Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال چهارم، شماره دوم، پاییز و زمستان ۱۴۰۵، صفحات ۶۰ - ۴۹

تشخیص خویشاوندی چهره با استفاده از ترکیب یادگیری خودنظارتی و ترنسفورمر بصری

نسرین رسولی[۱]، کمال میرزایی[۲*]، محسن سرداری زارچی[۳]

[۱] دانشجوی دکتری، گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران

[۲] استادیار، گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران

[۳] دانشیار، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه میبد، میبد، ایران

مقاله پژوهشی

چکیده

تشخیص روابط خویشاوندی از روی چهره به دلیل تغییرات ناشی از سن، ژست، نور و شرایط غیرکنترل‌شده، چالشی اساسی در بینایی ماشین است. وابستگی روش‌های موجود به داده‌های برچسب‌خورده محدود، تعمیم‌پذیری آن‌ها را کاهش‌داده‌است. در این مقاله، یک چارچوب یادگیری عمیق چندمرحله‌ای برای یادگیری شباهت‌های موروثی چهره ارائه‌می‌شود که ترکیبی از یادگیری خودنظارتی و مدل‌سازی رابطه‌ای مبتنی بر ترنسفورمرهای بصری است. در مرحله اول، بازنمایی‌های مقاوم چهره با آموزش خودنظارتی و بدون برچسب خویشاوندی استخراج‌می‌شوند. سپس در مرحله نظارت‌شده، این بازنمایی‌ها با سازوکار توجه بین‌فردی پالایش‌شده و شباهت‌های ژنتیکی ظریف تقویت‌می‌گردند. روش پیشنهادی بر روی مجموعه داده‌های مرجع ارزیابی‌شده است. نتایج تجربی نشان‌می‌دهد روش پیشنهادی به‌دقت ۹۴/۲٪ در مجموعه‌داده کنترل‌شده **KinFaceW-I** و دقت ۸۷/۶٪ در مجموعه غیرکنترل‌شده **KinFaceW-II/FIW** می‌رسد که نسبت به **ViT** بدون خودنظارتی به‌ترتیب بهبود ۴/۸٪ و ۶/۴٪ نشان‌میدهد. افت عملکرد در انتقال از محیط کنترل‌شده به غیرکنترل‌شده تنها ۶/۶٪ است که کمتر از روش‌های مقایسه‌ای (۸/۲٪ تا ۱۱/۲٪) است. ترکیب یادگیری خودنظارتی با مدل‌سازی توجه‌محور، وابستگی به داده‌های برچسب‌خورده را کاهش و کارایی را در محیط‌های واقعی افزایش‌می‌دهد.

تاریخ دریافت:
۱۴۰۴/۱۲/۰۲

تاریخ پذیرش:
۱۴۰۵/۰۴/۰۱

کلیدواژه‌ها:
تشخیص خویشاوندی، یادگیری خود نظارتی، بینایی ماشین، ترنسفورمرهای بصری، یادگیری شباهت

نویسنده‌مسئول:
kamal.mirzaie@iau.ac.ir

doi : 10.22034/ABMIR.2026.24353.1227

E-ISSN: 2821-2037

۱- مقدمه

تشخیص روابط خویشاوندی[1] بر اساس اطلاعات چهره، یکی از مسائل داغ و درعین‌حال چالشی در حوزه بینایی ماشین و هوش مصنوعی به‌شمارمی‌آید. برخلاف مسئله شناسایی هویت، که هدف آن تطبیق دقیق یک فرد با نمونه‌های قبلی است، در تشخیص خویشاوندی تمرکز بر کشف شباهت‌های ظریف، پراکنده و اغلب غیرمستقیمی است که ریشه در ویژگی‌های ژنتیکی و ساختارهای موروثی چهره دارند. این شباهت‌ها ممکن است در اجزای مختلف صورت نظیر فرم استخوان‌بندی، تناسب اجزا یا الگوهای بافتی ظاهرشوند و به‌صورت یکنواخت یا آشکار قابل‌مشاهده نباشند [۱].

تشخیص خودکار روابط خویشاوندی از روی تصاویر چهره، یکی از مسائل مهم و رو به رشد در حوزه بینایی ماشین با کاربردهای گسترده اجتماعی و انسانی است. این فناوری می‌تواند در یافتن کودکان گمشده، بازسازی شجره‌نامه‌های خانوادگی، تأیید هویت در اسناد رسمی، تحلیل داده‌های مردم‌شناسی و حتی جنایی‌شناسی نقش کلیدی ایفا کند. برخلاف روش‌های سنتی که به آزمایش‌های گران و زمان‌بر DNA متکی هستند، تشخیص خویشاوندی مبتنی بر چهره یک راهکار سریع، غیرتهاجمی و کم‌هزینه ارائه می‌دهد. با گسترش شبکه‌های اجتماعی و پایگاه‌های داده تصویری عظیم، نیاز به سامانه‌های خودکاری که بتوانند روابط خانوادگی را تنها بر اساس تصویر چهره تشخیص دهند، بیش‌ازپیش احساس می‌شود. بااین‌حال، به دلیل تغییرات ظاهری ناشی از سن، ژست، نورپردازی و شرایط غیرایده‌ال تصویربرداری، این مسئله همچنان یکی از چالش‌های باز در حوزه بینایی ماشین محسوب می‌شود و پژوهش در این زمینه از اهمیت بالایی برخورداراست [۲]. در چنین شرایطی، روش‌های کلاسیک مبتنی بر ویژگی‌های دست‌ساز[2] و حتی بسیاری از مدل‌های یادگیری عمیق نظارت‌شده، با افت عملکرد و ضعف در تعمیم‌پذیری مواجه می‌شوند. یکی از دلایل اصلی این مسئله، وابستگی شدید این روش‌ها به مجموعه‌داده‌های برچسب‌خورده محدود و بعضاً سوگیرانه است که توان پوشش تمامی حالات ممکن در دنیای واقعی را ندارند [۳].

در سال‌های اخیر، به‌کارگیری شبکه‌های عمیق و چارچوب‌های یادگیری جفتی نظیر شبکه‌های سیامی[3] و سه‌تایی[4]، پیشرفت‌های قابل‌توجهی در تشخیص خویشاوندی ایجاد کرده است. بااین‌حال، این رویکردها اغلب به داده‌های آموزشی گسترده و برچسب‌خورده متکی هستند و در مواجهه با تغییر دامنه داده یا انتقال به مجموعه‌داده‌های جدید، پایداری مطلوبی از خود نشان نمی‌دهند. این محدودیت‌ها، پژوهشگران را به سمت بهره‌گیری از الگوهای یادگیری کم‌نظارت و بدون‌نظارت سوق داده است [٤].

یادگیری خودنظارتی[5] به‌عنوان یکی از رویکردهای نوظهور در یادگیری ماشین، امکان استخراج بازنمایی‌های معنادار از داده‌های خام و بدون برچسب را فراهم می‌سازد. این رویکرد با تعریف وظایف پیش‌متنی و استفاده از حجم عظیمی از داده‌های بدون‌نظارت، قادر است ویژگی‌هایی مقاوم در برابر تغییرات ظاهری و نویزهای محیطی بیاموزد. چنین ویژگی‌هایی می‌توانند پایه‌ای مناسب برای مسائل سطح بالا مانند تشخیص روابط خویشاوندی فراهم کنند، به‌ویژه در شرایطی که داده‌های برچسب‌خورده کمیاب یا پرهزینه هستند [۵].

از سوی دیگر، معماری‌های مبتنی بر ترنسفورمرهای بصری[6] با اتکا بر سازوکار توجه[7]، توانایی مدل‌سازی وابستگی‌های بلندبرد و روابط ساختاری میان نواحی مختلف تصویر را دارند. این ویژگی، آن‌ها را به گزینه‌ای مناسب برای تحلیل چهره و استخراج الگوهای پراکنده و غیرمحلی تبدیل کرده است. ترکیب این معماری‌ها با سازوکارهای توجه بین‌فردی، امکان یادگیری رابطه‌ای میان دو تصویر چهره را فراهم می‌کند؛ به‌گونه‌ای‌که شباهت‌ها و تفاوت‌های معنادار میان آن‌ها به‌صورت صریح مدل‌سازی شوند [٦].

در این پژوهش، با تکیه‌بر این مشاهدات، یک چارچوب جدید برای یادگیری شباهت‌های موروثی چهره ارائه‌می‌شود که از آموزش

[1] Kinship verification
[2] Hand-crafted features
[3] Siamese
[4] Triplet
[5] Self-supervised learning
[6] Vision Transformers
[7] Attention

خودنظارتی برای استخراج بازنمایی‌های عمومی[1] و از مدل‌سازی توجه‌محور برای تقویت روابط بین‌چهره‌ای بهره می‌برد. هدف اصلی، کاهش وابستگی به داده‌های برچسب‌خورده، افزایش تعمیم‌پذیری مدل در محیط‌های غیرکنترل‌شده و بهبود دقت تشخیص روابط خویشاوندی است. در ادامه، جزئیات روش پیشنهادی، داده‌های مورداستفاده و نتایج تجربی به‌صورت نظام‌مند ارائه و تحلیل خواهند شد. نوآوری اصلی این پژوهش نه در معرفی یک معماری کاملاً جدید، بلکه در طراحی یک چارچوب یکپارچه و هدفمند برای مسئله تشخیص خویشاوندی نهفته است؛ چارچوبی که در آن، بازنمایی‌های عمومی آموخته‌شده به‌صورت خودنظارتی، به‌طور خاص برای مدل‌سازی روابط موروثی پالایش می‌شوند. این رویکرد، شکاف میان یادگیری بازنمایی عمومی و استدلال رابطه‌محور در حوزه تشخیص خویشاوندی را تا حد قابل‌توجهی کاهش می‌دهد. نوآوری اصلی این را می‌توان در موارد زیر خلاصه کرد:

۱. ارائه یک چارچوب دومرحله‌ای که برای اولین بار یادگیری خودنظارتی را با ترنسفورمر بصری و مکانیسم توجه بین‌فردی برای مسئله تشخیص خویشاوندی چهره ترکیب می‌کند.

۲. استفاده از آموزش خودنظارتی بر روی داده‌های بدون برچسب برای استخراج بازنمایی‌های عمومی و مقاوم چهره، به‌طوری‌که وابستگی به داده‌های برچسب‌خورده خویشاوندی به میزان قابل‌توجهی کاهش می‌یابد.

۳. طراحی پودمان توجه بین‌فردی که به‌طور خاص شباهت‌های ظریف و پراکنده ژنتیکی میان جفت چهره‌ها را مدل‌سازی می‌کند؛ برخلاف روش‌های پیشین که صرفاً از شباهت کسینوسی یا فاصله اقلیدسی استفاده می‌کردند.

٤. کاهش شکاف میان یادگیری بازنمایی عمومی و استدلال رابطه‌محور در مسئله تشخیص خویشاوندی، که در پژوهش‌های قبلی کمتر موردتوجه قرارگرفته است.

ساختار باقی‌مانده این مقاله به‌صورت زیر سازمان‌دهی شده است: در بخش ۲، مرور جامعی بر پژوهش‌های پیشین در زمینه تشخیص خویشاوندی چهره، شامل روش‌های مبتنی بر ویژگی‌های دست‌ساز، شبکه‌های عصبی کانولوشنی، یادگیری خودنظارتی و ترنسفورمرهای بصری ارائه‌می‌شود. بخش ۳ به‌تفصیل روش پیشنهادی شامل تعریف مسئله، مرحله یادگیری خودنظارتی بازنمایی چهره، معماری ترنسفورمر بصری، مدل‌سازی رابطه‌ای میان جفت چهره‌ها و تابع هدف نهایی را تشریح می‌کند. همچنین در این بخش، مجموعه‌داده‌های مورداستفاده و پارامترهای پیاده‌سازی معرفی می‌گردند. بخش ٤ نتایج تجربی، تحلیل‌های کمی و کیفی، مقایسه با روش‌های پایه و پیشرفته و بحث پیرامون یافته‌ها را شامل می‌شود. درنهایت، بخش ۵ نتیجه‌گیری کلی پژوهش، جمع‌بندی دستاوردها و ارائه مسیرهای پیشنهادی برای تحقیقات آتی را ارائه می‌دهد.

۲- مرور پیشینه تحقیق

تشخیص روابط خویشاوندی از طریق تصاویر چهره، به‌عنوان یکی از مسائل تخصصی در حوزه تحلیل چهره مطرح است که در ماهیت با مسئله شناسایی هویت تفاوت بنیادین دارد. درحالی‌که شناسایی هویت بر تطبیق دقیق افراد تمرکز دارد، تشخیص خویشاوندی جستجوی شباهت‌های موروثی غیرمستقیم و ظریف را هدف قرار می‌دهد [۷]. رویکردهای اولیه در این حوزه عمدتاً بر ویژگی‌های دست‌ساز مانند الگوهای دودویی محلی[2] [۸]، هیستوگرام گرادیان‌های جهت‌دار[3][۹] و توصیف‌گر [۱۰] تکیه داشتند که از نواحی مختلف چهره استخراج و سپس با استفاده از طبقه‌بندهایی مانند ماشین بردار پشتیبان[4][۱۱] یا جنگل تصادفی پردازش می‌شدند. اگرچه این روش‌ها از سادگی مفهومی برخوردار بودند، ولی در مواجهه با تغییرات عملی در زاویه دید، سن، حالات چهره و شرایط نوری، از قدرت تعمیم‌پذیری ضعیفی رنج می‌بردند. ظهور یادگیری عمیق انقلابی در این زمینه ایجاد کرد. شبکه‌های عصبی کانولوشنی[5] (CNN) با توانایی یادگیری ویژگی‌های سلسله‌مراتبی و انتزاعی، دقت تشخیص را به میزان قابل‌توجهی بهبود بخشیدند [۱۲]. معماری‌های جفتی و سه‌تایی مبتنی بر CNN، امکان یادگیری رابطه‌ای میان جفت تصاویر را فراهم کردند

[1] General representations
[2] Local Binary Patterns
[3] Histogram of Oriented Gradients
[4] Support Vector Machines
[5] Convolutional Neural Networks

[۱۳]. باوجود دستیابی به نتایج قابل‌قبول روی مجموعه‌داده‌های معیاری مانند KinFaceW-I[1]، KinFaceW-II[2] و FIW[3]، این مدل‌ها در شرایط واقعی و غیرکنترل‌شده با محدودیت مواجه شدند. وابستگی شدید به داده‌های برچسب‌خورده و حساسیت به سوگیری موجود در داده‌های آموزشی، از موانع اصلی تعمیم‌پذیری آن‌ها به‌شمارمی‌رود.

برای غلبه بر محدودیت‌های ذاتی یادگیری نظارت‌شده، پژوهش‌های اخیر به سمت یادگیری خودنظارتی گرایش یافته‌اند [۱۳]. این رویکرد با تعریف وظایف پیش‌متنی بر روی حجم انبوهی از داده‌های بدون برچسب، قادر به یادگیری بازنمایی‌های معنادار و عمومی است [٦]. چارچوب‌هایی مانند SimCLR، MoCo و BYOL در یادگیری بازنمایی چهره موفقیت‌های چشمگیری نشان داده‌اند [١٤]. بااین‌حال، کاربرد مستقیم این روش‌های عمومی در تشخیص خویشاوندی با چالش مواجه است، زیرا تشخیص خویشاوندی نیازمند مدل‌سازی ویژگی‌های ظریف، پراکنده و غیرمحلی است که به‌صورت موروثی بین نسل‌ها منتقل می‌شوند، امری که روش‌های عمومی مقابله‌ای به‌طور خاص برای استخراج آن طراحی نشده‌اند.

به‌موازات این تحولات، ترنسفورمرهای بصری (ViT)[4] به‌عنوان جایگزینی قدرتمند برای شبکه‌های کانولوشنی ظهور کرده‌اند **[۱۵]**. **ViT دسته‌ای از معماری‌های یادگیری عمیق هستند که برای پردازش تصاویر، از سازوکار توجه[5] به‌جای لایه‌های کانولوشنی استفاده می‌کنند. در این معماری، تصویر ورودی به قطعات کوچک و غیرهم‌پوشان تقسیم‌شده و هر قطعه به یک بردار تعبیه‌شده تبدیل می‌شود. سپس این بردارها همراه با بردار موقعیتی، به یک شبکه ترنسفورمر متشکل از چندین لایه توجه چندسری[6] و شبکه‌های پیشخور**[7] داده می‌شوند. سازوکار توجه به مدل اجازه می‌دهد تا وابستگی‌های بلندبرد و سراسری میان نواحی مختلف تصویر را به‌طور مستقیم مدل‌سازی کند؛ ویژگی‌ای که در شبکه‌های کانولوشنی به دلیل ماهیت محلی آن‌ها به‌سختی قابل‌دستیابی است. این قابلیت، ترنسفورمرهای بصری را برای مسائلی مانند تشخیص خویشاوندی که در آن شباهت‌های ژنتیکی به‌صورت پراکنده و غیرمحلی در سراسر چهره ظاهر می‌شوند، بسیار مناسب ساخته است. نقطه قوت این معماری‌ها در توانایی مدل‌سازی وابستگی‌های بلندبرد و سراسری در تصویر از طریق مکانیزم توجه، نهفته است [١٦]. برخلاف CNN ها که بر پیوندهای محلی متکی هستند، ViT ها با استفاده از توجه چندسری[8] می‌توانند ارتباط بین نواحی دور از هم چهره (مانند ارتباط فرم چانه با پیشانی یا تقارن‌های بین‌نسلی) را مستقیماً مدل کنند [١٧]. اگرچه کاربرد موفق ViT ها در مسائل عمومی تحلیل چهره همچون شناسایی هویت یا تشخیص احساس از CNN ها پیشی گرفته است [١٨]، ولی تطبیق و به‌کارگیری هدفمند این معماری برای مسئله تشخیص خویشاوندی هنوز در مراحل ابتدایی قراردارد.

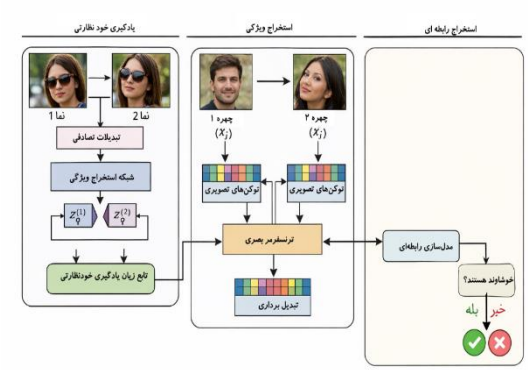
با بررسی پیشینه تحقیق، شکاف پژوهشی آشکاری، مشاهده می‌شود که باوجود بهینه‌بودن هر یک از روش‌های یادشده، چارچوب یکپارچه‌ای که قدرت استدلال سراسری ViT ها را با بهره‌وری از برچسب یادگیری خودنظارتی ترکیب کرده‌باشد، مشاهده نگردید. روش‌های موجود یا بر یادگیری نظارت‌شده با ViT متمرکزند یا از یادگیری خودنظارتی بهره می‌برند.

۳- روش‌ها و داده‌های تحقیق

در این پژوهش، یک چارچوب یادگیری عمیق چندمرحله‌ای برای تشخیص روابط خویشاوندی مبتنی بر تصاویر چهره ارائه می‌شود که باهدف استخراج بازنمایی‌های پایدار، تعمیم‌پذیر و رابطه‌محور طراحی‌شده است. ایده اصلی روش پیشنهادی بر این فرض استوار است که شباهت‌های موروثی چهره، الگوهایی ظریف، پراکنده و غیرمحلی هستند که تنها با تکیه‌بر یادگیری نظارت‌شده و داده‌های محدود به‌خوبی قابل مدل‌سازی نیستند. بنابراین، در چارچوب پیشنهادی، ابتدا از یادگیری خودنظارتی برای کسب دانش عمومی

[1] https://www.kinfacew.com/dataset/kinfacew-i-zip
[2] https://www.kinfacew.com/dataset/kinfacew-ii-zip
[3] https://fulab.sites.northeastern.edu/fiw-download/
[4] Vision Transformer
[5] Attention
[6] Multi-Head Self-Attention
[7] Feed-Forward Networks
[8] Multi-Head Self-Attention

از ساختار چهره انسانی استفاده‌شده و سپس این دانش در یک فرآیند نظارت‌شده، برای مدل‌سازی روابط خویشاوندی پالایش می‌شود. نمای کلی روش در شکل (۱) نشان داده‌شده است.



شکل (۱): نمای کلی روش پیشنهادی

۳-۱ تعریف مسئله و نمادگذاری ریاضی

فرض می‌شود مجموعه داده‌ای از تصاویر چهره به‌صورت رابطه (۱) در اختیار است:

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \quad (۱)$$

که در آن $\mathbf{x}_i \in \mathbb{R}^{H\times W\times C}$ نشان‌دهنده یک تصویر چهره با ارتفاع H، عرض W و تعداد کانال‌های رنگی C(معمولاً برابر ۳) است. برای مرحله تشخیص خویشاوندی، مجموعه‌ای از جفت تصاویر به‌همراه برچسب رابطه (۲) تعریف می‌شود:

$$\mathcal{D} = \{(x_i, x_j, y_{ij})\} \quad (۲)$$

که در آن:

$$y_{ij} = \begin{cases} 1 & \text{اگر } x_i \text{ و } x_j \text{ خویشاوند باشند} \\ 0 & \text{در غیر اینصورت} \end{cases} \quad (۳)$$

هدف، یادگیری تابع تصمیم‌گیری پارامتریزه‌شده $f_\theta(\cdot)$ است همان‌طور که در رابطه (٤) نمایش داده‌شده است:

$$\hat{y}_{ij} = f_\theta(x_i, x_j) \quad (٤)$$

بطوریکه اختلاف میان $\hat{y}_{ij}$ و y_{ij} کمینه‌شود.

۳-۲ مرحله اول: یادگیری خودنظارتی بازنمایی چهره

در مرحله نخست، از مجموعه بزرگی از تصاویر چهره بدون استفاده از برچسب‌های خویشاوندی بهره گرفته می‌شود. برای هر تصویر ورودی x، دو نمای متفاوت از طریق اعمال تبدیلات تصادفی داده‌افزا مانند رابطه (۵) تولید می‌گردد:

$$x^{(1)} = t_1(x), x^{(2)} = t_2(x) \quad (۵)$$

که $t(\cdot)$ شامل عملیات‌هایی نظیر برش تصادفی، تغییر مقیاس، وارونگی افقی، تغییر روشنایی و محوشدگی است. هدف از این تبدیلات، وادارکردن مدل به یادگیری ویژگی‌هایی مستقل از تغییرات سطحی تصویر است. مرحله یادگیری خودنظارتی با استفاده از مجموعه‌ای بزرگ از تصاویر چهره بدون برچسب خویشاوندی انجام‌شده است که از دیتاست عمومی VGGFace2[1] استخراج گردیده‌اند. این مجموعه شامل هویت‌هایی مستقل از دیتاست‌های ارزیابی بوده و هیچ‌گونه هم‌پوشانی فردی با داده‌های مرحله آزمون ندارد. بدین ترتیب، از بروز نشت اطلاعات و سوگیری در فرآیند ارزیابی جلوگیری شده است.

هر یک از نماها مطابق با رابطه (٦) توسط شبکه استخراج ویژگی $g_\phi(\cdot)$ به فضای ویژگی نگاشت می‌شوند:

$$\mathbf{z}^{(1)} = g_\phi(x^{(1)}), \mathbf{z}^{(2)} = g_\phi(x^{(2)}) \quad (٦)$$

تابع زیان خودنظارتی توسط رابطه (۷) به‌گونه‌ای تعریف می‌شود که شباهت میان بازنمایی‌های مربوط به یک تصویر افزایش‌یافته و شباهت آن‌ها با سایر نمونه‌ها کاهش‌یابد:

$$\mathcal{L}_{ssl} = -\log \frac{\exp(\text{sim}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(\mathbf{z}^{(1)}, \mathbf{z}_k)/\tau)} \quad (۷)$$

که در آن $\text{sim}(\cdot)$ شباهت کسینوسی، τ پارامتر دما، و K تعداد نمونه‌های موجود در دسته آموزشی است. در تابع زیان خودنظارتی رابطه (۷)، مقدار پارامتر دما (τ) برابر با ۰/۰۷ تنظیم شده‌است.

این فرآیند موجب می‌شود مدل بازنمایی‌هایی بیاموزد که نسبت به نویزهای محیطی مقاوم بوده و ساختارهای اساسی چهره را کدگذاری کنند.

۳-۳ استخراج ویژگی مبتنی بر ترنسفورمر بصری

برای مدل‌سازی روابط فضایی و ساختاری چهره، از معماری ترنسفورمر بصری استفاده می‌شود. در این معماری، تصویر ورودی براساس رابطه (۸) به N قطعه غیرهم‌پوشان با اندازه ثابت تقسیم می‌شود:

[1] https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/

$$x \to \{p_1, p_2, \dots, p_N\} \quad (۸)$$

برای سادگی نمادگذاری، پس از عملیات تخت‌سازی نیز همچنان از نماد p_i برای نمایش بردار متناظر با قطعه iام استفاده می‌شود. هر قطعه نمایش داده شده در رابطه (۹) به یک بردار تعبیه‌شده نگاشت‌شده و همراه با بردار موقعیتی وارد شبکه می‌شود:

$$z_i = E_{pi} + e^i_{pos} \quad (۹)$$

در رابطه فوق، p_i بردار متناظر با قطعه iام تصویر پس‌ازآنجام عملیات تخت‌ساز است. هر قطعه تصویر دارای ابعاد $P \times P \times C$ است که در آن P اندازه هر ضلع قطعه و C تعداد کانال‌های رنگی تصویر است. با اعمال عملیات تخت‌سازی، تمامی مقادیر پیکسل‌های موجود در قطعه به یک بردار یک‌بعدی تبدیل می‌شوند. ازآنجاکه تعداد کل عناصر موجود در هر قطعه برابر با $P \times P \times C$ است، طول بردار حاصل برابر با P^2C خواهد بود. بنابراین بردار قطعه در فضای R^{P^2C} قرارمی‌گیرد. به‌عنوان‌مثال، در این پژوهش اندازه هر قطعه برابر با 16×16 پیکسل و تعداد کانال‌های رنگی برابر با 3 در نظر گرفته‌شده است؛ درنتیجه بعد بردار حاصل برابر با $16 \times 16 \times 3 = 768$ خواهد بود. این بردار توسط ماتریس تعبیه خطی Eبه فضای ویژگی با بعد D نگاشت‌شده و پس از افزودن بردار موقعیتی e^i_{pos}، بردار نهایی z_i به‌عنوان ورودی لایه‌های ترنسفورمر مورداستفاده‌قرارمی‌گیرد. سازوکار توجه به‌صورت رابطه (۱۰) تعریف می‌شود:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (۱۰)$$

که در آن Q، K و V به‌ترتیب ماتریس‌های پرسش، کلید و مقدار هستند. این سازوکار به مدل اجازه می‌دهد وابستگی‌های بلندبرد میان نواحی مختلف چهره را بیاموزد؛ ویژگی‌ای که برای تشخیص شباهت‌های ژنتیکی پراکنده بسیار حیاتی است.

برای مدل‌سازی روابط فضایی و ساختاری چهره، از معماری ترنسفورمر بصری استفاده می‌شود. در این معماری، تصویر ورودی به قطعات غیرهم‌پوشان با اندازه ثابت تقسیم‌شده و پس از تبدیل خطی و افزودن رمزگذاری موقعیت، از چندین لایه توجه چندسری عبورمی‌کند. سازوکار توجه به مدل اجازه می‌دهد تا وابستگی‌های بلندبرد میان نواحی مختلف چهره را به‌طور مستقیم مدل‌سازی کند؛ ویژگی‌ای که برای تشخیص شباهت‌های ژنتیکی پراکنده در مسئله خویشاوندی حیاتی است. مشخصات دقیق معماری ترنسفورمر بصری به‌کاررفته در این پژوهش، در جدول (۱) ارائه‌شده است. این معماری بر اساس نسخه استانداردViT-Base/16 با اعمال تغییرات مناسب برای مسئله تشخیص خویشاوندی طراحی‌شده است.

جدول (۱): مشخصات ترانسفورمر بصری روش پیشنهادی

پارامتر	مقدار
اندازه ورودی/ قطعه/ تعداد قطعات	۲۲٤*۲۲٤ /۱٦*۱٦/ ۱۹٦
ابعاد (پنهان /MLP / هرسر)	٦٤/ ۳۰۷۲ / ۷٦۸
لایه‌ها/ سرها/ فعال سازی	GELU/۱۲/ ۱۲
نرمالیزاسیون / اتصال باقیمانده	LayerNorm(pre-LN) /دارد
توکن/ رمزگذاری موقعیت Dropout/CLS]]	۰/۱ / دارد / قابل یادگیری
خروجی	بردار LayerNorm + [CLS] نهایی (بعد ۷٦۸)

٤-۳ مدل‌سازی رابطه‌ای میان جفت چهره‌ها

در مرحله نظارت‌شده، هر جفت تصویر (x_i, x_j) به‌صورت مجزا از شبکه استخراج ویژگی بیان‌شده در رابطه (۱۱) عبورداده‌می‌شود:

$$\mathrm{h}_i = g_\phi(x_i), \mathrm{h}_j = g_\phi(x_j) \quad (۱۱)$$

برای مدل‌سازی کنش متقابل میان این دو بازنمایی، یک پودمان توجه بین‌فردی همانند رابطه (۱۲) اعمال می‌شود:

$$\mathrm{r}_{ij} = \mathrm{Attn}_{inter}(\mathrm{h}_i, \mathrm{h}_j) \quad (۱۲)$$

این بردار رابطه‌ای شامل اطلاعات مکمل و مقایسه‌ای دو چهره است و به‌عنوان ورودی به یک طبقه‌بند دودویی داده می‌شود.

در مرحله نظارت‌شده، تمامی اجزای مدل شامل شبکه استخراج ویژگی، پودمان توجه بین‌فردی و طبقه‌بند نهایی به‌صورت سرتاسر[1] و هم‌زمان آموزش داده می‌شوند. به این معنا که گرادیان خطا از خروجی نهایی تا لایه‌های اولیه شبکه استخراج ویژگی منتشرشده و تمامی وزن‌ها به‌طور یکپارچه به‌روزرسانی می‌گردند.

[1] End-to-End

۳-۵ تابع هدف نهایی

تابع زیان کلی مدل به‌صورت ترکیبی مانند رابطه (۱۳) تعریف می‌شود:

$$L_{\text{total}} = L_{\text{BCE}} + \lambda_1 L_{\text{distance}} + \lambda_2 L_{\text{contrastive}} \quad (۱۳)$$

که در آن L_{BCE} زیان آنتروپی متقاطع دودویی برای طبقه‌بندی نهایی است، L_{distance} زیان فاصله‌ای (مانند فاصله اقلیدسی) برای تنظیم فضای ویژگی و نزدیک کردن بازنمایی جفت‌های خویشاوند به‌یکدیگر است، و $L_{\text{contrastive}}$ زیان کنتراستیو مشابه رابطه (۷) است که در مرحله خودنظارتی نیز استفاده‌شده و برای حفظ بازنمایی‌های مقاوم به‌کارمی‌رود. مقادیر عددی فراپارامترها به اینصورت تعیین می‌شود که (۱) ضریب تنظیم زیان فاصله‌ای $\lambda_1 = 0.5$،(۲) زیان کنتراستیو $\lambda_2 = 0.3$، و (۳) دمای تابع زیان خودنظارتی $\tau = 0.07$.

نحوه انتخاب فراپارامترها به اینگونه است که کلیه فراپارامترهای مدل پیشنهادی از طریق جستجوی شبکه‌ای بر روی مجموعه اعتبارسنجی (۱۵٪ از داده‌های آموزشی) تعیین‌شده‌اند. محدوده جستجو برای هر فراپارامتر به‌صورت زیر در نظر گرفته شد: نرخ یادگیری در بازه $[10^{-5}, 10^{-3}]$، پارامتر دما (τ) در بازه $[0.01, 0.5]$، ضرایب λ_1 و λ_2 در بازه $[0.1, 0.9]$و نرخ دراپ‌اوت[1] در بازه $[0.05, 0.5]$. تعداد لایه‌های ترنسفورمر (۱۲ لایه) و تعداد سرهای توجه (۱۲ سر) بر اساس معماری استاندارد ViT-Base، انتخاب‌گردیده‌است. اندازه قطعه (۱۶ پیکسل) و بعد فضای پنهان (۷۶۸) نیز مطابق با معماری اصلی ViT در نظر گرفته‌شده است. درنهایت، مقادیری که بیشترین دقت را بر روی مجموعه اعتبارسنجی ارائه دادند، به‌عنوان مقادیر نهایی انتخاب شدند.

همان‌طور که در بخش ۳-۴ اشاره شد، آموزش مدل به‌صورت سرتاسر انجام می‌شود. مقادیر فراپارامترهای تابع هدف ترکیبی به این‌صورت تعیین‌شده است.

۳-۶ مجموعه‌داده‌ها

همان‌طور که در جدول (۲) فهرست شده است، در این پژوهش، برای ارزیابی عملکرد مدل پیشنهادی از دو مجموعه‌داده مرجع و پرکاربرد در حوزه تشخیص خویشاوندی استفاده‌شده است.

- **مجموعه داده اول(KinFaceW-I)**: این مجموعه داده شامل تصاویر چهره در شرایط کنترل‌شده (نورپردازی یکنواخت، پس‌زمینه ساده، وضعیت روبه‌جلو) است. تعداد کل تصاویر ۱۰۶۸ تصویر مربوط به ۵۰۰ فرد مختلف است که به‌صورت ۱۰۶۸ جفت تصویر والد- فرزند سازمان‌دهی شده‌اند. روابط موجود شامل پدر-پسر، پدر-دختر، مادر-پسر و مادر-دختر است.

جدول (۲): مشخصات مجموعه داده‌های مورداستفاده درروش پیشنهادی

مجموعه‌داده	نوع روابط	تعداد جفت‌ها	تعداد افراد	نوع محیط
KinFaceW-I	والد-فرزند	۱۰۶۸	۵۰۰	کنترل‌شده
KinFaceW-II / FIW	والد-فرزند، خواهر-برادر	۱۵۰۰۰	۷۰۰	غیرکنترل‌شده

- مجموعه داده دوم (KinFaceW-I / KinFaceW-II و FIW): این مجموعه داده شامل تصاویر چهره در محیط‌های غیرکنترل‌شده (در دنیای واقعی) با تنوع بالا در نورپردازی، ژست، سن، حالات چهره و کیفیت تصویر است. تعداد کل جفت‌ها ۱۵،۰۰۰ جفت شامل روابط والد-فرزند و خواهر-برادر است که از ۷۰۰ فرد مختلف گردآوری‌شده‌اند. این مجموعه داده چالش‌برانگیزتر بوده و برای ارزیابی تعمیم‌پذیری مدل در شرایط واقعی مناسب است. همچنین پارامترهای آموزشی مدل که متشکل از تعداد لایه‌ها، نرخ یادگیری وغیره است در جدول (۳) واردشده است.
- نتایج مربوط به مجموعه داده کنترل‌شده KinFaceW-I صرفاً در جدول (۴) و شکل (۲) که به افت عملکرد در انتقال از

[1] Dropout

محیط کنترل‌شده به غیرکنترل‌شده می‌پردازد، مورداستفاده قرارگرفته است.

جدول (۳): پارامترهای آموزشی مدل درروش پیشنهادی

پارامتر	مقدار
اندازه تصویر	۲۲٤*۲۲٤
تعداد لایه ترانسفورمر	۱۲
تعداد سرهای توجه	۱۲
بهینه‌ساز	AdamW
نرخ یادگیری	۱*۱۰^-٤
اندازه دسته	٦٤
تعداد دوره‌ها	۱۰۰
نرخ دراپ اوت	۰/۱

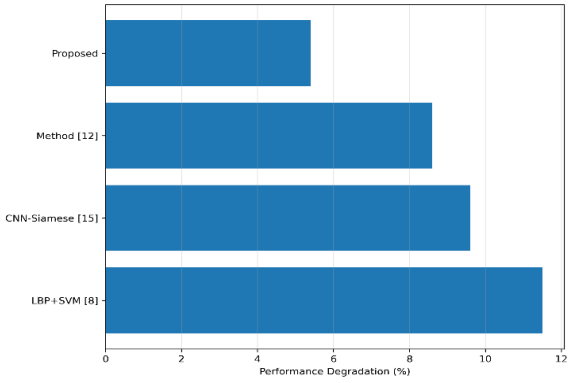
٤- نتایج و بحث

پیش از ارائه نتایج، لازم به یادآوری است که روش پیشنهادی از دو مرحله تشکیل‌شده است: مرحله اول یادگیری خودنظارتی بازنمایی‌های عمومی چهره بر روی مجموعه داده VGGFace2 و مرحله دوم پالایش این بازنمایی‌ها با استفاده از پودمان توجه بین‌فردی مبتنی بر ترنسفورمر بصری. تمام پیاده‌سازی‌ها با زبان Python و کتابخانه PyTorch نسخه ۱,۱۲ انجام‌شده و آموزش مدل بر روی دو GPU NVIDIA Tesla V100 با ۳۲ گیگابایت حافظه صورت گرفته است. در ادامه، ابتدا نتایج روش پیشنهادی با معیارهای دقت، در F1-score و AUC-PR ارائه و تحلیل می‌شود، سپس مقایسه با روش‌های دیگر در مراجع [۸،۱۲ و ۱۵]انجام‌شده و تحلیل مربوطه آورده می‌شود. لازم به توضیح است که کلیه نتایج گزارش‌شده در این بخش، مگر اینکه خلاف آن ذکر شود، برروی مجموعه داده غیرکنترل‌شده KinFaceW-II/FIW به‌دست‌آمده است.

جدول (٤): مقایسه دقت روش‌های مختلف در دو محیط کنترل‌شده و غیرکنترل‌شده

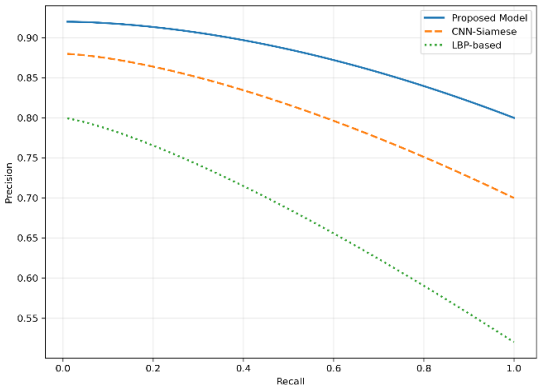
ردیف	روش	افت عملکرد (درصد)	دقت در محیط غیرکنترل‌شده(KinFaceW-II/FIW)	دقت در محیط کنترل‌شده(KinFaceW-I)
۱	LBP + SVM	٪۱۱/۲	٪٦۱/۳	٪۷۲/۵
۲	CNN سیامی	٪۹/۳	٪۷٦/۸	٪۸٦/۱
۳	ViT بدون خودنظارتی	٪۸/۲	٪۸۱/۲	٪۸۹/٤
٤	روش پیشنهادی	٪٦/٦	٪۸۷/٦	٪۹٤/۲

همان‌طور که در شکل (۲) مشاهده می‌شود، روش پیشنهادی با افت عملکرد تنها ٪٦/٦ در انتقال از محیط کنترل‌شده (KinFaceW-I) به محیط غیرکنترل‌شده (KinFaceW-II/FIW)، کمترین کاهش دقت را در میان تمام روش‌های مقایسه‌شده دارد. این در حالی است که روش مبتنی بر ویژگی‌های دست‌ساز LBP+SVM با افت ٪۱۱/۲ CNN سیامی با افت ٪۹/۳ و ViT بدون خودنظارتی با افت ٪۸/۲ کاهش دقت بسیار بیشتری را تجربه می‌کنند. این نتایج به‌وضوح نشان می‌دهد که ترکیب یادگیری خودنظارتی با معماری ترنسفورمر بصری، نه تنها دقت را در شرایط ایده‌ال افزایش می‌دهد، بلکه پایداری و تعمیم‌پذیری مدل را در مواجهه با تغییرات چالش‌برانگیز دنیای واقعی به‌طور قابل‌توجهی بهبود می‌بخشد.



شکل (۲): افت عملکرد نسبی روش‌های مختلف

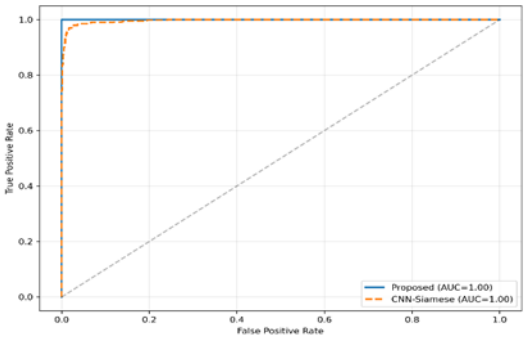
تنها گزارش یک مقدار دقت ثابت، برای ارزیابی طبقه بندها کافی نیست؛ بلکه پایداری عملکرد در آستانه‌های مختلف تصمیم‌گیری اهمیت اساسی دارد. همان‌گونه که در شکل (۳) مشاهده می‌شود، منحنی دقت–بازخوانی[1] مدل پیشنهادی در مقایسه با سایر روش‌ها، سطح بالاتری را در کل بازه بازخوانی حفظ می‌کند. این رفتار نشان می‌دهد که مدل نه‌تنها در تشخیص جفت‌های خویشاوند عملکرد مناسبی دارد، بلکه نرخ هشدارهای کاذب آن نیز کنترل‌شده باقی می‌ماند. در کاربردهایی مانند جست‌وجوی خویشاوندی در پایگاه‌های داده بزرگ، چنین توازنی میان دقت و بازخوانی اهمیت حیاتی دارد.



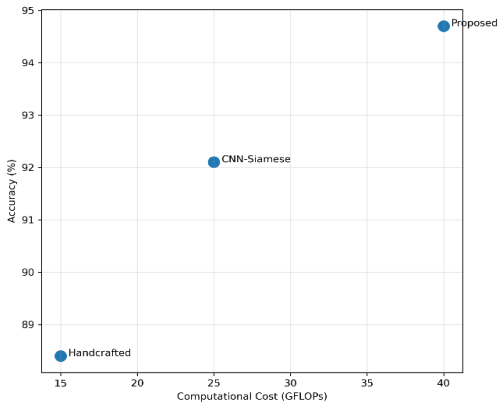
شکل (۳): منحنی‌های دقت – بازخوانی

برای بررسی کیفیت بازنمایی‌های آموخته‌شده، توزیع ویژگی‌های استخراج‌شده توسط مدل‌های مختلف در فضای کم‌بعد نمایش داده شده است. همان‌گونه که در شکل (٤) مشاهده می‌شود، بازنمایی‌های مدل پیشنهادی خوشه‌بندی منسجم‌تری میان جفت‌های خویشاوند ایجاد می‌کنند، درحالی‌که هم‌پوشانی قابل‌توجهی میان نمونه‌های مثبت و منفی درروش‌های مبتنی بر CNN مشاهده می‌شود. این جداسازی بهتر، نقش کلیدی در بهبود دقت نهایی و کاهش خطاهای تصمیم‌گیری ایفا می‌کند.

اگرچه معماری پیشنهادی ازنظر محاسباتی پیچیده‌تر از روش‌های مبتنی بر CNN است، بررسی نسبت عملکرد به هزینه نشان می‌دهد که این افزایش هزینه توجیه‌پذیر است. مطابق شکل (۵)، مدل پیشنهادی درازای افزایش محدود پیچیدگی، بهبود قابل‌توجهی در دقت ارائه می‌دهد. این نسبت، به‌ویژه در کاربردهایی که دقت اولویت بالاتری نسبت به زمان پاسخ دارد، اهمیت عملی بالایی دارد.



شکل (٤): نمایش بصری ابعاد پایین بازنمایی‌های ویژگی یادگرفته‌شد



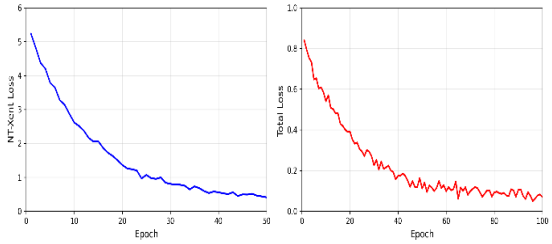
شکل (۵): تعادل عملکرد–پیچیدگی درروش‌های مختلف تأیید خویشاوندی

برای بررسی روند یادگیری و اطمینان از همگرایی مدل، نمودار مقدار تابع هزینه[2] در طول دوره‌های آموزش برای هر دو مرحله روش پیشنهادی در شکل (٦) ارائه‌شده است. در مرحله خودنظارتی، مقدار زیان NT-Xent از حدود ۵/۲ در دوره اول به حدود ۰/٤ در دوره پنجاهم کاهش‌یافته و پس‌ازآن تقریباً ثابت می‌ماند که نشان‌دهنده یادگیری موفق بازنمایی‌های عمومی است. در مرحله نظارت‌شده، زیان ترکیبی L_{total} از حدود ۰/۸ به کمتر از ۰/۱ در دوره شصتم کاهش می‌یابد و تا پایان دوره صدم پایدار

[1] Precision-Recall

[2] Loss

می‌ماند. عدم نوسان شدید در هر دو نمودار نشان‌دهنده انتخاب مناسب نرخ یادگیری و پایداری فرآیند آموزش است.



شکل (٦): نمودار مقدار تابع هزینه در حین آموزش مدل پیشنهادی

۵- نتیجه‌گیری

در این مقاله، یک چارچوب یادگیری عمیق برای مسئله تشخیص خویشاوندی مبتنی بر چهره ارائه شد که با ترکیب یادگیری خودنظارتی و مدل‌سازی توجه‌محور، به‌دنبال افزایش دقت، پایداری تصمیم‌گیری و تعمیم‌پذیری در شرایط واقعی است.

برخلاف بسیاری از مطالعات پیشین که به داده‌های برچسب‌خورده محدود و سناریوهای کنترل‌شده متکی هستند، چارچوب پیشنهادی ابتدا با استفاده از داده‌های بدون برچسب، بازنمایی‌های عمومی و مقاومی از ساختار چهره انسانی می‌آموزد و سپس این بازنمایی‌ها را برای استدلال رابطه‌ای میان جفت چهره‌ها پالایش می‌کند.

نتایج تجربی نشان داد که روش پیشنهادی با دستیابی به‌دقت ۹۴/۲٪ در محیط کنترل‌شده و ۸۷/٦٪ در محیط غیرکنترل‌شده، بهترین عملکرد را در میان روش‌های مقایسه‌ای ارائه کرده است. همچنین افت عملکرد این روش در انتقال به شرایط واقعی تنها ٦/٦٪ بوده که کمتر از سایر روش‌های موردبررسی است. به‌ویژه، کاهش افت عملکرد در محیط‌های غیرکنترل‌شده و جداسازی مؤثرتر فضای ویژگی‌ها نشان‌می‌دهد که یادگیری خودنظارتی نقش کلیدی در افزایش توان تعمیم‌پذیری مدل ایفا می‌کند. علاوه بر این، تحلیل نسبت عملکرد به هزینه محاسباتی نشان داد که افزایش پیچیدگی معماری مدل پیشنهادی در برابر بهبود قابل‌توجه دقت و پایداری تصمیم‌گیری، توجیه‌پذیر است. این ویژگی، چارچوب ارائه‌شده را به گزینه‌ای مناسب برای کاربردهای عملی تشخیص خویشاوندی، به‌ویژه در سناریوهای نزدیک به دنیای واقعی با داده‌های ناهمگون، تبدیل می‌کند. باوجود نتایج امیدوارکننده، این پژوهش همچنان با محدودیت‌هایی همراه است. در این مطالعه، ارزیابی مدل در وضعیت‌های میان‌مجموعه‌ای موردبررسی قرارنگرفته و تحلیل‌های رسمی نیز ارائه نشده است. از سوی‌دیگر، بررسی ترکیب مجموعه داده‌های مورداستفاده نشان‌می‌دهد که ازنظر نژاد و قومیت، داده‌ها عمدتاً محدود به جمعیت‌های قفقازی و آسیایی هستند و تنوع کافی برای نژادهای آفریقایی، اسپانیایی و سایرگروه‌ها وجود ندارد. همچنین ازنظر گروه‌های سنی، داده‌های مربوط به سالمندان (بالای ٦٠ سال) و خردسالان (زیر ۵ سال) به میزان کافی در مجموعه‌ها حضور ندارند. بنابراین، اگرچه مدل پیشنهادی بر روی داده‌های موجود عملکرد مناسبی دارد، تعمیم نتایج به جمعیت‌های متنوع‌تر نیازمند بررسی بیشتر بر روی مجموعه داده‌های متوازن‌تر است. بر اساس محدودیت‌های ذکرشده، انجام آزمایش‌های آموزش و آزمون بر روی دیتاست‌های ناهمگون، تحلیل عمیق‌تر و جامع‌تر و بررسی رفتار مدل در روابط خویشاوندی پیچیده‌تر، به‌عنوان مسیرهای مهم پژوهشی آینده پیشنهاد می‌شود.

References

[1] M. C. Mzoughi, N. B. Aoun, and S. J. T. V. C. Naouali, "A review on kinship verification from facial information," Visual Computer, vol. 41, no. 3, pp. 1789-1809, 2025. doi:10.1007/s00371-024-03493-1

[2] M. Almuashi, S. Z. Mohd Hashim, D. Mohamad, M. H. Alkawaz, A. J. M. T. Ali, and Applications, "Automated kinship verification and identification through human facial images: a survey," Multimedia Tools and Applications, vol. 76, no. 1, pp. 265-307, 2017. doi:10.1007/s11042-015-3007-5

[3] J. P. Robinson et al., "Visual kinship recognition of families in the wild," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 11, pp. 2624-2637,2018. doi:10.1109/TPAMI.2018.2826549

[4] J. Yu, M. Li, X. Hao, and G. Xie, "Deep fusion siamese network for automatic kinship verification," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 892-899,2020. doi:10.1109/FG47880.2020.00127

[5] Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. J. T. Makedon, "A survey on contrastive self-supervised learning," Technologies, vol. 9, no. 1, p. 2, 2020. doi:10.3390/technologies9010002

[6] Khan et al., "A survey of the vision transformers and their CNN-transformer based variants," Artificial Intelligence Review, vol. 56, no. Suppl 3, pp. 2917-2970, 2023. doi:10.1007/S10462-023-10595-0

[7] G. Guo, X. J. I. T. o. I. Wang, and Measurement, "Kinship measurement on salient facial features," IEEE Transactions on Instrumentation and Measurement, vol. 61, no. 8, pp. 2322-2325, 2012. doi:10.1109/TIM.2012.2187468

[8] A. Chergui, S. Ouchtati, J. Sequeira, S. E. Bekhouche, and F. Bougourzi, "Kinship verification using BSIF and LBP," in 2018 International Conference on Signal, Image, Vision and their Applications (SIVA), pp. 1-5,2018. doi:10.1109/SIVA.2018.8661085

[9] Z. Zhou and Y. Zhou, "Cross-channel similarity based histograms of oriented gradients for color images," in 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 1621-1625, 2018. doi:10.1109/SMC.2019.8913933

[10] A. Puthenputhussery, Q. Liu, and Liu, "Sift flow based genetic fisher vector feature for kinship verification," in 2016 IEEE international conference on image processing (ICIP), pp. 2921-2925,2016. doi:10.1109/icip.2016.7532814

[11] N. Nader, F. E.-Z. El-Gamal, S. El-Sappagh, K. S. Kwak, and M. J. P. C. S. Elmogy, "Kinship verification and recognition based on handcrafted and deep learning feature-based techniques," PeerJ Computer Science, vol. 7, p. e735, 2021. doi:10.7717/peerj-cs.735

[12] M. M. Sadr, M. Khani, S. M. J. B. S. P. Tootkaleh, and Control, "Predicting athletic injuries with deep Learning: Evaluating CNNs and RNNs for enhanced performance and Safety," Biomedical Signal Processing and Control, vol. 105, p. 107692, 2025. doi:10.1016/j.bspc.2025.107692

[13] A. Menon, R. Yashwanthika, J. Shreya, K. Somasundaram, S. K. Thangavel, and S. Asawa, "Person Re-Identification Using Siamese Triplet Network," in 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL), 2025, pp. 391-397: IEEE. doi:10.1109/ICSADL65848.2025.10933063

[14] N. Giakoumoglou, T. Stathaki, and A. J. a. p. a. Gkelias, "A Review on Discriminative Self-supervised Learning Methods in Computer Vision," arXiv preprint arXiv:2405.04969, 2024. doi:10.48550/arXiv.2405.04969

[15] N. Fnu and A. Bansal, "Understanding the architecture of vision transformer and its variants: A review," in 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1-6,2024.doi:10.1109/ICIESTR60916.2024.10798341

[16] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. J. A. c. s. Shah, "Transformers in vision: A survey," in ACM Computing Surveys, vol. 54, no. 10s, pp. 1-41, 2022. doi:10.1145/3505244

[17] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. J. a. p. a. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. doi:10.18653/v1/p19-1580

[18] I. Fibriani, E. M. Yuniarno, R. Mardiyanto, M. H. J. I. J. o. I. E. Purnomo, and Systems, "SSCNetViT: A Hybrid Siamese Sequential Classification-ViT Model for Kinship Recognition in Indonesia Facial Microexpression," in International Journal of Intelligent Engineering and Systems, vol. 18, no. 1, 2025. doi: 10.22266/ijies2025.0229.59

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 4, No. 2, Autumn & Winter 2026, pp. 49 - 60

A Multi-Stage Deep Learning Approach for Facial Kinship Recognition Using Self-Supervision and Attention

Nasrin Rasouli[1], Kamal Mirzaie[2*], Mohsen Sardari Zarchi[3]

[1]Ph.D. student, Department of Computer Engineering, May.C., Islamic Azad University, Maybod, Iran

[2]Assistant Professor, Department of Computer Engineering, May.C., Islamic Azad University, Maybod, Iran

[3]Associate Professor, Department of Computer Engineering, Meybod University, Meybod, Iran

Article Information

Original Research Paper

Received:
2026 February 21

Accepted:
2026 June 22

Keywords:
Kinship recognition, self-supervised learning, vision transformer, inter-personal attention

Corresponding Author*:
kamal.mirzaie@iau.ac.ir

Abstract

Kinship recognition from facial images is a challenging problem in computer vision due to variations in age, pose, illumination, and unconstrained acquisition conditions. The strong reliance of existing methods on limited labeled kinship data significantly restricts their generalization capability in real-world scenarios. This paper proposes a multi-stage deep learning framework for learning hereditary facial similarities by integrating self-supervised representation learning with transformer-based relational modeling. In the first stage, robust and general facial representations are learned from a large collection of unlabeled face images through self-supervised training, without using any kinship annotations. In the second stage, these representations are refined in a supervised manner using an inter-personal attention mechanism, which explicitly enhances subtle genetic similarities between paired faces. The proposed approach is evaluated on benchmark kinship datasets under both controlled and unconstrained conditions. Experimental results demonstrate that the proposed method achieves 94.2% accuracy on the controlled dataset KinFaceW-I and 87.6% accuracy on the unconstrained dataset KinFaceW-II/FIW, corresponding to significant improvements of 4.8% and 6.4% compared to the ViT without self-supervision, respectively. Furthermore, the performance degradation of the proposed method when transferring from controlled to unconstrained environments is only 6.6%, which is substantially lower than that of comparative methods (ranging from 8.2% to 11.2%). The findings indicate that combining self-supervised pretraining with attention-based relational modeling provides an effective solution for reducing dependency on labeled data while improving the reliability and practical applicability of facial kinship recognition systems.

: 10.22034/ABMIR.2026.24353.1227

E-ISSN: 2821-2037