفاوت	⋆
GRID	۲۰۰۹	۲۵۰	۸	۱۲۷۵	متفاوت	
CAVIAR4ReID	۲۰۱۱	۷۲	۲	۱۲۲۰	متفاوت	⋆
PRID2011	۲۰۱۱	۹۳۴	۲	۲۴۵۴۱	۱۲۸ × ۶۴	⋆
WARD	۲۰۱۲	۷۰	۳	۴۷۸۶	۱۲۸ × ۴۸	⋆
CUHK01	۲۰۱۲	۹۷۱	۲	۳۸۸۴	۱۶۰ × ۶۰	⋆
CUHK02	۲۰۱۳	۱۸۱۶	۱۰ (۵ جفت)	۷۲۶۴	۱۶۰ × ۶۰	⋆
CUHK03	۲۰۱۴	۱۴۶۷	۱۰ (۵ جفت)	۱۳۱۶۴	متفاوت	⋆
RAiD	۲۰۱۴	۴۳	۴	۶۹۲۰	۱۲۸ × ۶۴	⋆
iLIDS-VID	۲۰۱۴	۳۰۰	۲	۴۲۴۹۵	متفاوت	⋆
Market1501	۲۰۱۵	۱۵۰۱	۶	۳۲۶۶۸	۱۲۸ × ۶۴	⋆
MARS	۲۰۱۶	۱۲۶۱	۶	۱۱۹۱۰۰۳	۲۵۶ × ۱۲۸	⋆
DukeMTMC-reID	۲۰۱۷	۱۸۱۲	۸	۳۶۴۴۱	متفاوت	⋆
MSMT17	۲۰۱۸	۴۱۰۱	۱۵	۱۲۶۴۴۱	متفاوت	⋆

۷.۳.۲ خلاصه و نتیجه‌گیری

در این فصل مروری بر کارهای دیگران در حوزه‌ی بازشناسایی شخص صورت گرفت. در ابتدا به بررسی تاثیر یادگیری عمیق در مسائل بینایی ماشین پرداخته شد و مدل‌های مشهور یادگیری عمیق معرفی شدند. باتوجه‌به تاثیر قابل توجه یادگیری عمیق در مسائل بینایی ماشین، می‌توان انتظار داشت که استفاده از تکنیک‌های یادگیری عمیق در حوزه‌ی بازشناسایی شخص نیز نتایج خوبی را به‌دست آورد.

در بخش بعدی این فصل، تاریخچه‌ی بازشناسایی شخص آورده شده است. دو مورد از مهم‌ترین چالش‌های بازشناسایی شخص یعنی کمبود داده‌های آموزشی و راه‌حل‌های ارائه شده برای آن مشکل و همچنین تعمیم‌پذیری مدل بازشناسایی شخص و وفق‌دهی دامنه در این حوزه بررسی شدند. با استفاده از تکنیک‌های یادگیری عمیق نتایج بسیار خوبی در حوزه‌ی بازشناسایی شخص بانظارت به‌دست آمده است. اما حوزه‌ی بازشناسایی شخص

بدون نظارت و وفق‌دهی دامنه در بازشناسایی شخص موضوعات جدیدتر و چالشی‌تری هستند. در این پایان‌نامه مدلی برای مسئله‌ی وفق‌دهی دامنه در بازشناسایی شخص ارائه خواهد شد. همچنین در این فصل توابع اتلاف متداول در بازشناسایی شخص معرفی شدند. در مدل پیشنهادی علاوه‌بر تابع اتلاف طبقه‌بندی داده‌های منبع، از تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف و تابع اتلاف سه‌گانه نیز استفاده می‌شود.

معیارهای ارزیابی مدل‌های بازشناسایی شخص نیز در این فصل بررسی شدند. برای ارزیابی مدل پیشنهادی از معیارهای CMC و mAP استفاده می‌شود. مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص در بخش آخر این فصل جمع‌بندی و مقایسه شدند. متداول‌ترین مجموعه‌داده‌های بزرگ-مقیاس این حوزه که برای آموزش مدل‌های عمیق مناسب هستند، مجموعه‌داده‌های CUHK03 ، Market1501 ، DukeMTMC-reID و MSMT17 می‌باشند. از میان این مجموعه‌های داده در اکثر مقالات وفق‌دهی دامنه در بازشناسایی شخص نتایج مدل روی duke $\rightarrow$ market و market $\rightarrow$ duke ذکر شده است. بنابراین به دلیل امکان مقایسه‌ی بهتر با سایر کارها در فرآیند آموزش مدل از دو مجموعه‌داده‌ی Market1501 و DukeMTMC-reID استفاده می‌شود.

فصل ۳

روش پیشنهادی

۱.۳ پیش‌گفتار

هدف این پژوهش، ارائه‌ی یک مدل بازشناسایی شخص است که هنگام آزمایش روی مجموعه‌داده‌ی بدون برچسب دامنه‌ی هدف، قابلیت تعمیم‌پذیری مناسبی داشته باشد. باتوجه‌به پرهزینه بودن فرآیند برچسب‌گذاری تصاویر آموزشی، کارآیی قابل قبول مدل روی مجموعه‌داده‌ی بدون برچسب آزمایشی بسیار با اهمیت است. در بخش اول این فصل به توضیح مقاله و مدل پایه پرداخته می‌شود و ایده‌ی یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف تبیین می‌شود. در بخش بعدی این فصل، ساختار مدل ارائه شده توضیح داده می‌شود. مدل پیشنهادی، با استفاده از سه تابع اتلاف معرفی شده، عملکرد قابل قبولی در وفق‌دهی دامنه‌ی بدون نظارت در حوزه‌ی بازشناسایی شخص دارد. در این مدل، از تابع اتلاف طبقه‌بندی داده‌های دارای برچسب دامنه‌ی منبع، تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف و تابع اتلاف سه‌گانه استفاده می‌شود. به‌منظور یادگیری تغییرناپذیری نسبت به همسایه‌ها، دو استراتژی بررسی می‌شوند.

۲.۳ مدل پایه

در مقاله‌ی پایه [۱۱۲] روشی برای وفق‌دهی دامنه‌ی بدون نظارت در مسئله‌ی بازشناسایی شخص ارائه شده است. منظور از وفق‌دهی دامنه‌ی بدون نظارت این است که یک مجموعه‌داده‌ی دارای برچسب و یک مجموعه‌داده‌ی بدون برچسب وجود دارند و هدف این است که مدل آموزش دیده، عملکرد مناسبی روی مجموعه‌داده‌ی بدون برچسب داشته باشد. در اکثر روش‌های وفق‌دهی دامنه سعی می‌شود که به‌نحوی دانش را از مجموعه‌داده‌ی دارای برچسب به مجموعه‌داده‌ی بدون برچسب منتقل کنند. اصطلاحاً به مجموعه‌داده‌ای که دارای داده‌های برچسب‌گذاری شده است مجموعه‌داده‌ی منبع و به مجموعه‌داده‌ای که داده‌های بدون برچسب دارد، مجموعه‌داده‌ی هدف گفته می‌شود.

به‌دلیل فاصله‌ی دامنه‌های مجموعه‌داده‌های مختلف ناشی از تفاوت میزان روشنایی، تفاوت سرعت حرکت افراد، تفاوت پس‌زمینه‌ها و ...، هنگام آزمایش مدل روی مجموعه‌داده‌ای متفاوت با مجموعه‌داده‌ی دارای برچسب آموزشی، عملکرد مدل به شدت کاهش می‌یابد. باتوجه‌به پرهزینه‌بودن فرآیند برچسب‌گذاری تصاویر و همچنین افت شدید عملکرد مدل آموزش دیده در هنگام آزمایش روی یک مجموعه‌داده‌ی جدید، حوزه‌ی وفق‌دهی دامنه در بازشناسایی شخص از اهمیت زیادی برخوردار است. در بخش ۳.۳.۲ در رابطه با موضوع وفق‌دهی دامنه در مسئله‌ی بازشناسایی شخص بحث شده است.

۱.۲.۳ حافظه‌ی نمونه

در مقاله‌ی پایه [۱۱۲] ، مدلی ارائه شده است که از دامنه‌ی دارای برچسب منبع و دامنه‌ی بدون برچسب هدف یاد می‌گیرد. در اکثر مقالات ارائه شده در حوزه‌ی وفق‌دهی دامنه در بازشناسایی شخص، سعی می‌شود که به‌نحوی فاصله‌ی دامنه‌های هدف و منبع را کاهش دهند، درحالی‌که ویژگی‌های درون-دامنه‌ای در دامنه‌ی هدف درنظر گرفته نمی‌شوند. ویژگی‌های درون-دامنه‌ای در دامنه‌ی هدف می‌توانند تأثیر زیادی بر عملکرد مدل در هنگام آزمایش داشته باشند. در این مقاله سه ویژگی درون-دامنه‌ای در دامنه‌ی هدف بررسی می‌شوند. برای رسیدن به این هدف، ساختاری با عنوان حافظه‌ی نمونه[1] (ExM) برای ذخیره کردن ویژگی‌های دامنه‌ی هدف معرفی شده است.

به‌منظور بهبود قابلیت تعمیم‌پذیری شبکه روی مجموعه‌ی آزمایشی هدف، ویژگی‌های به‌روز تمامی تصاویر

[1] examplar memory

دامنه‌ی هدف در ExM ذخیره می‌شوند. ساختار ExM یک ساختار کلید-مقداری[۲] است. در ExM، هر قسمت ویژگی نرمال شده‌ی خروجی آخرین لایه‌ی تماماً متصل شبکه‌ی CNN را به‌عنوان کلید ذخیره می‌کند و برچسب مربوط به آن تصویر را در بخش مقدار خود ذخیره می‌کند. اگر تعداد تصاویر دامنه‌ی هدف برابر با N_t باشد، ExM دارای N_t بخش می‌باشد که هر بخش ویژگی و برچسب یک تصویر هدف را ذخیره کرده است. از آنجایی که تصاویر دامنه‌ی هدف بدون برچسب می‌باشند، فرض می‌شود که هر تصویر متعلق به کلاس واحدی است و برای ساده‌سازی، برچسب هر تصویر برابر با اندیس مربوط به آن بخش در ExM درنظر گرفته می‌شود. در هر تکرار فرآیند آموزش، تصویر آموزشی دامنه‌ی هدف $(x_{t,i})$ وارد شبکه‌ی عمیق بازشناسایی می‌شود و ویژگی نرمال شده‌ی آن $(f(x_{t,i}))$ از لایه‌ی تماماً متصل شبکه‌ی CNN به‌دست می‌آید. در هنگام پس‌انتشار، ویژگی موجود در ExM برای نمونه‌ی آموزشی $x_{t,i}$ به‌روزرسانی می‌شود:

$$K[i] \leftarrow \alpha K[i] + (۱ - \alpha) f(x_{t,i}) \tag{۱.۳}$$

که در آن $K[i]$ کلید ExM برای تصویر $x_{t,i}$ در iامین بخش می‌باشد. α مقداری بین ۰ و ۱ دارد و نرخ به‌روزرسانی را کنترل می‌کند.

۲.۲.۳ یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف

هنگام انتقال دانش از دامنه‌ی منبع به دامنه‌ی هدف، اگر تفاوت‌های درون-دامنه‌ای دامنه‌ی هدف درنظر گرفته شوند، بهبود قابل توجهی در عملکرد به‌وجود می‌آید. درمقاله‌ی پایه [۱۱۲] ، سه ویژگی اساسی تغییرناپذیری نسبت به نمونه‌ها[۳] ، تغییرناپذیری نسبت به دوربین‌ها[۴] و تغییرناپذیری نسبت به همسایه‌ها[۵] در دامنه‌ی هدف بررسی شده‌اند.

- **تغییرناپذیری نسبت به نمونه‌ها:** در مجموعه‌داده‌ی بازشناسایی شخص، ظاهر هر تصویر می‌تواند متفاوت با ظاهر تصویر دیگر باشد، حتی اگر هر دو تصویر متعلق به یک شخص باشند. ویژگی تغییرناپذیری نسبت به نمونه‌ها از طریق یادگیری تشخیص هر تصویر واحد، به مدل بازشناسایی شخص اعمال می‌شود. این

[2] key-value structure
[3] exemplar-invariance
[4] camera-invariance
[5] neighborhood-invariance

موضوع به مدل بازشناسایی این امکان را می‌دهد که ویژگی‌های ظاهری فرد را دریافت کند. برای رسیدن به این هدف، هر نمونه تصویر دامنه‌ی هدف از سایر نمونه‌های دامنه‌ی هدف جدا درنظر گرفته می‌شود. اگر N_t تصویر در دامنه‌ی هدف موجود باشد، فرض می‌شود که به تعداد N_t کلاس مختلف نیز وجود دارد و هرتصویر در کلاس خودش طبقه‌بندی می‌شود. برای هر تصویر هدف $x_{t,i}$ ابتدا شباهت کسینوسی بین ویژگی $f(x_{t,i})$ و ویژگی‌های ذخیره شده در ExM محاسبه می‌شود. سپس احتمال پیش‌بینی شده‌ی[۶] تعلق $x_{t,i}$ به کلاس i با استفاده از تابع Softmax محاسبه می‌شود:

$$p(i|x_{t,i}) = \frac{exp(K[i]^T f(x_{t,i})/\beta)}{\sum_{j=1}^{N_t} exp(K[i]^T f(x_{t,i})/\beta))} \tag{۲.۳}$$

به‌طوری‌که در آن β عددی در بازه‌ی $(0, 1]$ است و مقیاس توزیع را تنظیم می‌کند. هدف تغییرناپذیری نسبت به نمونه‌ها، کمینه کردن احتمال درست‌نمایی لگاریتمی منفی[۷] روی تصاویر آموزشی دامنه‌ی هدف می‌باشد.

$$L_{ei} = -\log p(i|x_{t,i}) \tag{۳.۳}$$

مثالی از تغییرناپذیری نسبت به نمونه‌ها در شکل ۱.۳ نمایش داده شده است.

شکل ۱.۳: در تغییرناپذیری نسبت به نمونه‌ها، هر نمونه جدا از نمونه‌های دیگر نگه داشته می‌شود. [۱۱۲]

[6] predicted probability
[7] negative log-likelihood

- **تغییرناپذیری نسبت به دوربین‌ها:** یکی از چالش‌های مجموعه‌داده‌های بازشناسایی شخص، تنوع سبک دوربین‌هایی است که تصاویر را ثبت کرده‌اند. تصاویر موجود از یک هویت که توسط دوربین‌های مختلفی گرفته شده‌اند ممکن است تفاوت‌های قابل توجهی باهم داشته باشند. به همین دلیل تفاوت سبک دوربین‌ها یکی از فاکتورهای مهم در مسئله‌ی بازشناسایی شخص است. برای مقابله با این مسئله، یک روش داده‌افزایی روی دامنه‌ی هدف انجام شده است بدین‌ترتیب که اگر تعداد C دوربین در دامنه‌ی هدف وجود دارد، تعداد $C-1$ تصویر از هر تصویر موجود، به سبک دوربین‌های مختلف با استفاده مدل CamStyle [۱۱۴] تولید می‌شوند. هویت تصاویر تولید شده با هویت تصویر اصلی یکسان خواهد بود. بنابراین تابع اتلاف تغییرناپذیری نسبت به دوربین‌ها به‌صورت زیر می‌باشد:

$$L_{ci} = -\log p(i|\hat{x}_{t,i}) \tag{۴.۳}$$

که در آن منظور از $\hat{x}_{t,i}$ یک نمونه از تصاویر تغییر سبک داده شده‌ی تصویر $x_{t,i}$ است. بنابراین تصاویر موجود از یک نمونه با سبک دوربین‌های مختلف، به‌هم نزدیک می‌شوند. مثالی از تغییرناپذیری نسبت به دوربین‌ها در شکل ۲.۳ نمایش داده شده است.

شکل ۲.۳: تصویر یک نمونه و تصاویر تولید شده از آن نمونه با سبک دوربین‌های مختلف، به‌هم نزدیک می‌شوند درحالی‌که تصاویر تولید شده از یک هویت از تصاویر هویت‌های دیگر دور نگه داشته می‌شوند. [۱۱۲]

- **تغییرناپذیری نسبت به همسایه‌ها:** در دامنه‌ی هدف از هر هویت تعدادی تصویر موجود است اما به‌دلیل در دسترس نبودن برچسب‌های تصاویر، نمونه‌های مثبت مربوط به یک هویت واحد، قابل تشخیص

نیستند. اگر این نمونه‌های مثبت در فرآیند آموزش قابل استخراج باشند، می‌توان مدل بازشناسایی با قابلیت تعمیم‌پذیری بیشتری روی دامنه‌ی هدف داشت. برای رسیدن به این مقصود، ابتدا شباهت کسینوسی بین $f(x_{t,i})$ و ویژگی‌های ذخیره شده در کلیدهای ExM محاسبه می‌شود. سپس k نزدیک‌ترین همسایه به نمونه‌ی $x_{t,i}$ از میان کلیدهای ExM پیدا شده و اندیس‌های آن‌ها در $M(x_{t,i}, k)$ ذخیره می‌شوند. k اندازه‌ی $M(x_{t,i}, k)$ می‌باشد و نزدیک‌ترین نمونه در $M(x_{t,i}, k)$، i است. برای درنظرگرفتن این ویژگی در مدل بازشناسایی، باید تصویر هدف $x_{t,i}$ متعلق به کلاس‌های کاندیداهای موجود در $M(x_{t,i}, k)$ باشد. بنابراین وزن احتمال تعلق $x_{t,i}$ به کلاس j به‌شکل زیر است:

$$w_{i,j} = \begin{cases} \frac{۱}{k}, & j \neq i \\ ۱, & j = i \end{cases}, \forall j \in M(x_{t,i}, k) \tag{۵.۳}$$

مقصود تغییرناپذیری نسبت به همسایه‌ها به‌صورت یک تابع اتلاف برچسب-نرم[8] فرمول‌بندی می‌شود:

$$L_{ni} = -\sum_{j \neq i} w_{i,j} \log p(j|x_{t,i}), \forall j \in M(x_{t,i}, k) \tag{۶.۳}$$

مثالی از تغییرناپذیری نسبت به همسایه‌ها در شکل ۳.۳ نمایش داده شده است.

شکل ۳.۳: یک نمونه و همسایه‌های آن به‌هم نزدیک می‌شوند. [۱۱۲]

[8] soft-label loss

با درنظر گرفتن سه ویژگی بررسی شده در دامنه‌ی هدف، تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف به‌صورت زیر تعریف می‌شود:

$$L_{tgt} = -\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{j} w_{i,j}\log p(j|x^*_{t,i}) \tag{۷.۳}$$

که در آن $x^*_{t,i}$ تصویری است که به‌صورت تصادفی از اجتماع مجموعه‌ی تصویر $x_{t,i}$ و تصاویر تغییر سبک‌داده شده‌ی متناظرش نمونه‌برداری می‌شود و $j \in M(x^*_{t,i}, k)$ می‌باشد. n_t تعداد تصاویر دامنه‌ی هدف در یک دسته است. در فرمول ۱۴.۳ اگر $i = j$ باشد، شبکه به یادگیری تغییرناپذیری نسبت به نمونه‌ها و یادگیری تغییرناپذیری نسبت به دوربین‌ها مجهز است و به نمونه‌ی $x^*_{t,i}$ برچسب کلاس خودش را اختصاص می‌دهد. اما اگر $i \neq j$ باشد، شبکه به یادگیری تغییرناپذیری نسبت به همسایه‌ها مجهز است و $x^*_{t,i}$ را به همسایه‌های موجود در $M(x^*_{t,i}, k)$ نزدیک می‌کند.

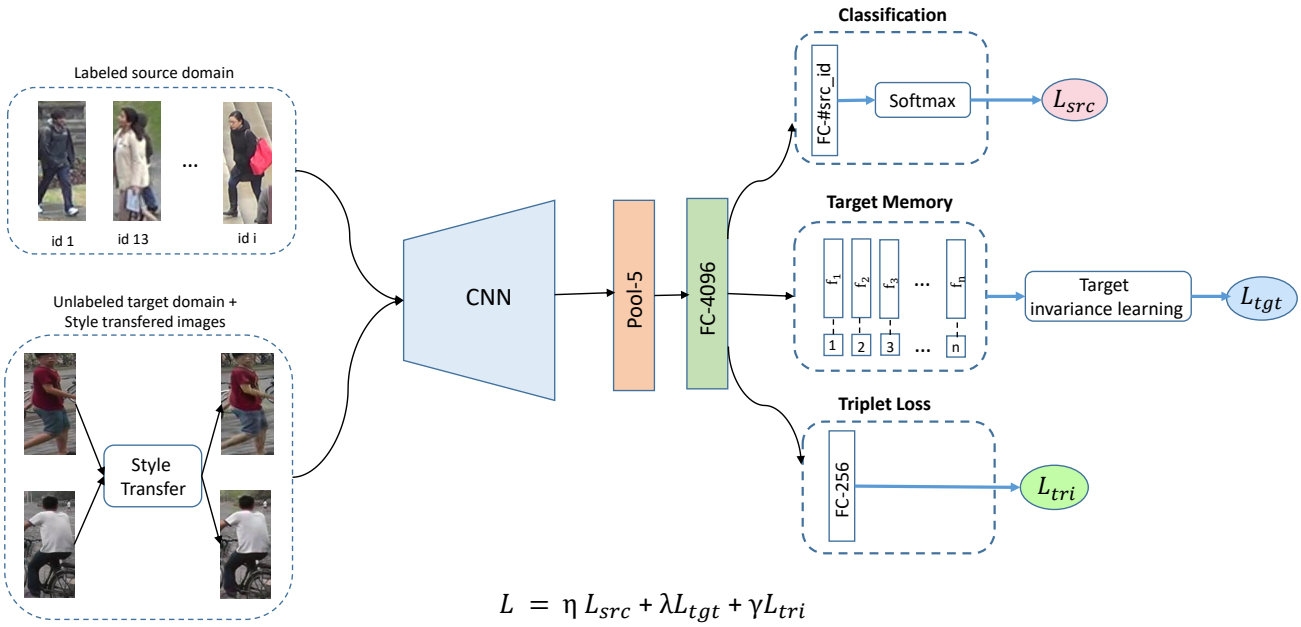
تابع اتلاف کلی شبکه از ترکیب تابع اتلاف دامنه‌ی منبع و تابع اتلاف دامنه‌ی هدف تشکیل می‌شود. میزان تأثیر تابع اتلاف منبع و تابع اتلاف هدف توسط ضریب λ که عددی در بازه‌ی $(0, 1]$ است کنترل می‌شود.

$$L = (1-\lambda)L_{src} + \lambda L_{tgt} \tag{۸.۳}$$

۳.۳ روش پیشنهادی

در روش پیشنهادی، از ایده‌ی حافظه‌ی نمونه و یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، که در مقاله‌ی [۱۱۲] معرفی شده‌اند، استفاده شده است. معرفی ساختار و جزئیات این حافظه و فرآیند یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف در بخش ۲.۳ با جزئیات کامل مورد بحث قرار گرفته است.

در معماری ارائه شده، علاوه‌بر تابع اتلاف طبقه‌بندی داده‌های منبع (L_{src}) و تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف (L_{tgt})، یک تابع اتلاف سه‌گانه (L_{tri}) نیز برای محاسبه‌ی تابع اتلاف نهایی شبکه استفاده می‌شود. اجزای مدل پیشنهادی در شکل ۴.۳ نمایش داده شده است.



شکل ۴.۳: مدل از داده‌های دارای برچسب دامنه‌ی منبع و داده‌های بدون برچسب دامنه‌ی هدف یاد می‌گیرد. مدل شامل سه بخش است:۱)طبقه‌بندی داده‌های دارای برچسب دامنه‌ی منبع ۲) یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف ۳) تابع اتلاف سه‌گانه

- **تابع اتلاف طبقه‌بندی داده‌های منبع** (L_{src})

منظور از تابع اتلاف طبقه‌بندی داده‌های منبع، همان تابع اتلاف شناسایی است که از طریق آنتروپی متقاطع محاسبه می‌شود. با دراختیار داشتن داده‌های دارای برچسب دامنه‌ی منبع، می‌توان به فرآیند آموزش به‌عنوان یک مسئله‌ی طبقه‌بندی نگاه کرد. اگر تصویر ورودی x_i برچسب y_i را داشته باشد، احتمال تعلق برچسب y_i به داده‌ی x_i با تابع Softmax کدگذاری می‌شود. سپس تابع اتلاف طبقه‌بندی داده‌های منبع با استفاده از آنتروپی متقاطع به‌دست می‌آید:

$$L_{src} = -\frac{۱}{n_s}\sum_{i=۱}^{n_s}\log(p(y_i|x_i)) \qquad (۹.۳)$$

که در آن n_s تعداد تصاویر آموزشی دارای برچسب در یک دسته می‌باشد.

- **تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف** (L_{tgt})

برای بهبود عملکرد مدل هنگام آزمایش روی مجموعه‌داده‌ی هدف، می‌توان در هنگام انتقال دانش از مجموعه‌داده‌ی دارای برچسب منبع به مجموعه‌داده‌ی بدون برچسب هدف، ویژگی‌های درون-دامنه‌ای در دامنه‌ی هدف را بررسی کرد. یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف باعث مقاوم شدن مدل آموزش دیده، دربرابر تغییرات درون-دامنه‌ای در دامنه‌ی هدف می‌شود. درنتیجه عملکرد مدل هنگام آزمایش روی مجموعه‌داده‌ی هدف بهبود می‌یابد. در مقاله‌ی پایه [۱۱۲] ، سه ویژگی درون-دامنه‌ای در دامنه‌ی هدف مورد بررسی قرار گرفته است که در بخش ۲.۲.۳ با جزئیات بیان شده‌اند.

در تغییرناپذیری نسبت به نمونه‌ها، هر تصویر نمونه در دامنه‌ی هدف، مجزا از سایر نمونه‌ها درنظر گرفته می‌شود و باعث می‌شود که مدل بازشناسایی شخص بازنمایی ظاهری تصویر فرد را دریافت کند. در تغییرناپذیری نسبت به دوربین‌ها، به ازای هر تصویر موجود در دامنه‌ی هدف، تصاویر جدیدی با سبک سایر دوربین‌های دامنه‌ی هدف تولید می‌شوند و تصویر اصلی و تصاویر تولید شده‌ی متناظر با آن تصویر، به‌هم نزدیک می‌شوند. این موضوع باعث می‌شود که مدل آموزش دیده نسبت به تغییر سبک دوربین‌ها در دامنه‌ی هدف مقاوم‌تر شود. در تغییرناپذیری نسبت به همسایه‌ها، یک تصویر نمونه و مشابه‌ترین نمونه‌های موجود در دامنه‌ی هدف به آن تصویر نمونه، به یکدیگر نزدیک می‌شوند. با درنظر گرفتن یادگیری تغییرناپذیری نسبت به نمونه‌ها، تغییرناپذیری نسبت به دوربین‌ها و تغییرناپذیری نسبت به همسایه‌ها در دامنه‌ی هدف، تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف مشابه فرمول ۱۴.۳ تعریف می‌شود:

$$L_{tgt} = -\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{j} w_{i,j}\log p(j|x^*_{t,i}) \tag{۱۴.۳}$$

که در آن $x^*_{t,i}$ تصویری است که به‌صورت تصادفی از اجتماع مجموعه‌ی تصویر $x_{t,i}$ و تصاویر تغییر سبک‌داده شده‌ی متناظرش نمونه‌برداری می‌شود و $j \in M(x^*_{t,i}, k)$ می‌باشد. n_t تعداد تصاویر دامنه‌ی هدف در یک دسته‌ی آموزشی است. اگر $i = j$ باشد، شبکه به یادگیری تغییرناپذیری نسبت به نمونه‌ها و یادگیری تغییرناپذیری نسبت به دوربین‌ها مجهز است و به نمونه‌ی $x^*_{t,i}$ برچسب کلاس خودش را اختصاص می‌دهد. اما اگر $i \neq j$ باشد، شبکه به یادگیری تغییرناپذیری نسبت به همسایه‌ها نیز مجهز است و $x^*_{t,i}$ را به همسایه‌های موجود در $M(x^*_{t,i}, k)$ نزدیک می‌کند.

- **تابع اتلاف سه‌گانه** (L_{tri})

تابع اتلاف سه‌گانه، نیاز به مجموعه‌ای از سه‌تایی‌ها دارد که تشکیل شده‌اند از: یک تصویر لنگر $x_{1,i}$، یک نمونه‌ی مثبت $x_{2,i}$ (هویت $x_{1,i}$ و $x_{2,i}$ یکسان است.) و یک نمونه‌ی منفی $x_{3,i}$ (هویت $x_{1,i}$ و $x_{3,i}$ باهم متفاوت می‌باشد.). هدف تابع اتلاف سه‌گانه، کمینه‌کردن تفاوت‌های درون-کلاسی و بیشینه‌کردن تفاوت‌های بیرون-کلاسی می‌باشد. تابع اتلاف سه‌گانه، بر اساس این ایده عمل می‌کند که فاصله‌ی بین زوج‌های مثبت باید کمتر از فاصله‌ی بین زوج‌های منفی باشد. با فرض وجود سه‌تایی $< x_{1,i}, x_{2,i}, x_{3,i} >$ مدل سعی می‌کند شرط زیر را برقرار کند:

$$||f(x_{1,i}) - f(x_{2,i})||^2 < ||f(x_{1,i}) - f(x_{3,i})||^2 \qquad (10.3)$$

که در آن $f(x_{1,i})$ نمایانگر ویژگی مربوط به تصویر $x_{1,i}$ است که توسط شبکه تولید می‌شود. $||.||^2$ نمایانگر نرمال‌سازی $L2$ است.

با این فرض می‌توان تابع اتلاف سه‌گانه را به‌صورت زیر تعریف کرد:

$$L_{triplet}(X) = \sum_{i=1}^{n} \max\{||f(x_{1,i}) - f(x_{2,i})||^2 - ||f(x_{1,i}) - f(x_{3,i})||^2 + \rho, 0\} \quad (11.3)$$

که در آن X مجموعه‌ای از سه‌تایی ها است و $X_i =< x_{1,i}, x_{2,i}, x_{3,i} >$. n تعداد سه‌تایی‌های موجود در مجموعه‌ی X است و پارامتر ρ، ثابت حاشیه نام دارد.

در مسئله‌ی بازشناسایی شخص، هویت‌های موجود در دو مجموعه‌داده‌ی منبع و هدف کاملاً متفاوت هستند. درواقع هر نمونه از دامنه‌ی منبع $(x_{s,i})$ و هر نمونه از دامنه‌ی هدف $(x_{t,i})$ متعلق به کلاس‌های مختلفی می‌باشند و یک زوج منفی را تشکیل می‌دهند. ازطرفی، تصاویر تولید شده با سبک دوربین‌های مختلف از یک تصویر دامنه‌ی هدف $(x_{t^*,i})$ ، هویتی مشابه با تصویر اصلی $(x_{t,i})$ دارند. بنابراین یک تصویر از دامنه‌ی هدف با یکی از تصاویر تغییرسبک‌داده شده‌ی متناظرش، یک زوج مثبت را می‌سازند. ایده‌ی تابع اتلاف سه‌گانه‌ی پیشنهادی، از مقاله‌ی [۱۱۱] برگرفته شده است. با داشتن سه‌تایی‌هایی تشکیل شده از یک نمونه از دامنه‌ی منبع، یک نمونه از دامنه‌ی هدف و یک نمونه از تصاویر تولید شده‌ی متناظر

با نمونه‌ی دامنه‌ی هدف، می‌توان تابع اتلاف سه‌گانه را تعریف نمود:

$$L_{tri} = L_{triplet}(\{x_{t,i}\}_{i=1}^{n_t}, \{x_{t^*,i}\}_{i=1}^{n_t^*}, \{x_{s,i}\}_{i=1}^{n_s},) \tag{۱۲.۳}$$

نمونه‌هایی از سه‌تایی‌های مدنظر در تصویر ۵.۳ نمایش داده شده‌اند.

شکل ۵.۳: مثال‌های از سه‌تایی‌های مناسب برای تابع اتلاف سه‌گانه L_{tri}

استفاده از تابع اتلاف سه‌گانه‌ی فرمول ۱۲.۳ ، باعث می‌شود که نمونه‌های دامنه‌ی منبع از نمونه‌های دامنه‌ی هدف دور شوند و همچنین نمونه‌های دامنه‌ی هدف به نمونه‌های تولید شده‌ی متناظر خود نزدیک شوند.

- **تابع اتلاف کلی شبکه** (L)

در روش پیشنهادی تابع اتلاف کلی سیستم به‌صورت زیر تعریف می‌شود:

$$L = \eta L_{src} + \lambda L_{tgt} + \gamma L_{tri} \tag{۱۳.۳}$$

η، λ و γ اعداد ثابتی هستند و میزان تأثیر توابع اتلاف L_{src}، L_{tgt} و L_{tri} را در تابع اتلاف نهایی، کنترل می‌کنند.

در تابع اتلاف نهایی شبکه، L_{src} باعث می‌شود مدل ساختار و ویژگی‌های درون-دامنه‌ای در دامنه‌ی منبع را یاد بگیرد و نسبت به تغییرات درون-دامنه‌ای در دامنه‌ی هدف که مشابه با تغییرات درون-دامنه‌ای در دامنه‌ی منبع می‌باشند، مقاوم‌تر شود. L_{tgt} مدل را نسبت به تغییرات درون-دامنه‌ای در دامنه‌ی هدف شامل تغییرات دوربین‌ها، تغییرات نمونه‌ها و تغییرات همسایه‌ها مقاوم می‌سازد و باعث می‌شود مدل از دامنه‌ی بدون برچسب هدف نیز ویژگی‌ها و بازنمایی‌هایی را یاد بگیرد. درنتیجه قابلیت تعمیم‌پذیری مدل هنگام آزمایش روی مجموعه‌داده‌ی هدف بهبود یابد. L_{tri} با استفاده از تابع اتلاف سه‌گانه، مدل را نسبت به تغییرات دوربین‌ها در دامنه‌ی هدف مقاوم می‌سازد و باعث می‌شود مدل تفاوت‌های بین دامنه‌ای در دامنه‌های منبع و هدف را یاد بگیرد.

۱.۳.۳ بررسی استراتژی انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها

برای انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها، دو استراتژی مورد آزمایش قرار می‌گیرند. در استراتژی اول برای هر تصویر دامنه‌ی هدف، پس از محاسبه‌ی شباهت ویژگی آن تصویر ($f(x_{t,i})$) با ویژگی‌های ذخیره شده در حافظه، تعداد k نزدیک‌ترین همسایه از میان نمونه‌های موجود در حافظه، انتخاب می‌شود. درنتیجه برای تمامی تصاویر دامنه‌ی هدف به تعداد ثابت k همسایه انتخاب خواهد شد. این روش مشابه با روش انتخاب همسایه‌ها در مقاله‌ی پایه [۱۱۲] است که در بخش ۲.۲.۳ توضیح داده شده است. با این استراتژی، تابع اتلاف تغییرناپذیری نسبت به همسایه‌ها از فرمول ۶.۳ محاسبه می‌شود و تابع اتلاف L_{tgt} مشابه با فرمول ۱۴.۳ خواهد بود:

$$L_{tgt} = -\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{j} w_{i,j}\log p(j|x^*_{t,i}) \tag{۱۴.۳}$$

استراتژی دوم برای انتخاب همسایه‌ها این است که یک آستانه‌ای درنظر گرفته شود و پس از محاسبه‌ی شباهت ویژگی تصویر مدنظر با ویژگی‌های ذخیره شده در حافظه، نمونه‌هایی که به حد کافی به تصویر مدنظر نزدیک هستند به‌عنوان همسایه‌های آن انتخاب می‌شوند. مشکلی که استراتژی دوم دارد این است که تعداد همسایه‌های

یک تصویر می‌تواند متغیر باشد. اگر یک تصویر تعداد همسایه‌های زیادی داشته باشد، مجموع اتلاف بین آن تصویر و همسایه‌هایش بسیار زیاد می‌شود. اگر تصویری تعداد همسایه‌های کمی داشته باشد، مجموع اتلاف بین آن تصویر و همسایه‌هایش می‌تواند بسیار کم شود. این عدم توازن می‌تواند باعث عملکرد ضعیف مدل بازشناسایی شود. در مقاله‌ی [۱۹] راهکاری برای این مسئله ارائه شده است. برای مقابله با این مشکل، یک جریمه درنظر گرفته می‌شود. اگر تعداد همسایه‌های تصویر $x_{t,i}$ برابر با $||w_i||_۱$ باشد ، جریمه‌ی $۱/||w_i||_۱ \log(||w_i||_۱)$ وارد فرمول محاسبه‌ی تابع اتلاف L_{tgt} می‌شود.

$$L_{tgt} = -\frac{۱}{n_t} \sum_{i=۱,||w_i||_۱ \geq ۲}^{n_t} \frac{۱}{||w_i||_۱ \log(||w_i||_۱)} \sum_j w_{i,j} \log p(j|x^*_{t,i}) \qquad (۱۴.۳)$$

که در آن $w_{i,j} \in \{۰,۱\}$ است. اگر $w_{i,j} = ۱$ باشد یعنی $x_{t,j}$ به‌عنوان همسایه‌ی $x_{t,i}$ انتخاب شده است. اگر $||w_i||_۱ = ۱$ باشد، مخرج عبارت جریمه بی‌معنا خواهد بود و بیانگر این است که تصویر مربوطه به غیر از خودش هیچ همسایه‌ای ندارد. بنابراین از آن تصویر به‌عنوان نمونه‌ی آموزشی استفاده نخواهد شد. اگر تصویری تعداد همسایه‌های زیادی داشته باشد، جریمه‌ی $۱/||w_i||_۱ \log(||w_i||_۱)$ باعث کاهش اتلاف بین آن تصویر و همسایه‌هایش می‌شود. با این استراتژی، تعداد همسایه‌ها در طی فرایند آموزش، معقول و متوازن می‌شود.

الگوریتم ۱.۳ نحوه‌ی انتخاب همسایه‌ها در استراتژی اول انتخاب همسایه در یادگیری تغییرناپذیری نسبت به همسایه‌ها و الگوریتم ۲.۳ نحوه‌ی انتخاب همسایه‌ها در استراتژی دوم انتخاب همسایه در یادگیری تغییرناپذیری نسبت به همسایه‌ها را بیان می‌کنند. به‌منظور بررسی تأثیر نحوه‌ی انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها، در فصل آینده نتایج آزمایشات انجام شده روی مدل با استفاده از هر دو استراتژی آورده می‌شود.

۴.۳ خلاصه و نتیجه‌گیری

در این فصل ابتدا مدل پایه معرفی شده است. در مدل پایه سه ویژگی درون–دامنه‌ای در دامنه‌ی هدف مورد بررسی قرار می‌گیرند. درنتیجه عملکرد مدل هنگام آزمایش روی مجموعه‌داده‌ی بدون برچسب دامنه‌ی هدف بهبود می‌یابد. تابع اتلاف کلی شبکه در مدل پایه، از تابع اتلاف طبقه‌بندی داده‌های برچسب‌گذاری شده‌ی

دامنه‌ی منبع و تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف تشکیل می‌شود. در یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، ویژگی‌های تغییرناپذیری نسبت به نمونه‌ها، تغییرناپذیری نسبت به دوربین‌ها و تغییرناپذیری نسبت به همسایه‌ها درنظر گرفته شده‌اند. در یادگیری تغییرناپذیری نسبت به نمونه‌ها، هر نمونه در دامنه‌ی هدف مستقل از سایر نمونه‌ها فرض می‌شود. در یادگیری تغییرناپذیری نسبت به دوربین‌ها، یک تصویر از دامنه‌ی هدف و تصاویر تولیدشده از آن تصویر با سبک سایر دوربین‌ها به‌هم نزدیک می‌شوند. بدین‌ترتیب مدل نسبت به تفاوت سبک دوربین‌ها در دامنه‌ی هدف مقاوم‌تر می‌شود. در یادگیری تغییرناپذیری نسبت به همسایه‌ها، برای هر تصویر از دامنه‌ی هدف، تعداد k نزدیک‌ترین نمونه‌ها به آن تصویر، به‌عنوان همسایه‌های آن تصویر درنظرگرفته می‌شوند.

پس از تشریح مدل پایه و اجزای آن، روش پیشنهادی توضیح داده شده است. در روش پیشنهادی به‌منظور استخراج ویژگی از داده‌های دامنه‌ی منبع و دامنه‌ی هدف، از شبکه‌ی پیش‌آموزش‌دیده‌ی ResNeXt-50 استفاده شده است. در روش پیشنهادی علاوه بر تابع اتلاف طبقه‌بندی داده‌های منبع و تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، از یک تابع اتلاف سه‌گانه نیز استفاده شده است. سه‌تایی‌های این تابع اتلاف سه‌گانه از یک نمونه تصویر از دامنه‌ی هدف (تصویر لنگر)، یک نمونه تصویر از دامنه‌ی منبع (نمونه‌ی منفی) و یک نمونه تصویر از تصاویر تولیدشده از تصویر لنگر به سبک سایر دوربین‌های دامنه‌ی هدف (نمونه‌ی مثبت)، تشکیل شده‌اند. بدین‌ترتیب این تابع اتلاف سه‌گانه علاوه‌بر یادگیری تفاوت‌های درون-دامنه‌ای در دامنه‌ی هدف، تفاوت‌های بین دامنه‌های منبع و هدف را نیز یاد می‌گیرد.

برای روش انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها، دو استراتژی مختلف بیان شده است. استراتژی اول مشابه با مقاله‌ی پایه است و در آن برای هر تصویر از دامنه‌ی هدف، به تعداد k تصویر همسایه انتخاب می‌شود. اما در استراتژی دوم، تصاویری که فاصله‌ی آن‌ها از یک نمونه تصویر دامنه‌ی هدف، کمتر از مقدار آستانه‌ی μ باشد، به‌عنوان همسایه‌های آن تصویر انتخاب می‌شوند.

الگوریتم ۱.۳ استراتژی اول انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها

Require: $\{x_{t,i}\}_{i=1}^{n_t}$: Unlabeled target data,
$\{x_{s,i}, y_{s,i}\}_{i=1}^{n_s}$: Labeled source data,
k: Number of candidate positive samples in neighborhood-invariance learning,
α: Update rate for ExM,
E: Number of epochs.

```
for e = 0 to E − 1 do
    for i = 1 to n_t do
        extract feature f_i
        calculate distance d(K, f_i)
        sort d(K, f_i)
        select k more similar images to x_{t,i} and add their features into M(x_{t,i}, k)
        for j = 1 to n_t do
            if K[j] ∈ M(x_{t,i}, k) then
                w_{i,j} = 1/k; //selected
            else
                w_{i,j} = 0; //removed
            end if
        end for
    end for
    //model training
    train the model with labeled source data and unlabeled target data.
    //keys in feature memory ExM update
    for j = 1 to n_t do
        K[i] ← αK[i] + (1 − α)f_i
    end for
end for
```

الگوریتم ۲.۳ استراتژی دوم انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها

Require: $\{x_{t,i}\}_{i=1}^{n_t}$: Unlabeled target data,
$\{x_{s,i}, y_{s,i}\}_{i=1}^{n_s}$: Labeled source data,
μ: Similarity threshold,
α: Update rate for ExM,
E: Number of epochs.

```
for e = 0 to E − 1 do
    for i = 1 to n_t do
        extract feature f_i
        calculate distance d(K, f_i)
        for j = 1 to n_t do
            if d(K[j], f_i) ≤ μ then
                w_{i,j} = 1; //selected
            else
                w_{i,j} = 0; //removed
            end if
        end for
    end for
    //model training
    train the model with labeled source data and unlabeled target data.
    //keys in feature memory ExM update
    for j = 1 to n_t do
        K[i] ← αK[i] + (1 − α)f_i
    end for
end for
```

فصل ۴

نتایج علمی

۱.۴ پیش‌گفتار

در این فصل، ابتدا نحوه‌ی آماده‌سازی مجموعه‌های داده توضیح داده می‌شود. سپس تنظیمات مدل و آزمایش و مقادیر پارامترها بیان می‌گردد. نتایج اجراهای متعدد مدل پیشنهادی و تحلیل پارامترهای مختلف مدل در این فصل آورده شده است.

۲.۴ مجموعه‌های داده

مجموعه‌داده‌های DukeMTMC-reID و Market1501 برای آزمایش مدل پیشنهادی استفاده شده‌اند. نمونه‌هایی از تصاویر مجموعه‌داده‌های Market1501 و DukeMTMC-reID در شکل ۱.۴ نمایش داده شده است. در ادامه درمورد این دو مجموعه‌داده توضیحاتی ارائه شده و سپس نحوه‌ی آماده‌سازی داده‌ها توضیح داده‌می‌شود.

شکل ۱.۴: نمونه‌هایی از تصاویر مجموعه‌داده‌های Market1501 و DukeMTMC-reID

- مجموعه‌داده Market1501

مجموعه‌داده‌ی Market1501 در سال ۲۰۱۵ برای مسئله‌ی بازشناسایی شخص ارائه شده است. تصاویر این مجموعه‌داده در مقابل فروشگاهی در دانشگاه Tisinghua ثبت شده‌اند. در مجموع شش دوربین تصاویر را ثبت کردند. پنج دوربین دارای رزولوشن بالا و یک دوربین دارای رزولوشن پایین هستند. تصاویر هویت‌های این مجموعه‌داده حداقل توسط دو دوربین و حداکثر توسط شش دوربین ثبت شده‌اند. مجموعه‌داده‌ی Market1501 دارای ۳۲۶۶۸ تصویر از ۱۵۰۱ هویت می‌باشد. معمولاً ۷۵۱ هویت برای آموزش مورداستفاده قرار می‌گیرند و ۷۵۰ هویت نیز به‌منظور آزمایش مدل استفاده می‌شوند. تصاویر این مجموعه‌داده به سه دسته تقسیم می‌شوند، ۱۲۹۳۶ تصویر از ۷۵۱ هویت برای آموزش، ۱۹۷۳۲ تصویر از ۷۵۰ هویت در گالری و ۳۳۶۸ تصویر از همان ۷۵۰ هویت موجود در گالری، به‌عنوان پرس‌وجو موجود هستند.

- مجموعه‌داده DukeMTMC-reID

مجموعه‌داده‌ی DukeMTMC-reID زیرمجموعه‌ای از مجموعه‌داده‌ی DukeMTMC است که برای

وظیفه‌ی بازشناسایی شخص برپایه‌ی تصویر، ارائه شده است. مجموعه‌داده‌ی DukeMTMC شامل ۵۸ دقیقه ویدیوی با رزولوشن بالا از هشت دوربین مختلف می‌باشد. در DukeMTMC-reID تصاویر عابرین پیاده از ویدیوها در هر ۱۲۰ فریم، برش داده شده‌اند. در مجموع ۳۶۴۱۱ تصویر محدود شده‌ی دارای برچسب در این مجموعه‌داده وجود دارد. تصاویر ۱۴۰۴ هویت توسط بیش از دو دوربین ثبت شده‌اند و تصاویر ۴۰۸ هویت فقط توسط یک دوربین ثبت شده‌اند. ۷۰۲ هویت برای آموزش مورداستفاده قرار می‌گیرند و ۷۰۲ هویت دیگر برای آزمایش مدل استفاده می‌شوند. تصاویر موجود در این مجموعه‌داده نیز همانند مجموعه‌داده‌ی Market1501 به سه دسته تقسیم می‌شوند، ۱۶۵۲۲ تصویر از ۷۰۲ هویت برای آموزش، ۱۷۶۶۱ تصویر از ۷۰۲ هویت دیگر در گالری و ۲۲۲۸ تصویر از همان ۷۰۲ هویت موجود در گالری، به‌عنوان پرس‌وجو موجود هستند.

۱.۲.۴ آماده‌سازی داده‌ها

در فرآیند آموزش و آزمایش مدل، مجموعه‌داده‌ی DukeMTMC-reID به‌عنوان مجموعه‌داده‌ی دارای برچسب دامنه‌ی منبع و مجموعه‌داده‌ی Market1501 به‌عنوان مجموعه‌داده‌ی بدون برچسب دامنه‌ی هدف درنظر گرفته شدند. در مدل پیشنهادی، فرآیند آموزش با استفاده از دو مجموعه‌داده‌ی منبع و هدف صورت می‌گیرد و آزمایش روی مجموعه‌داده‌ی بدون برچسب هدف انجام می‌شود.

داده‌های آموزشی از دامنه‌ی منبع و هدف نرمال‌سازی شده و روش‌های داده‌افزایی پاک کردن تصادفی و جابه‌جایی افقی تصادفی[1] روی آن‌ها اعمال می‌شوند. داده‌های آزمایشی نیز نرمال‌سازی شده و ابعاد داده‌های آموزشی و آزمایشی به ۱۲۸ × ۲۵۶ تغییر می‌کند.

در دامنه‌ی هدف شماره‌ی دوربینی که هر تصویر را ثبت کرده است در دسترس است. به‌منظور یادگیری تغییرناپذیری نسبت به دوربین‌ها در دامنه‌ی هدف، اگر به تعداد C دوربین در دامنه‌ی هدف موجود باشد، برای هر تصویر دامنه‌ی هدف، به تعداد $C - ۱$ تصویر جدید، به سبک سایر دوربین‌ها، تولید می‌شود. تولید تصاویر جدید به سبک سایر دوربین‌ها با به کارگیری StarGAN [۱۳] به‌منظور آموزش یک مدل CamStyle [۱۱۴] در دامنه‌ی هدف محقق می‌شود. بنابراین تصاویر جدید تولید شده نیز به تصاویر آموزشی اضافه می‌شوند. در شکل ۲.۴ نمونه‌هایی از تصاویر مجموعه‌داده‌ی DukeMTMC-reID و تصاویر تولید شده از آن‌ها به سبک سایر دوربین‌ها،

[1] random horizontal flipping

نمایش داده شده است.

شکل ۲.۴: نمونه‌هایی از تصاویر مجموعه‌داده‌ی DukeMTMC-reID و تصاویر تولید شده از آن‌ها به سبک سایر دوربین‌ها

۳.۴ تنظیمات آزمایش

در ابتدا داده‌های آموزشی و آزمایشی آماده‌سازی می‌شوند. داده‌های آموزشی از دامنه‌ی هدف و منبع نرمال‌سازی شده و داده‌افزایی روی آن‌ها اعمال می‌شود. همچنین اندازه‌ی تصاویر ورودی به ۱۲۸ × ۲۵۶ تبدیل می‌شود. پس از آماده‌سازی داده‌ها فرآیند آموزش مدل آغاز می‌شود. برای آموزش مدل، از مدل پیش‌آموزش دیده‌ی ResNeXt-50 [۹۱] که پارامترهای آن روی مجموعه‌داده‌ی ImageNet [۱۶] آموزش دیده‌اند، استفاده شده است. به‌منظور صرفه‌جویی در حافظه‌ی GPU ، دو لایه‌ی باقی‌مانده‌ی[۲] ابتدایی ثابت می‌شوند. پس از لایه‌ی Pool-5 در ResNeXt-50 یک لایه‌ی تماماً متصل با اندازه‌ی ۴۰۹۶ اضافه شده و نرخ برون‌اندازی با احتمال ۰٫۵، نرمال‌سازی دسته‌ای و ReLU پس از آن لایه اعمال می‌شوند.

پس از مرحله‌ی استخراج ویژگی، معماری دارای سه شاخه‌ی مختلف می‌شود. شاخه‌ی اول مربوط به طبقه‌بندی داده‌های دارای برچسب دامنه‌ی منبع می‌باشد. برای این هدف یک لایه‌ی تماماً متصل با اندازه‌ی تعداد هویت‌های موجود در دامنه‌ی منبع، به مدل اضافه می‌شود. تابع اتلاف مربوط به شاخه‌ی اول همان L_{src} است.

[2] residual layer

میزان تأثیر این تابع اتلاف در تابع اتلاف نهایی شبکه، با پارامتر η کنترل می‌شود که مقدار آن در استراتژی اول برابر با ۰٫۶ و در استراتژی دوم در تنظیمات duke $\rightarrow$ market برابر با ۰٫۴ و در تنظیمات market $\rightarrow$ duke برابر با ۰٫۶ درنظر گرفته شده است.

شاخه‌ی دوم مربوط به یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف می‌باشد. ویژگی‌های استخراج شده از تصاویر دامنه‌ی هدف، در حافظه‌ی نمونه ذخیره می‌شوند. نرخ به‌روزرسانی حافظه‌ی نمونه، α در استراتژی اول برابر با ۰٫۶ و در استراتژی دوم روی duke $\rightarrow$ market برابر با ۰٫۴ و روی market $\rightarrow$ duke برابر با ۰٫۶ درنظر گرفته شده است. با افزایش دوره[3]، مقدار α نیز افزایش می‌یابد. در پنج دوره‌ی اول، از یادگیری تغییرناپذیری نسبت نمونه‌ها و یادگیری تغییرناپذیری نسبت به دوربین‌ها استفاده می‌شود. پس از پنج دوره، یادگیری تغییرناپذیری نسبت به همسایه‌ها نیز اضافه می‌شود. در استراتژی اول انتخاب همسایه‌ها، مقدار k (تعداد نمونه‌های مثبت همسایگی) برابر با ۶ درنظر گرفته شده است. در استراتژی دوم انتخاب همسایه‌ها، مقدار μ (مقدار آستانه‌ی شباهت یک تصویر با همسایه‌هایش) برابر با ۰٫۵۵ درنظر گرفته شده است. درواقع در استراتژی اول، برای هر تصویر به تعداد ثابت k همسایه انتخاب می‌شود درحالی‌که در استراتژی دوم، دو تصویر به‌عنوان همسایه انتخاب می‌شوند اگر شباهت کسینوسی آن‌ها از مقدار آستانه‌ی μ بیشتر باشد. بنابراین در استراتژی دوم، تعداد همسایه‌ها ثابت نیست. تابع اتلاف مربوط به شاخه‌ی دوم همان L_{tgt} است. میزان تأثیر این تابع اتلاف در تابع اتلاف نهایی شبکه با پارامتر λ کنترل می‌شود. که مقدار آن در استراتژی اول برابر با ۰٫۴ و در استراتژی دوم در تنظیمات duke $\rightarrow$ market برابر با ۰٫۶ و در تنظیمات market $\rightarrow$ duke برابر با ۰٫۴ درنظر گرفته شده است.

شاخه‌ی سوم مربوط به تابع اتلاف سه‌گانه است. در این شاخه یک لایه‌ی تماماً متصل با اندازه‌ی ۲۵۶ اضافه می‌شود. خروجی این لایه به‌عنوان ویژگی مورداستفاده در محاسبه‌ی تابع اتلاف سه‌گانه کاربرد دارد. مقدار پارامتر حاشیه‌ای ρ در تابع اتلاف سه‌گانه برابر ۰٫۳ درنظر گرفته شده است. میزان تأثیر این تابع اتلاف سه‌گانه در تابع اتلاف نهایی شبکه با پارامتر γ کنترل می‌شود. مقدار پارامتر γ در هر دو استراتژی برابر با ۰٫۵ است.

برای آموزش مدل از بهینه‌ساز SGD با مومنتوم برابر با ۰٫۹ و کاهش وزنی[4] برابر با ۴×۱۰^{-۵} استفاده شده است. در ۴۰ دوره‌ی اول نرخ یادگیری برای لایه‌های پایه در ResNeXt-50 برابر با ۰٫۰۱ و برای سایر لایه‌ها برابر با ۰٫۱ می‌باشد. پس از ۴۰ دوره نرخ یادگیری بر ۱۰ تقسیم می‌شود. اندازه‌ی دسته برای داده‌های منبع و هدف برابر ۱۲۸ و برای داده‌های سه‌گانه برابر ۶۴ درنظر گرفته می‌شود.

[3] epoch
[4] weight-decay

خروجی مدل بازشناسایی شخص برای هر تصویر پرس‌وجو از دامنه‌ی هدف، یک لیستی از تصاویر است که باید هویتی مشابه با هویت پرس‌وجو داشته باشند. در شکل ۳.۴ مثالی از دو نمونه تصویر پرس‌و جو و چند نمونه‌ی ابتدای لیست بازیابی شده قابل مشاهده است.

شکل ۳.۴: دو نمونه تصویر پرس‌وجو و چند نمونه‌ی ابتدای لیست بازیابی شده

۱.۳.۴ مشخصات سخت‌افزاری و نرم‌افزاری

مشخصات سخت‌افزاری و نرم‌افزاری بستری که آزمایش‌ها روی آن انجام گرفته است عبارتند از :

کارت گرافیک : Tesla P100 16 GB

میزان حافظه : 25 GB

فضای ذخیره‌سازی : 68 GB

نسخه‌ی python : 3.6.9

نسخه‌ی pytorch : 1.7.0

۴.۴ نتایج آزمایش

در جدول ۱.۴ و جدول ۲.۴ نتایج چهار مرتبه اجرای مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها و چهار مرتبه اجرای مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها برای ۶۰ دوره در تنظیمات duke → market و market → duke آورده شده است. از آنجایی که نتیجه‌ی مدل پایه در مقاله‌ی پایه [۱۱۲] برای ۶۰ دوره ثبت شده، به‌منظور مقایسه‌ی بهتر، نتیجه‌ی مدل پیشنهادی نیز برای ۶۰ دوره آورده شده است.

همان‌طور که در جدول ۱.۴ و جدول ۲.۴ قابل مشاهده است، استفاده از استراتژی دوم انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها، موجب بهبود عملکرد مدل می‌شود. البته مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها نیز نسبت به مدل پایه [۱۱۲] ، عملکرد بهتری دارد. این بهبود عملکرد نسبت به مدل پایه [۱۱۲] ، حاصل بهره بردن از تابع اتلاف سه‌گانه برای یادگیری تفاوت‌های بین-دامنه‌ای میان دو دامنه‌ی منبع و هدف و همچنین یادگیری تفاوت‌های درون-دامنه‌ای در دامنه‌ی هدف می‌باشد.

جدول ۱.۴: نتایج چندین اجرای مدل پیشنهادی با استراتژی اول و دوم انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها در تنظیمات duke → market

DukeMTMC-reID → Market1501						
R-1(%)	R-5(%)	R-10(%)	R-20(%)	mAP(%)	epochs	
77.4	**89.1**	**92.7**	**95.3**	**46**	60	strategy1
77.6	88.8	92.3	95.1	45.6	60	
76.8	88.8	92.3	95.1	46	60	
76.6	88.6	92	94.1	45.3	60	
84.5	**93.1**	**95.7**	**97.2**	63	60	strategy2
84.6	93.1	95.7	97.2	62.8	60	
84.4	92.9	95.4	97.2	62.4	60	
84.1	93	95.7	97.2	**63.1**	60	

جدول ۲.۴: نتایج چندین اجرای مدل پیشنهادی با استراتژی اول و دوم انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها در تنظیمات market → duke

Market1501 → DukeMTMC-reID						
R-1(%)	R-5(%)	R-10(%)	R-20(%)	mAP(%)	epochs	
64	**76.8**	**80.9**	**84.1**	**41.1**	60	strategy1
64	76.7	80.2	83.4	41.1	60	
64	76	79.9	83.5	40.9	60	
64.4	75.4	80.1	83.8	40.2	60	
70.1	**80.8**	**84.4**	**87.1**	**49.1**	60	strategy2
70.6	80.1	84.1	86.8	48.9	60	
69.8	80.8	84	86.4	48.9	60	
69.6	80.6	83.8	86	48.8	60	

۱.۰.۴.۴ نتایج استفاده از معماری‌های مختلف CNN در عملکرد مدل

به‌منظور بررسی تأثیر معماری CNN در عملکرد مدل، آزمایش‌های متعددی انجام شده است. در جدول ۳.۴ نتایج آزمایش‌های انجام شده با معماری‌های ResNet-50 ، ResNeXt-50 و WideResNet-50 آورده شده است. در هر سه معماری، به‌منظور صرفه‌جویی در حافظه‌ی GPU ، دو لایه‌ی باقی‌مانده‌ی ابتدایی ثابت شدند. معمولاً در مدل‌های بازشناسایی شخص از معماری ResNet-50 برای استخراج ویژگی استفاده می‌شود. در مدل پیشنهادی نتایج آزمایشات نشان می‌دهد که عملکرد مدل هنگام استفاده از ResNeXt-50 بهتر از عملکرد هنگام استفاده از ResNet-50 می‌باشد. علاوه‌براین، پارامترهای قابل آموزش ResNeXt-50 کمتر از ResNet-50 است و این مسئله می‌تواند نقطه‌ی قوت محسوب شود. عملکرد مدل هنگام استفاده از WideResNet-50 نیز مناسب است اما به‌دلیل تعداد پارامترهای بیشتری که نسبت به دو معماری دیگر دارد، در مدل پیشنهادی استفاده نمی‌شود.

جدول ۳.۴: نتایج آزمایش مدل با به کارگیری معماری‌های مختلف. در سه سطر ابتدایی جدول نتایج اجرای مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها و در سه سطر آخر جدول نتایج اجرای مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها آورده شده است.

CNN Architecture	Trainable Parameters	Market→Duke (%) R-1	Market→Duke (%) mAP	Duke→Market (%) R-1	Duke→Market (%) mAP	
ResNet-50	≈ 36 million	76.7	45.5	63.4	40.8	strategy1
ResNeXt-50	≈ 35 million	**77.4**	**46**	**64**	**41.1**	strategy1
WideResNet-50	≈77 million	76	43.2	63.8	41	strategy1
ResNet-50	≈36 million	83.3	60.5	67.7	47.3	strategy2
ResNeXt-50	≈35 million	**84.5**	**63**	**70.1**	**49.1**	strategy2
Wide-ResNet-50	≈77 million	83.8	62.3	69.9	48.7	strategy2

۲.۰.۴.۴ بررسی میزان مصرف حافظه‌ی GPU

مدل پیشنهادی و مدل‌های ECN [۱۱۲] ، AE [۱۹] و GPP [۱۱۳] از نظر یادگیری ویژگی‌های درون–دامنه‌ای در دامنه‌ی هدف رویکرد تقریبا مشابهی دارند. تفاوت عمده‌ی آن‌ها در استراتژی‌های انتخاب همسایه برای نمونه‌های دامنه‌ی هدف می‌باشد. مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها عملکرد بهتری از مدل‌های ECN و AE دارد. اما مدل GPP نسبت به مدل پیشنهادی، به‌ویژه در هنگام آزمایش روی مجموعه‌داده‌ی DukeMTMC-reID ، دقت بالاتری دارد. مدل GPP از ساختار گرافی برای انتخاب همسایه‌ها در دامنه‌ی هدف بهره می‌برد. ساختار گراف قادر به استخراج روابط پیچیده در داده‌ها می‌باشد. به همین دلیل عملکرد این مدل هنگام آزمایش روی مجموعه‌داده‌ی DukeMTMC-reID، که مجموعه‌داده‌ی چالشی‌تری نسبت به مجموعه‌داده‌ی Market1501 است، بهبود قابل توجهی دارد. اما مدل GPP از لحاظ پیچیدگی و میزان مصرف حافظه‌ی GPU تفاوت قابل ملاحظه‌ای با مدل پیشنهادی دارد.

در جدول ۴.۴ مدل‌های ECN ، AE ، GPP و مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها از نظر عملکرد و میزان مصرف حافظه‌ی GPU با هم مقایسه شده‌اند. از نتایج ثبت شده در جدول ۴.۴ می‌توان نتیجه گرفت که مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها با وجود عملکرد خوبی که دارد میزان مصرف حافظه‌ی آن نیز قابل قبول است. همچنین استفاده از تابع اتلاف سه‌گانه در مدل پیشنهادی میزان مصرف حافظه‌ی GPU را

به میزان اندکی افزایش داده ولی نتیجه را بهبود بخشیده است.

جدول ۴.۴: مقایسه‌ی عملکرد و میزان مصرف حافظه‌ی GPU

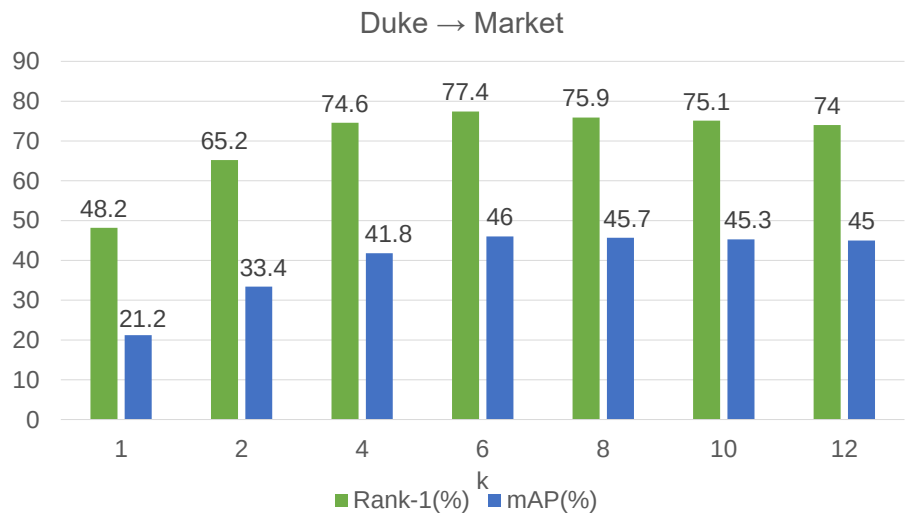
Method	GPU RAM Usage (MB)	Duke → Market (%)		Market→Duke (%)	
		R-1	mAP	R-1	mAP
ECN [112]	≈7200	75.1	43	63.3	40.4
AE [19]	≈7000	81.6	58	67.9	46.7
GPP [113]	≈9800	84.1	63.8	74	54.4
Ours ($L_{src} + L_{tgt}$)	≈7000	82.5	60.6	68.8	47.5
Ours ($L_{src} + L_{tgt} + L_{tri}$)	≈7100	84.5	63	70.1	49.1

۳.۰.۴.۴ بررسی تأثیر پارامتر کلیدی در دو استراتژی انتخاب همسایه

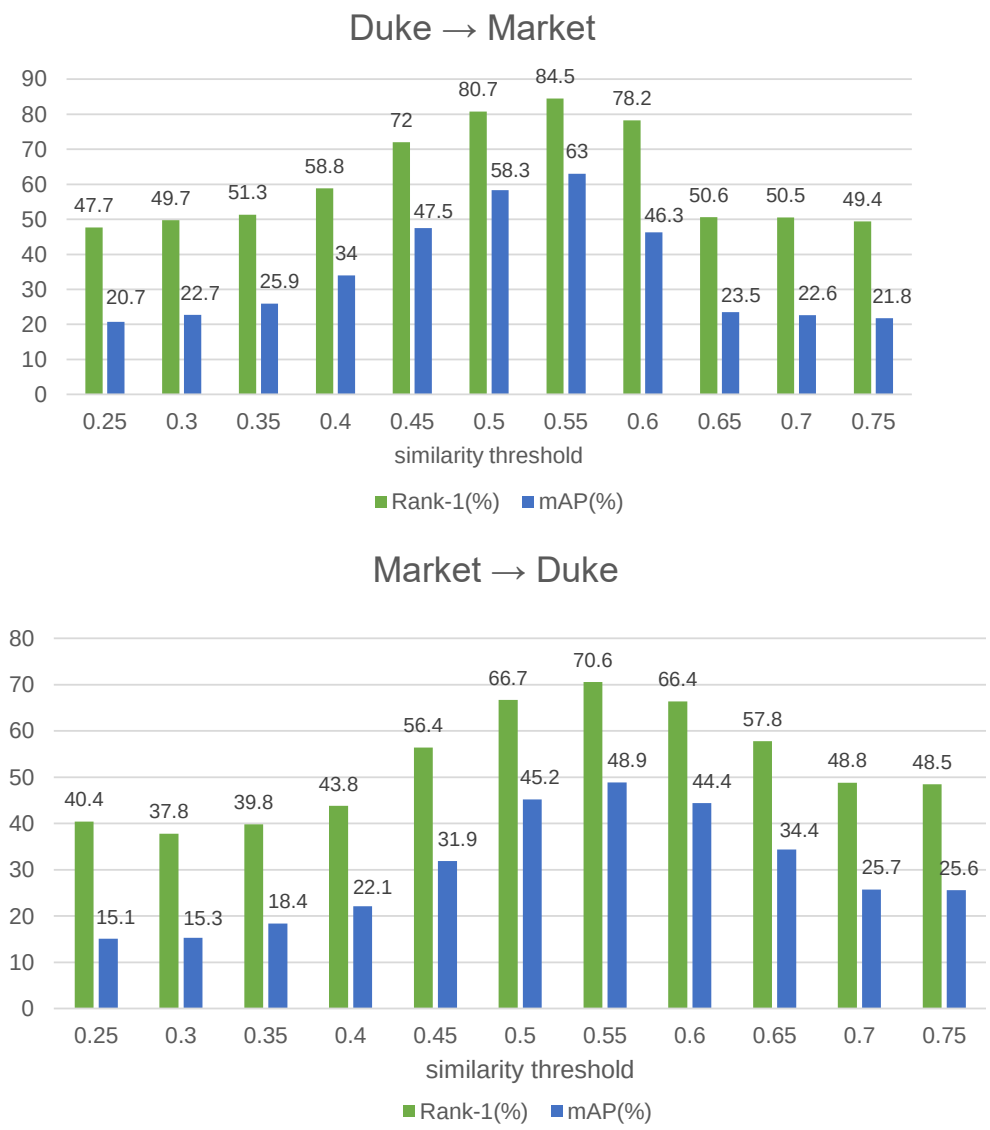
در استراتژی اول انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها، تعداد نمونه‌های مثبت همسایه‌ها (k) یک پارامتر کلیدی محسوب می‌شود. در شکل ۴.۴ نمودار نتایج اجرای آزمایش‌ها روی duke → market به ازای مقادیر مختلف تعداد نمونه‌های مثبت همسایه‌ها در استراتژی اول آورده شده است. به ازای تعداد $k = ۶$ بهترین نتایج حاصل شده است. هنگامی که $k = ۱$ باشد یعنی در یادگیری تغییرناپذیری نسبت به همسایه‌ها فقط خود تصویر درنظر گرفته شده و هیچ نمونه‌ی دیگری به‌عنوان همسایه انتخاب نمی‌شود. در این حالت در یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، یادگیری تغییرناپذیری نسبت به همسایه‌ها اعمال نشده است و همان‌طور که در نمودار شکل ۴.۴ مشخص است عملکرد مدل به شدت کاهش یافته است. حتی هنگامی‌که فقط یک نمونه به‌عنوان همسایه‌ی هر تصویر در دامنه‌ی هدف درنظرگرفته شود ($k = ۲$)، عملکرد مدل نسبت به حالتی که از یادگیری تغییرناپذیری نسبت به همسایه‌ها استفاده نمی‌شود ($k = ۱$)، بهبود قابل توجهی پیدا می‌کند. نتایج این نمودار نمایانگر اهمیت اعمال یادگیری تغییرناپذیری نسبت به همسایه‌ها در یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف می‌باشد.

در استراتژی دوم انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها، مقدار آستانه‌ی درنظر گرفته شده برای فاصله‌ی بین یک تصویر و همسایه‌هایش (μ) یک پارامتر کلیدی می‌باشد. در شکل ۵.۴ نمودار نتایج اجرای آزمایش‌ها به ازای مقادیر مختلف پارامتر μ در تنظیمات duke → market و market → duke آورده

شده است. همان‌طور که در نمودار شکل ۵.۴ مشخص است به ازای $\mu = ۰/۵۵$ بهترین نتایج حاصل شده است. هنگامی که مقدار آستانه خیلی کم و یا زیاد باشد، عملکرد مدل کاهش می‌یابد. انتخاب مقدار آستانه‌ی درنظر گرفته شده برای انتخاب همسایه‌ها در استراتژی دوم، اهمیت زیادی دارد.



شکل ۴.۴: نتایج اجرای آزمایش‌ها با مقادیر مختلف k برای یادگیری تغییرناپذیری نسبت به همسایه‌ها در استراتژی اول انتخاب همسایه‌ها روی duke → market



شکل ۵.۴: نتایج اجرای آزمایش‌ها با مقادیر مختلف μ برای یادگیری تغییرناپذیری نسبت به همسایه‌ها در استراتژی دوم انتخاب همسایه‌ها

۵.۴ مقایسه‌ی روش پیشنهادی با سایر روش‌ها

در جدول ۵.۴ عملکرد مدل پیشنهادی با ۱۴ روش بازشناسایی شخص مقایسه شده است. روش‌های CAMEL [۹۸]، PUL [۲۱] و UMDL [۶۴] روش‌های بدون نظارت هستند که در فرآیند آموزش از یک مجموعه‌داده‌ی دارای برچسب منبع برای مقداردهی اولیه مدل استفاده می‌کنند اما طی یادگیری ویژگی‌های دامنه‌ی هدف، از برچسب‌های دامنه‌ی منبع چشم‌پوشی می‌کنند. روش پیشنهادی با استراتژی اول و دوم انتخاب همسایه‌ها نسبت به روش CAMEL [۹۸] ، که از روش‌های بدون نظارت PUL [۲۱] و UMDL [۶۴] موفق‌تر است، عملکرد بسیار بهتری دارد. روش پیشنهادی ما با استراتژی دوم انتخاب همسایه‌ها دقت مرتبه ۱ را به میزان ۳۰ درصد و mAP را به میزان ۳۶/۷ درصد روی duke → market و دقت مرتبه ۱ را به میزان ۲۹/۸ درصد و mAP را به میزان ۲۹/۳ درصد روی market → duke افزایش داده است. روش‌های PTGAN [۸۹] ، SPGAN [۱۷] ، TJ-AIDL [۸۷] ، CamStyle [۱۱۵] ، HHL [۱۱۱] ، ARN [۵۰] ، ATNet [۵۱] ، DAL [۶۶] ، ECN [۱۱۲] ، AE [۱۹] و GPP [۱۱۳] ازجمله روش‌های موفق ارائه شده برای وفق‌دهی دامنه‌ی بدون نظارت در مسئله‌ی بازشناسایی شخص در سال‌های مختلف می‌باشند.

روش HHL [۱۱۱] از تابع اتلاف سه‌گانه‌ای مشابه با تابع اتلاف سه‌گانه‌ی استفاده شده در مدل پیشنهادی بهره برده است. اما مدل پیشنهادی با استراتژی‌های اول و دوم انتخاب همسایه‌ها توانسته است دقت مرتبه ۱ روی duke → market را به ترتیب ۱۵/۲ درصد و ۲۲/۳ درصد نسبت به دقت مرتبه ۱ در مدل HHL بهبود دهد. همچنین در تنظیمات duke → market معیار mAP در مدل پیشنهادی با استراتژی‌های اول و دوم انتخاب همسایه‌ها نسبت به معیار mAP در مدل HHL به ترتیب ۱۴/۶ درصد و ۳۱/۶ درصد بهبود داشته است. در تنظیمات market → duke مدل پیشنهادی با استراتژی اول و دوم انتخاب همسایه‌ها دقت مرتبه ۱ را به ترتیب ۱۷/۱ درصد و ۲۳/۲ درصد و میزان mAP را به ترتیب ۱۳/۹ درصد و ۲۱/۹ درصد بهبود داده است.

استراتژی دوم انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها، مشابه استراتژی انتخاب همسایه در مدل AE [۱۹] است. اما دقت مرتبه ۱ و معیار mAP در مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها روی duke → market نسبت به دقت مرتبه ۱ و معیار mAP در مدل AE ، ۲/۹ درصد و ۵ درصد بهبود داشته است. همچنین در تنظیمات market → duke مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها نسبت به مدل AE دقت مرتبه ۱ و معیار mAP را به ترتیب به میزان ۲/۲ درصد و ۲/۴ درصد بهبود داده است.

مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها و مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها نسبت

به مدل پایه [۱۱۲] توانسته اند در تنظیمات duke → market دقت مرتبه ۱ را به ترتیب ۲٫۳ درصد و ۹٫۴ درصد بهبود دهند. همچنین در این تنظیمات معیار mAP در مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها و مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها نسبت به mAP در مدل پایه، به ترتیب ۳ درصد و ۲۰ درصد بهبود یافته است. در تنظیمات market → duke نیز عملکرد مدل پیشنهادی با استراتژی اول و دوم انتخاب همسایه‌ها از عملکرد مدل پایه در این تنظیمات بهتر است. دقت مرتبه ۱ به ترتیب به میزان ۰٫۷ درصد و ۶٫۸ درصد افزایش یافته و معیار mAP نیز به میزان ۰٫۷ درصد و ۸٫۷ درصد بهبود یافته است.

همان‌طور که در جدول ۵.۴ قابل مشاهده است عملکرد مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها از تمامی مدل‌ها به‌غیر از AE [۱۹] بهتر است. عملکرد مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها نیز از تمامی روش‌های موجود در جدول به‌غیر از GPP [۱۱۳] بهتر می‌باشد. یکی از دلایل اصلی بهبود عملکرد در روش پیشنهادی، درنظرگرفتن ویژگی‌های درون-دامنه‌ای در دامنه‌ی هدف، طی فرآیند آموزش می‌باشد. یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف باعث می‌شود که مدل آموزش دیده نسبت به تغییرات درون-دامنه‌ای در دامنه‌ی هدف مقاوم شود و هنگام آزمایش روی دامنه‌ی هدف، عملکرد بهتری داشته باشد. علاوه‌بر تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، در روش پیشنهادی از یک تابع اتلاف سه‌گانه نیز بهره برده شده است. استفاده از تابع اتلاف سه‌گانه سبب می‌شود که علاوه‌بر تغییرات درون-دامنه‌ای در دامنه‌ی هدف، تفاوت‌های بین دامنه‌های منبع و هدف نیز یاد گرفته شود. بدین‌ترتیب مدل می‌تواند دانش یادگرفته شده از دامنه‌ی برچسب‌گذاری‌شده‌ی منبع را به دامنه‌ی بدون برچسب هدف منتقل کند. یکی دیگر از دلایل بهبود عملکرد مدل پیشنهادی، استفاده از شبکه‌ی پیش‌آموزش دیده‌ی ResNeXt به‌منظور استخراج ویژگی است. در مقاله‌ی پایه [۱۱۲] و اکثر مقاله‌های بازشناسایی شخص، از شبکه‌ی ResNet برای استخراج ویژگی استفاده شده است.

مدل‌های ECN ، AE و GPP علاوه بر تلاش برای کم کردن فاصله‌ی بین دامنه‌های منبع و هدف، سعی می‌کنند ویژگی‌های دامنه‌ی هدف را نیز طی فرآیند آموزش یاد بگیرند. این سه روش، استراتژی‌های انتخاب همسایه‌ی متفاوتی دارند. مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها از روش‌های ECN و AE روی هر دو مجموعه‌داده‌ی Market1501 و DukeMTMC-reID عملکرد بهتری دارد. مقدار معیار CMC در تنظیمات duke → market در مدل پیشنهادی از تمامی روش‌های جدول بهتر است. اما در این تنظیمات معیار mAP در مدل GPP به میزان ۰٫۸ درصد نسبت به مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها بالاتر است. همچنین در تنظیمات market → duke روش GPP عملکرد بهتری از مدل ما دارد. دلیل عملکرد بهتر مدل GPP نسبت به مدل پیشنهادی، استفاده از ساختار گرافی در استراتژی انتخاب همسایه‌ها در این مدل می‌باشد. اما از لحاظ

میزان پیچیدگی مدل و میزان مصرف حافظه‌ی GPU تفاوت قابل ملاحظه‌ای با مدل پیشنهادی ما دارد. در جدول ۴.۴ میزان حافظه‌ی مصرفی روش‌های ECN ، AE ، GPP و روش پیشنهادی آورده شده است. روش GPP نسبت به روش پیشنهادی با استراتژی دوم انتخاب همسایه‌ها مقدار ۲۷۰۰MB حافظه‌ی GPU بیشتری مصرف می‌کند. بنابراین روش پیشنهادی ما علاوه‌بر نتایج قابل رقابت میزان مصرف حافظه‌ی معقولی نیز دارد.

جدول ۵.۴: مقایسه‌ی روش پیشنهادی با سایر روش‌های وفق‌دهی دامنه‌ی بدون نظارت در بازشناسایی شخص

Method	Reference	Duke → Market (%)				Market → Duke (%)			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
UMDL [64]	CVPR 16	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
CAMEL [98]	ICCV 17	54.5	73.1	-	26.3	40.3	57.6	-	19.8
PTGAN [89]	CVPR 18	38.6	-	66.1	-	27.4	-	50.7	-
PUL [21]	TOMM 18	45.5	60.7	66.7	20.5	30	43.4	48.5	16.4
SPGAN [17]	CVPR 18	51.5	70.1	76.8	22.8	41.1	56.6	63	22.3
TJ-AIDL [87]	CVPR 18	58.2	74.8	81.1	26.5	44.3	59.6	65	23
CamStyle [115]	TIP 18	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
HHL [111]	ECCV 18	62.2	78.8	84	31.4	46.9	61	66.7	27.2
ARN [50]	CVPR 18	70.3	80.4	86.3	39.4	60.2	73.9	79.5	33.4
ATNet [51]	CVPR 19	55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
DAL [66]	ICCV 19	64.3	-	-	34.5	55.4	-	-	36.7
ECN [112] (base model)	CVPR 19	75.1	87.6	91.6	43	63.3	75.8	80.4	40.4
AE [19]	TOMM 20	81.6	91.9	94.6	58	67.9	79.2	83.6	46.7
GPP [113]	PAM 20	84.1	92.8	95.4	**63.8**	**74**	**83.7**	**87.4**	**54.4**
Ours (strategy1)	-	77.4	89.1	92.7	46	64	76.8	80.9	41.1
Ours (strategy2)	-	**84.5**	**93.1**	**95.7**	63	70.1	80.8	84.1	49.1

۶.۴ خلاصه و نتیجه‌گیری

در فرآیند آموزش و آزمایش مدل پیشنهادی، از مجموعه‌داده‌های DukeMTMC-reID و Market1501 به‌عنوان مجموعه‌داده‌های منبع و هدف استفاده شده است. تنظیمات آزمایش و مقادیر پارامترهای مختلف در اجرای آزمایشات با هردو استراتژی انتخاب همسایه‌ها، در این فصل آورده شده است. سه معماری ResNet-50، ResNeXt-50 و WideResNet-50 در مدل پیشنهادی مورد آزمایش قرار گرفته‌اند. نتایج اجرای آزمایش‌ها نشان می‌دهد که معماری ResNeXt-50 باوجود اینکه تعداد پارامترهای قابل آموزش کمتری نسبت به دو معماری دیگر دارد، عملکرد بهتری در مدل پیشنهادی دارد. همچنین با انجام آزمایش‌های متعددی، تأثیر پارامترهای کلیدی k و μ در استراتژی اول و استراتژی دوم انتخاب همسایه‌ها بررسی شد. به ازای $k = ۶$ بهترین نتایج در اجرای آزمایش‌ها با استراتژی اول انتخاب همسایه‌ها و به ازای $\mu = ۰/۵۵$ بهترین نتایج در اجرای آزمایش‌ها با استراتژی دوم انتخاب همسایه‌ها به‌دست می‌آید. در بخش آخر این فصل، نتایج اجرای مدل با چندین روش موفق دیگر در حوزه‌ی بازشناسایی شخص مقایسه شده است.

فصل ۵

خلاصه، بحث، نتیجه‌گیری و کارهای آینده

۱.۵ خلاصه

یکی از ابزارهای نظارت بر مکان‌های عمومی، دوربین‌های امنیتی نظارتی می‌باشد. این دوربین‌ها در بخش‌های مختلف یک محیط نصب می‌شوند و به‌طور پیوسته از محدوده‌ی نظارت خود، تصویر و یا ویدیو تهیه می‌کنند. به‌دلیل حجم بالای تصاویر ثبت شده توسط دوربین‌ها، امکان تحلیل و بررسی آن‌ها در هنگام نیاز، توسط اپراتور انسانی وجود ندارد. بنابراین مسئله‌ی بازشناسایی شخص اتوماتیک نقش مهمی در حفظ امنیت مکان‌های عمومی دارد. در مسئله‌ی بازشناسایی شخص، تصویر یک فرد به‌عنوان پرس‌وجو به سیستم داده می‌شود و تصاویر ثبت شده از آن شخص توسط دوربین‌های غیرهم‌پوشان، توسط سیستم بازیابی می‌شوند. مجموعه‌داده‌های حوزه‌ی بازشناسایی شخص چالش‌های زیادی دارند. تفاوت شرایط روشنایی، تفاوت زاویه‌ی دید دوربین‌ها، تفاوت حالات قرارگیری افراد، کیفیت پایین تصاویر، عدم مشخص بودن چهره‌ی افراد و... از چالش‌های این مجموعه‌های داده می‌باشند.

در سال‌های اخیر استفاده از تکنیک‌های یادگیری عمیق در بازشناسایی شخص، نتایج موفقیت‌آمیزی را در این حوزه به‌وجود آورده است. در حوزه‌ی بازشناسایی شخص بانظارت پژوهش‌های زیادی انجام شده و نتایج بسیار خوبی حاصل شده است. اما موضوع بازشناسایی شخص بدون نظارت و موضوع وفق‌دهی دامنه در بازشناسایی شخص، حوزه‌های چالشی‌تر و جدیدتری نسبت به بازشناسایی شخص بانظارت می‌باشند. در این پژوهش سعی شد که به موضوع وفق‌دهی دامنه‌ی بدون نظارت در بازشناسایی شخص پرداخته شود.

درمدل ارائه شده، داده‌های دارای برچسب دامنه‌ی منبع و داده‌های بدون برچسب دامنه‌ی هدف، به‌عنوان داده‌های آموزشی استفاده شدند. در هنگام آزمایش، مدل باید قادر باشد روی مجموعه‌داده‌ی بدون برچسب هدف عملکرد قابل قبولی داشته باشد. در روش پیشنهادی روی ویژگی‌های درون-دامنه‌ای در دامنه‌ی هدف تمرکز می‌شود. پس از استخراج ویژگی داده‌های آموزشی توسط شبکه‌ی ResNeXt-50 پیش‌آموزش‌دیده روی مجموعه‌داده‌ی ImageNet ، داده‌های دارای برچسب دامنه‌ی منبع طبقه‌بندی می‌شوند. ویژگی‌های داده‌های بدون برچسب دامنه‌ی هدف، در یک حافظه‌ی نمونه ذخیره شده و یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، روی آن‌ها اعمال می‌شود. تغییرناپذیری نسبت به نمونه‌ها، تغییرناپذیری نسبت به دوربین‌ها و تغییرناپذیری نسبت به همسایه‌ها در دامنه‌ی هدف بررسی می‌شوند. همچنین برای یادگیری تغییرناپذیری نسبت به همسایه‌ها، دو استراتژی مختلف برای انتخاب همسایه‌ها، مورد آزمایش قرار می‌گیرند. بدین‌ترتیب مدل در هنگام آزمایش روی مجموعه‌داده‌ی هدف، نسبت به تغییرات درون-دامنه‌ای در دامنه‌ی هدف مقاوم‌تر می‌شود.

در روش ارائه شده از یک تابع اتلاف سه‌گانه برای یادگیری تفاوت‌های درون-دامنه‌ای دامنه‌ی هدف و تفاوت‌های بین دامنه‌ای میان دامنه‌های منبع و هدف استفاده شده است. در این تابع سه‌گانه، یک تصویر از دامنه‌ی هدف به‌عنوان تصویر لنگر انتخاب می‌شود و یک تصویر تولید شده از آن تصویر با سبک سایر دوربین‌ها به‌عنوان نمونه‌ی مثبت و یک تصویر از دامنه‌ی منبع به‌عنوان نمونه‌ی منفی برای آن تصویر لنگر انتخاب می‌شوند. درنتیجه تابع اتلاف نهایی شبکه از مجموع تابع اتلاف طبقه‌بندی داده‌های دامنه‌ی منبع، تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف و تابع اتلاف سه‌گانه تشکیل می‌شود.

۲.۵ بحث

بازشناسایی شخص یکی از مسائل چالشی و پیچیده در حوزه‌ی بینایی ماشین است. در سال‌های اخیر با استفاده از روش‌های یادگیری عمیق نتایج موفقیت‌آمیزی در بازشناسایی شخص حاصل شده است . در این پایان‌نامه مسئله‌ی بازشناسایی شخص و به‌طور خاص وفق‌دهی دامنه‌ی بدون نظارت در مسئله‌ی بازشناسایی شخص مورد بررسی قرار گرفته است.

باوجود عملکرد موفقیت‌آمیز مدل‌های بازشناسایی شخص ، هنگام آزمایش مدل روی مجموعه‌داده‌ی بدون برچسب متفاوت با مجموعه‌داده‌ی آموزشی برچسب‌گذاری شده، عملکرد مدل به شدت کاهش پیدا می‌کند.

ازطرفی فرآیند برچسب‌گذاری تصاویر در سناریوهای واقعی بسیار پرهزینه و تقریباً غیرممکن است. بنابراین موضوع بازشناسایی شخص بدون نظارت و وفق‌دهی دامنه‌ی بدون نظارت در بازشناسایی شخص بسیار مهم می‌باشد.

در روش پیشنهادی، مدلی ارائه شده است که در فرآیند آموزش از مجموعه‌داده‌ی دارای برچسب منبع و مجموعه‌داده‌ی بدون برچسب هدف استفاده کرده و عملکرد مناسبی هنگام آزمایش روی مجموعه‌داده‌ی دامنه‌ی هدف دارد. در این روش علاوه‌بر تلاش برای کاهش فاصله‌ی دامنه‌های منبع و هدف، سعی می‌شود که با یادگیری ویژگی‌های درون-دامنه‌ای در دامنه‌ی هدف، مدل را نسبت به تغییرات در دامنه‌ی هدف مقاوم سازد و درنتیجه عملکرد مدل هنگام آزمایش روی دامنه‌ی هدف بهبود یابد. در فرآیند آموزش، سه ویژگی تغییرناپذیری نسبت به نمونه‌ها، تغییرناپذیری نسبت به دوربین‌ها و تغییرناپذیری نسبت به همسایه‌ها در دامنه‌ی هدف بررسی می‌شوند.

در یادگیری تغییرناپذیری نسبت به نمونه‌ها، هر تصویر متعلق به کلاس خودش و جدا از سایر نمونه‌ها فرض می‌شود. در یادگیری تغییرناپذیری نسبت به دوربین‌ها، فرض می‌شود که یک تصویر و نمونه‌های تولید شده از آن تصویر با سبک سایر دوربین‌ها، متعلق به یک کلاس هستند. در یادگیری تغییرناپذیری نسبت به دوربین‌ها، فرض می‌شود که یک تصویر و همسایه‌های آن متعلق به کلاس واحدی می‌باشند. شیوه‌ی انتخاب همسایه‌ها در یادگیری تغییرناپذیری نسبت به همسایه‌ها در عملکرد مدل تاثیرگذار است. در استراتژی اول انتخاب همسایه‌ها تعداد k نزدیک‌ترین نمونه به تصویر، به‌عنوان همسایه‌های آن تصویر درنظر گرفته می‌شوند. در استراتژی دوم انتخاب همسایه‌ها، نمونه‌هایی که فاصله‌ی آن‌ها از تصویر از یک مقدار آستانه‌ای کمتر باشد به‌عنوان همسایه‌های آن تصویر انتخاب می‌شوند. اجرای مدل با استراتژی دوم انتخاب همسایه‌ها نسبت به اجرای مدل با استراتژی اول انتخاب همسایه‌ها عملکرد بهتری دارد.

در مدل پیشنهادی علاوه‌بر تابع اتلاف طبقه‌بندی داده‌های منبع و تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، یک تابع اتلاف سه‌گانه نیز استفاده شده است.در این تابع سه‌گانه علاوه‌بر درنظر گرفتن تغییرات درون-دامنه‌ای در دامنه‌ی هدف، تفاوت‌های بین دامنه‌های منبع و هدف نیز بررسی می‌شود. تابع اتلاف سه‌گانه موجب بهبود عملکرد مدل پیشنهادی نسبت به مدل پایه [۱۱۲] شده است.

از نقاط ضعف مدل پیشنهادی می‌توان به پیچیدگی و تعداد زیاد پارامترهای آن اشاره نمود. مدل پیشنهادی باوجود عملکرد خوبی که دارد، از لحاظ تعداد پارامترهای قابل آموزش، مدلی نسبتاً پیچیده محسوب می‌شود. استفاده از تکنیک‌هایی مثل کاهش ابعاد لایه‌های اضافه‌شده و ... به‌منظور کاهش پیچیدگی مدل، به‌نحوی‌که عملکرد مدل کاهش نیابد کار ارزشمندی خواهد بود.

۳.۵ نتیجه‌گیری

در روش پیشنهادی سعی شده است که مدلی با تعمیم‌پذیری بالا روی دامنه‌ی هدف بدون برچسب ارائه شود. درواقع مدلی برای وفق‌دهی دامنه‌ی بدون نظارت در حوزه‌ی بازشناسایی شخص ارائه شده است. در این مدل، داده‌های دارای برچسب دامنه‌ی منبع و داده‌های بدون برچسب دامنه‌ی هدف، باهم برای آموزش مدل استفاده می‌شوند و مدل در هنگام آزمایش روی داده‌های بدون برچسب دامنه‌ی هدف، عملکرد بسیار خوبی دارد. به‌منظور استخراج ویژگی داده‌های دامنه‌های منبع و هدف، از شبکه‌ی پیش‌آموزش دیده‌ی ResNeXt-50 استفاده شده است که نسبت به شبکه‌ی ResNet-50 تعداد پارامترهای قابل آموزش کمتر و عملکرد بهتری دارد.

در مدل پیشنهادی، از تابع اتلاف طبقه‌بندی داده‌های منبع، تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف و تابع اتلاف سه‌گانه استفاده شده است. در تابع اتلاف یادگیری تغییرناپذیری‌ها در دامنه‌ی هدف، برای نحوه‌ی انتخاب همسایه‌ها، دو استراتژی مورد آزمایش قرار گرفته‌اند. مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها توانسته روی duke $\rightarrow$ market در رتبه‌ی ۱معیار CMC مقدار ۷۷/۴ درصد و مقدار mAP ۴۶ درصد را به‌دست آورد. همچنین مدل پیشنهادی با استراتژی اول انتخاب همسایه‌ها، در تنظیمات market $\rightarrow$ duke در رتبه‌ی ۱معیار CMC مقدار ۶۴ درصد و مقدار mAP ۴۱/۱ درصد را به‌دست آورده است.

مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها توانسته روی duke $\rightarrow$ market در رتبه‌ی ۱ معیار CMC مقدار ۸۴/۵ درصد و مقدار mAP ۶۳ درصد را به‌دست آورد. همچنین مدل پیشنهادی با استراتژی دوم انتخاب همسایه‌ها، در تنظیمات market $\rightarrow$ duke در رتبه‌ی ۱معیار CMC مقدار ۷۰/۱ درصد و مقدار mAP ۴۹/۱ درصد را به‌دست آورده است. این مقادیر به‌دست آمده نتایج بسیار خوبی در این حوزه به‌شمار می‌روند.

۴.۵ کارهای آینده

مسئله‌ی بازشناسایی شخص یکی از مسائل جدید و پیچیده در حوزه‌ی بینایی ماشین است. باوجود موفقیت‌های زیادی که در این حوزه حاصل شده است هنوز مسئله‌ی بازشناسایی شخص چالش‌های زیادی دارد. حوزه‌ی وفق‌دهی دامنه‌ی بدون نظارت یکی از شاخه‌های بسیار مهم در بازشناسایی شخص می‌باشد. در این بخش ایده‌هایی که می‌توانند موجب بهبود عملکرد مدل پیشنهادی شوند، بیان می‌شوند.

- **روش‌های داده‌افزایی با استفاده از شبکه‌ی مولدتخاصمی برای دامنه‌ی هدف**

در مدل ارائه شده، در دامنه‌ی هدف، به منظور یادگیری تغییرناپذیری نسبت به دوربین‌ها، از داده‌افزایی با استفاده از شبکه‌های مولد تخاصمی استفاده شده است. می‌توان از روش‌های دیگر داده‌افزایی با استفاده از شبکه‌ی مولد تخاصمی، مثل تولید تصاویر جدیدی از افراد با حالات قرارگیری مختلف و ... نیز بهره برد.

- **مجموعه‌داده‌های متنوع**

می‌توان مدل ارائه شده را روی مجموعه‌داده‌های بیشتری از حوزه‌ی بازشناسایی شخص آزمایش نمود و به‌منظور بهبود عملکرد مدل، از ترکیب مجموعه‌های داده برای آموزش مدل استفاده کرد.

- **استفاده از معماری‌های جدیدتر CNN**

می‌توان از معماری‌های جدیدتر CNN مانند EfficientNet به منظور استخراج ویژگی‌های دقیق‌تر استفاده نمود.

- **کاهش پیچیدگی مدل**

مدل‌های عمیق بازشناسایی شخص معمولاً پیچیدگی زیادی دارند. مدل ارائه شده در این پژوهش نیز دارای لایه‌ها و انشعابات متعددی است. بنابراین تعداد پارامترهای مدل زیاد است و آموزش مدل نیاز به تجهیزات سخت‌افزاری و زمان زیادی دارد. کاهش پیچیدگی مدل به‌نحوی‌که عملکرد آن کاهش پیدا نکند، کار ارزشمندی خواهد بود.

مراجع

[1] Ackley, David H, Hinton, Geoffrey E, and Sejnowski, Terrence J. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[2] Ballard, Dana H. Modular learning in neural networks. In *AAAI*, pages 279–284, 1987.

[3] Bazzani, Loris, Cristani, Marco, Perina, Alessandro, Farenzena, Michela, and Murino, Vittorio. Multiple-shot person re-identification by hpe signature. In *2010 20th International Conference on Pattern Recognition*, pages 1413–1416. IEEE, 2010.

[4] Bengio, Yoshua. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[5] Bromley, Jane, Guyon, Isabelle, LeCun, Yann, Säckinger, Eduard, and Shah, Roopak. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[6] Carreira-Perpinan, Miguel A and Hinton, Geoffrey E. On contrastive divergence learning. In *Aistats*, volume 10, pages 33–40. Citeseer, 2005.

[7] Chen, Dapeng, Xu, Dan, Li, Hongsheng, Sebe, Nicu, and Wang, Xiaogang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2018.

[8] Chen, Guangyi, Lin, Chunze, Ren, Liangliang, Lu, Jiwen, and Zhou, Jie. Self-critical attention learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9637–9646, 2019.

[9] Chen, Weihua, Chen, Xiaotang, Zhang, Jianguo, and Huang, Kaiqi. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017.

[10] Cheng, De, Gong, Yihong, Zhou, Sanping, Wang, Jinjun, and Zheng, Nanning. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[11] Cheng, Dong Seon, Cristani, Marco, Stoppa, Michele, Bazzani, Loris, and Murino, Vittorio. Custom pictorial structures for re-identification. In *Bmvc*, volume 1, page 6. Citeseer, 2011.

[12] Cho, KyungHyun, Raiko, Tapani, Ilin, Alexander, and Karhunen, Juha. A two-stage pre-training algorithm for deep boltzmann machines. In *International Conference on Artificial Neural Networks*, pages 106–113. Springer, 2013.

[13] Choi, Yunjey, Choi, Minje, Kim, Munyoung, Ha, Jung-Woo, Kim, Sunghun, and Choo, Jaegul. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[14] Chung, Dahjung, Tahboub, Khalid, and Delp, Edward J. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1983–1991, 2017.

[15] Das, Abir, Chakraborty, Anirban, and Roy-Chowdhury, Amit K. Consistent re-identification in a camera network. In *European conference on computer vision*, pages 330–345. Springer, 2014.

[16] Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[17] Deng, Weijian, Zheng, Liang, Ye, Qixiang, Kang, Guoliang, Yang, Yi, and Jiao, Jianbin. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2018.

[18] Dikmen, Mert, Akbas, Emre, Huang, Thomas S, and Ahuja, Narendra. Pedestrian recognition with a learned metric. In *Asian conference on Computer vision*, pages 501–512. Springer, 2010.

[19] Ding, Yuhang, Fan, Hehe, Xu, Mingliang, and Yang, Yi. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1):1–19, 2020.

[20] Dong, Xuanyi, Yan, Yan, Ouyang, Wanli, and Yang, Yi. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018.

[21] Fan, Hehe, Zheng, Liang, Yan, Chenggang, and Yang, Yi. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18, 2018.

[22] Farenzena, Michela, Bazzani, Loris, Perina, Alessandro, Murino, Vittorio, and Cristani, Marco. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2360–2367. IEEE, 2010.

[23] Ge, Yixiao, Li, Zhuowan, Zhao, Haiyu, Yin, Guojun, Yi, Shuai, Wang, Xiaogang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in neural information processing systems*, pages 1222–1233, 2018.

[24] Gheissari, Niloofar, Sebastian, Thomas B, and Hartley, Richard. Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1528–1535. IEEE, 2006.

[25] Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[26] Gray, Douglas and Tao, Hai. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[27] Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, and Courville, Aaron C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[28] Guo, Yiluan and Cheung, Ngai-Man. Efficient and deep person re-identification using multi-level similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2335–2344, 2018.

[29] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[30] Hebb, Donald Olding. *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall, 1949.

[31] Hermans, Alexander, Beyer, Lucas, and Leibe, Bastian. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[32] Hinton, Geoffrey E, Osindero, Simon, and Teh, Yee-Whye. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[33] Hinton, Geoffrey E and Salakhutdinov, Ruslan R. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[34] Hinton, Geoffrey E and Salakhutdinov, Russ R. A better way to pretrain deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2012.

[35] Hirzer, Martin, Beleznai, Csaba, Roth, Peter M, and Bischof, Horst. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[36] Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[37] Huang, Timothy and Russell, Stuart. Object identification in a bayesian context. In *IJCAI*, volume 97, pages 1276–1282, 1997.

[38] Huang, Yan, Xu, Jingsong, Wu, Qiang, Zheng, Zhedong, Zhang, Zhaoxiang, and Zhang, Jian. Multi-pseudo regularized label for generated data in person re-identification. *IEEE Transactions on Image Processing*, 28(3):1391–1403, 2018.

[39] Hubel, David H and Wiesel, Torsten N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[40] Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[41] Jordan, Michael I. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier, 1997.

[42] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[43] Lavi, Bahram, Serj, Mehdi Fatan, and Ullah, Ihsan. Survey on deep learning techniques for person re-identification task. *arXiv preprint arXiv:1807.05284*, 2018.

[44] LeCun, Yann, Boser, Bernhard E, Denker, John S, Henderson, Donnie, Howard, Richard E, Hubbard, Wayne E, and Jackel, Lawrence D. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[45] LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[46] Leng, Qingming, Ye, Mang, and Tian, Qi. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2019.

[47] Li, Wei and Wang, Xiaogang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.

[48] Li, Wei, Zhao, Rui, and Wang, Xiaogang. Human reidentification with transferred metric learning. In *Asian conference on computer vision*, pages 31–44. Springer, 2012.

[49] Li, Wei, Zhao, Rui, Xiao, Tong, and Wang, Xiaogang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[50] Li, Yu-Jhe, Yang, Fu-En, Liu, Yen-Cheng, Yeh, Yu-Ying, Du, Xiaofei, and Frank Wang, Yu-Chiang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–178, 2018.

[51] Liu, Jiawei, Zha, Zheng-Jun, Chen, Di, Hong, Richang, and Wang, Meng. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7202–7211, 2019.

[52] Liu, Jinxian, Ni, Bingbing, Yan, Yichao, Zhou, Peng, Cheng, Shuo, and Hu, Jianguo. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[53] Liu, Ming-Yu and Tuzel, Oncel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[54] Loy, Chen Change, Liu, Chunxiao, and Gong, Shaogang. Person re-identification by manifold ranking. In *2013 IEEE International Conference on Image Processing*, pages 3567–3571. IEEE, 2013.

[55] Luo, Hao, Gu, Youzhi, Liao, Xingyu, Lai, Shenqi, and Jiang, Wei. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[56] Lv, Jianming, Chen, Weihang, Li, Qing, and Yang, Can. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7948–7956, 2018.

[57] Lv, Jianming and Wang, Xintong. Cross-dataset person re-identification using similarity preserved generative adversarial networks. In *International Conference on Knowledge Science, Engineering and Management*, pages 171–183. Springer, 2018.

[58] Ma, Liqian, Jia, Xu, Sun, Qianru, Schiele, Bernt, Tuytelaars, Tinne, and Van Gool, Luc. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017.

[59] Martinel, Niki and Micheloni, Christian. Re-identify people in wide area camera network. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 31–36. IEEE, 2012.

[60] McCulloch, Warren S and Pitts, Walter. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[61] McLaughlin, Niall, Del Rincon, Jesus Martinez, and Miller, Paul. Data-augmentation for reducing dataset bias in person re-identification. In *2015 12th IEEE International conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2015.

[62] Montavon, Grégoire and Müller, Klaus-Robert. Deep boltzmann machines and the centering trick. In *Neural Networks: Tricks of the Trade*, pages 621–637. Springer, 2012.

[63] Ni, Xingyang, Fang, Liang, and Huttunen, Heikki. Adaptivereid: Adaptive l2 regularization in person re-identification. *arXiv preprint arXiv:2007.07875*, 2020.

[64] Peng, Peixi, Xiang, Tao, Wang, Yaowei, Pontil, Massimiliano, Gong, Shaogang, Huang, Tiejun, and Tian, Yonghong. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1315, 2016.

[65] Perez, Luis and Wang, Jason. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[66] Qi, Lei, Wang, Lei, Huo, Jing, Zhou, Luping, Shi, Yinghuan, and Gao, Yang. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8080–8089, 2019.

[67] Quan, Ruijie, Dong, Xuanyi, Wu, Yu, Zhu, Linchao, and Yang, Yi. Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3750–3759, 2019.

[68] Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[69] Rosenblatt, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[70] Salakhutdinov, Ruslan and Hinton, Geoffrey. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455, 2009.

[71] Salakhutdinov, Ruslan and Larochelle, Hugo. Efficient learning of deep boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 693–700, 2010.

[72] Schroff, Florian, Kalenichenko, Dmitry, and Philbin, James. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[73] Schwartz, William Robson and Davis, Larry S. Learning discriminative appearance-based models using partial least squares. In *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*, pages 322–329. IEEE, 2009.

[74] Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[75] Smolensky, Paul. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

[76] Song, Jifei, Yang, Yongxin, Song, Yi-Zhe, Xiang, Tao, and Hospedales, Timothy M. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 719–728, 2019.

[77] Song, Liangchen, Wang, Cheng, Zhang, Lefei, Du, Bo, Zhang, Qian, Huang, Chang, and Wang, Xinggang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, page 107173, 2020.

[78] Sovrasov, Vladislav and Sidnev, Dmitry. Building computationally efficient and well-generalizing person re-identification models with metric learning. *arXiv preprint arXiv:2003.07618*, 2020.

[79] Sun, Yifan, Zheng, Liang, Yang, Yi, Tian, Qi, and Wang, Shengjin. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[80] Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[81] Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[82] Tan, Mingxing and Le, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[83] Varior, Rahul Rama, Haloi, Mrinal, and Wang, Gang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016.

[84] Varior, Rahul Rama, Shuai, Bing, Lu, Jiwen, Xu, Dong, and Wang, Gang. A siamese long short-term memory architecture for human re-identification. In *European conference on computer vision*, pages 135–153. Springer, 2016.

[85] Voulodimos, Athanasios, Doulamis, Nikolaos, Doulamis, Anastasios, and Protopapadakis, Eftychios. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[86] Wang, Guangcong, Lai, Jianhuang, Huang, Peigen, and Xie, Xiaohua. Spatial-temporal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8933–8940, 2019.

[87] Wang, Jingya, Zhu, Xiatian, Gong, Shaogang, and Li, Wei. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018.

[88] Wang, Taiqing, Gong, Shaogang, Zhu, Xiatian, and Wang, Shengjin. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016.

[89] Wei, Longhui, Zhang, Shiliang, Gao, Wen, and Tian, Qi. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[90] Werbos, Paul. Beyond regression:" new tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*, 1974.

[91] Xie, Saining, Girshick, Ross, Dollár, Piotr, Tu, Zhuowen, and He, Kaiming. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[92] Xu, Yuanlu, Ma, Bingpeng, Huang, Rui, and Lin, Liang. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 937–940, 2014.

[93] Yang, Qize, Yu, Hong-Xing, Wu, Ancong, and Zheng, Wei-Shi. Patch-based discriminative feature learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3633–3642, 2019.

[94] Ye, Mang, Shen, Jianbing, Lin, Gaojie, Xiang, Tao, Shao, Ling, and Hoi, Steven CH. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020.

[95] Yi, Dong, Lei, Zhen, Liao, Shengcai, and Li, Stan Z. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014.

[96] Yi, Zili, Zhang, Hao, Tan, Ping, and Gong, Minglun. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[97] Younes, Laurent. On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics: An International Journal of Probability and Stochastic Processes*, 65(3-4):177–228, 1999.

[98] Yu, Hong-Xing, Wu, Ancong, and Zheng, Wei-Shi. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 994–1002, 2017.

[99] Zagoruyko, Sergey and Komodakis, Nikos. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[100] Zajdel, Wojciech, Zivkovic, Zoran, and Krose, Ben JA. Keeping track of humans: Have i seen this person before? In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 2081–2086. IEEE, 2005.

[101] Zhang, Xinyu, Cao, Jiewei, Shen, Chunhua, and You, Mingyu. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8222–8231, 2019.

[102] Zheng, Liang, Bie, Zhi, Sun, Yifan, Wang, Jingdong, Su, Chi, Wang, Shengjin, and Tian, Qi. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[103] Zheng, Liang, Shen, Liyue, Tian, Lu, Wang, Shengjin, Wang, Jingdong, and Tian, Qi. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[104] Zheng, Meng, Karanam, Srikrishna, Wu, Ziyan, and Radke, Richard J. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5735–5744, 2019.

[105] Zheng, Wei-Shi, Gong, Shaogang, and Xiang, Tao. Person re-identification by probabilistic relative distance comparison. In *CVPR 2011*, pages 649–656. IEEE, 2011.

[106] Zheng, Wei-Shi, Gong, Shaogang, and Xiang, Tao. Transfer re-identification: From person to set-based verification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2657. IEEE, 2012.

[107] 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 Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

[108] Zheng, Zhedong, Zheng, Liang, and Yang, Yi. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):1–20, 2017.

[109] Zheng, Zhedong, Zheng, Liang, and Yang, Yi. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[110] Zhong, Zhun, Zheng, Liang, Kang, Guoliang, Li, Shaozi, and Yang, Yi. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[111] Zhong, Zhun, Zheng, Liang, Li, Shaozi, and Yang, Yi. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188, 2018.

[112] Zhong, Zhun, Zheng, Liang, Luo, Zhiming, Li, Shaozi, and Yang, Yi. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019.

[113] Zhong, Zhun, Zheng, Liang, Luo, Zhiming, Li, Shaozi, and Yang, Yi. Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[114] Zhong, Zhun, Zheng, Liang, Zheng, Zhedong, Li, Shaozi, and Yang, Yi. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[115] Zhong, Zhun, Zheng, Liang, Zheng, Zhedong, Li, Shaozi, and Yang, Yi. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2018.

[116] Zhou, Kaiyang, Yang, Yongxin, Cavallaro, Andrea, and Xiang, Tao. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019.

[117] Zhou, Sanping, Wang, Fei, Huang, Zeyi, and Wang, Jinjun. Discriminative feature learning with consistent attention regularization for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8040–8049, 2019.

[118] Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, and Efros, Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Abstract

Person reidentification problem is a challenging computer vision task in which there is a query image captured by one camera and the goal is to retrieve other images of that person from the gallery set captured by multiple non-overlapping cameras. Images in person reidentification datasets are captured by surveillance cameras in public places and have different challenges. In recent years using deep learning approach, good results have been achieved in the field of person reidentification. Despite these achievements, the performance of the model usually decreases during testing the model on different unlabeled datasets. Unsupervised domain adaptation is a solution to this problem. The current thesis proposes a well-generalized model for unsupervised domain adaptation in person reidentification. The model uses both labeled source dataset and unlabeled target dataset during training and has a good performance during testing on an unlabeled target domain. ResNeXt-50 network is used for feature extraction. It has fewer trainable parameters than ResNet-50 network. To make the model robust to the target domain's variations, 3 invariance properties are considered during training over the target dataset. Exemplar invariance, camera invariance, and neighborhood invariance are making the invariance learning loss in the target domain. Two strategies are examined for selecting neighbors in neighborhood invariance learning. The final loss function of the network is consists of classification loss for labeled source domain, invariance learning loss for the target domain, and a triplet loss. The triplet loss considers the intra-domain variations in the target domain and the inter-domain variations between source and target domains. The proposed model with strategy 1 for selecting neighbors in neighborhood invariance achieves 77.4 % in rank1 accuracy , 89.1% in rank5 accuracy, and 46% for mAP on duke $\rightarrow$ market setting. It also achieves 64 % in rank1 accuracy, 76.8% in rank5 accuracy, and 41.1% for mAP on market $\rightarrow$ duke setting. The proposed model with strategy 2 for selecting neighbors in neighborhood invariance achieves 84.5 % in rank1 accuracy, 93.1% in rank5 accuracy, and 63% for mAP. It also achieves 70.1 % in rank1 accuracy, 80.8% in rank5 accuracy, and 49.1% for mAP on market $\rightarrow$ duke setting.

Keywords Person Reidentification, Deep Learning, Domain Adaptation, Convolutional Neural Network, Generalizable Model

University of Tehran
College of Farabi
Faculty of Engineering
Department of Computer Engineering

Proposing a Generalizable Model for Person Re-identification Using Deep Learning Approach

A Thesis Submitted to the Graduate Studies Office
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
in Information Technology Engineering - Information Technology

By:
Saba Sadat Faghih Imani

Supervisor:
Dr. Kazim Fouladi

Advisor:
Dr. Hossein Aghababa

September 2020