

SSLA-Net: معماری یادگیری معنایی - مکانی برای شناسایی مقاوم نقص‌های سطحی فولاد

معصومه رضائی^{۱*}، منصوره رضائی^۲، مهری رجائی^۳

^۱استادیار گروه مهندسی کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

^۲دکتری هوش مصنوعی، دانشکده مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران

^۳استادیار گروه مهندسی فناوری اطلاعات، دانشگاه سیستان و بلوچستان، زاهدان، ایران

مقاله پژوهشی

چکیده

در صنایع فولاد، شناسایی دقیق و سریع نقص‌های سطحی نقشی حیاتی در تضمین کیفیت و کاهش ضایعات ایفا می‌کند. با این حال، شباهت ظاهری میان الگوهای بافتی و پیچیدگی نواحی معیوب، کارایی روش‌های متداول بینایی ماشین را با محدودیت مواجه کرده است. در این پژوهش، یک چارچوب ترکیبی چندوجهی برای تشخیص و طبقه‌بندی نقص‌های سطحی فولاد ارائه شده است که با تلفیق مدل تشخیص ناحیه‌ای **Faster R-CNN** و نمایش‌های معنایی مدل **CLIP** از طریق ماژول توجه چندسری، ارتباط میان ویژگی‌های بصری و مفهومی را به صورت هم‌زمان مدل‌سازی می‌کند. این هم‌افزایی موجب افزایش دقت در شناسایی نواحی کوچک و پراکنده و همچنین بهبود تفکیک پذیری کلاس‌ها در فضای ویژگی شده است. در طراحی مدل، سه تابع ضرر مکمل شامل ضرر تشخیص ناحیه‌ای، ضرر هم‌ترازی ناحیه-متن و ضرر سطح تصویر به کار رفته‌اند تا یادگیری هم‌زمان سطوح مکانی و معنایی امکان‌پذیر گردد. نتایج تجربی روی مجموعه داده‌ی استاندارد **NEU-DET** نشان می‌دهند که مدل پیشنهادی به دقت ۱۰۰٪ در طبقه‌بندی سطح تصویر و $mAP@50 = 78.04\%$ در سطح ناحیه‌ای دست یافته است. این عملکرد برجسته نشان می‌دهد که یادگیری چندوجهی مبتنی بر **CLIP** می‌تواند دقت و پایداری تشخیص نقص را در شرایط واقعی صنعتی به طور چشمگیری ارتقا دهد.

تاریخ دریافت:

۱۴۰۴/۰۶/۲۲

تاریخ پذیرش:

۱۴۰۴/۰۸/۱۲

کلیدواژه‌ها:

یادگیری چندوجهی، شناسایی نقص‌های سطحی، توجه چندسری، نمایش معنایی-مکانی، بینایی ماشین صنعتی

نویسنده مسئول:

mrezaei@ece.usb.ac.ir

doi : 10.22034/ABMIR.2025.23624.1166

E-ISSN: [2821-2037](#)

© 2025. Published by Yazd University This is an open access article

under the CC BY 4.0 License (<https://creativecommons.org/licenses/by/4.0/>).



۱- مقدمه

فولاد به عنوان یک ماده حیاتی در صنایع استراتژیک از جمله خودروسازی، هوافضا، ساخت و ساز و تولید تجهیزات صنعتی شناخته می‌شود و نقشی کلیدی در کیفیت، ایمنی و دوام محصولات نهایی دارد. وجود هرگونه نقص سطحی نظیر ترک، خراش، پوسته شدگی یا فرورفتگی، منجر به کاهش استحکام مکانیکی، کاهش عمر مفید و بروز خرابی‌های جدی در خطوط تولید می‌گردد. از این رو، تشخیص سریع و دقیق نقص‌های سطح فولاد از الزامات فنی، اقتصادی و ایمنی محسوب می‌شود [۱]. علی‌رغم پیشرفت‌های اخیر در حوزه تشخیص خودکار عیوب سطحی، شناسایی نقص‌های کوچک، پراکنده و دارای بافت پیچیده همچنان چالشی اساسی محسوب می‌شود. این موضوع به‌ویژه در خطوط تولید صنعتی که نیازمند عملکرد پایدار و مقاوم هستند اهمیت بیشتری می‌یابد. به منظور پاسخ‌گویی به این نیازها، در این پژوهش یک چارچوب یادگیری چندوجهی به نام SSLA-Net پیشنهاد شده است که با ادغام اطلاعات مکانی و معنایی، توانایی مدل در تفکیک دقیق کلاس‌ها و شناسایی نقص‌های دشوار را افزایش می‌دهد. روش‌های سنتی تشخیص نقص، شامل بازرسی چشمی و روش‌های غیرمخرب فیزیکی و شیمیایی (مانند التراسونیک و ذرات مغناطیسی)، در محیط‌های صنعتی مدرن با محدودیت‌هایی مواجه هستند. این روش‌ها عمدتاً زمان‌بر، پرهزینه و دارای توانایی محدود در شناسایی نقص‌های کوچک و مبهم هستند [۲]. در نتیجه، تحقیقات به سمت کاربرد یادگیری عمیق و بینایی ماشین معطوف شده است. روش‌های مبتنی بر شبکه‌های کانولوشنی (CNN)، ترنسفورمرها و معماری‌های ترکیبی، پیشرفت‌های قابل توجهی در تشخیص و طبقه‌بندی نقص‌ها در مجموعه داده‌های استاندارد مانند NEU-DET [۳] و GC10-DET [۴] نشان داده‌اند [۵-۸]. با این حال، چالش‌هایی همچنان در زمینه تشخیص نقص‌های کوچک و پیچیده، مرزبندی دقیق کلاس‌های مشابه، و ایجاد نمایش‌های معنایی قوی برای نواحی تصویر وجود دارد.

به منظور رفع محدودیت‌های مذکور، این مطالعه یک چارچوب تلفیقی نوآورانه را ارائه می‌دهد. این چارچوب بر اساس ترکیب Faster R-CNN برای قابلیت‌های ناحیه محور، CLIP برای بازنمایی معنایی، و مکانیزم توجه چندسُرْطراحی شده است. در این روش، علاوه بر استخراج ویژگی‌های ROI^۳ از تصاویر با Faster R-CNN، از بردارهای تعبیه‌شده^۴ متنی تولیدشده توسط CLIP برای هر کلاس نقص نیز استفاده می‌شود. این بردارهای تعبیه شده به عنوان لنگرهای معنایی^۵ عمل کرده و هم‌ترازی بین ROI و مفهوم کلاس را فراهم می‌آورند. این رویکرد دقت شناسایی نقص‌های کوچک و مبهم را افزایش داده و امکان بازنمایی بهتر ویژگی‌های پیچیده و معنایی هر ناحیه تصویر را فراهم می‌آورد.

مزیت‌های کلیدی روش پیشنهادی به شرح زیر است:

(۱) دقت در تشخیص نقص‌های کوچک و پیچیده: ترکیب ویژگی‌های ROI از Faster R-CNN و بردارهای تعبیه‌شده معنایی CLIP، دقت تفکیک کلاس‌ها و شناسایی نقص‌های مبهم را بهبود می‌بخشد. (۲) هم‌ترازی معنایی میان تصویر و متن: استفاده از جملات راهنما با ساختار ساده و بردارهای معنایی از پیش آموزش‌دیده، ارتباطی مستقیم و معنادار میان نواحی تصویری و مفاهیم متنی برقرار می‌سازد. این هم‌ترازی، موجب افزایش دقت و پایداری در فرایند تشخیص می‌شود. (۳) نمایش غنی‌تر ویژگی‌ها با بهره‌گیری از سازوکار توجه چندگانه: مکانیزم توجه چند سر، امکان درک و ثبت وابستگی‌های چندبعدی میان نواحی تصویری و بردارهای معنایی را فراهم می‌سازد. نتیجه آن، نمایش نهایی ویژگی‌ها به صورت دقیق‌تر، جامع‌تر و با قدرت تفکیک بالاتر خواهد بود. (۴) بهبود فرایند آموزش با بهره‌گیری از تابع ضرر ترکیبی: استفاده هم‌زمان از ضرر مرتبط با تشخیص ناحیه‌ای و ضرر سطح کلان تصویر، موجب همگرایی بهتر مدل در مرحله آموزش شده و دقت نهایی را افزایش می‌دهد.

روش پیشنهادی با تلفیق ویژگی‌های ناحیه‌ای، بازنمایی معنایی و توجه چند سر، چارچوبی نوین برای تشخیص نقص‌های سطح

⁴ Embedding

⁵ Semantic anchors

¹ Convolutional Neural Network

² Multi-Head Attention

³ Region of Interest

در سال ۲۰۲۴، روشی مبتنی بر یادگیری عمیق برای تشخیص عیوب سطحی فولاد ارائه شده است [۹]. در این پژوهش، با بهره‌گیری از شبکه‌های عصبی کانولوشنی مختلف از جمله EfficientNetB3 [۱۰] و MobileNetV2 [۱۱] به مقایسه کارایی این معماری‌ها در شناسایی عیوب سطحی پرداختند. نتایج نشان داد که معماری‌های مورد استفاده توانایی بالایی در شناسایی عیوب سطحی دارند. مزیت اصلی این روش‌ها، اتوماسیون فرایند بازرسی و کاهش وابستگی به نیروی انسانی است که منجر به افزایش سرعت و دقت کنترل کیفیت در صنایع فولاد می‌شود. با این حال، پژوهش مذکور اشاره می‌کند که چالش‌هایی همچون نیاز به توازن داده‌ها و هزینه‌های محاسباتی در معماری‌های عمیق همچنان وجود دارد و برای بهبود بیشتر عملکرد، استفاده از روش‌هایی مانند افزایش داده یا بهره‌گیری از معماری‌های سبک‌تر پیشنهاد می‌شود. هان و کوتبای در سال ۲۰۲۵ روشی مبتنی بر یادگیری عمیق پیشنهاد کرده‌اند که شامل پیش‌پردازش تصاویر دیتاست NEU-DET با تکنیک‌هایی مانند عملیات مورفولوژیکی، اضافه کردن نویز گاوسی، تحلیل مؤلفه اصلی^۱ (PCA) برای کاهش بعد، و آستانه‌گذاری اوتسو برای جداسازی نقص‌ها از زمینه، و سپس آموزش مدل MobileNetV2 بر روی دیتاست جدید با تصاویر باینری و اندازه کوچک‌تر است [۲]. نتایج نشان‌دهنده دستیابی به دقت متوسط ۸۷٪ در طبقه‌بندی شش کلاس نقص است. این روش با کاهش اندازه داده‌ها و کاهش هزینه‌های محاسباتی، مزیت قابل‌توجهی در کاربردهای صنعتی ارائه می‌دهد؛ با این حال، وابستگی به تکنیک‌های پیش‌پردازش خاص ممکن است محدودیت‌هایی در تعمیم‌پذیری به دیتاست‌های متنوع‌تر ایجاد کند، و عملکرد آن در شرایط نوری پیچیده نیازمند ارزیابی بیشتر است. روش HCT-Det یک مدل مبتنی بر ترکیب سلسله‌مراتبی ویژگی‌های شبکه‌های کانولوشنی و ترنسفورمر برای تشخیص نقص سطح فولاد ارائه می‌دهد [۶]. این مدل از بلوک WSA-R^۲ برای به‌کارگیری توجه خودی پنجره‌ای با هدف کاهش هزینه محاسباتی، و از بلوک‌های رزیدوال موازی برای حفظ پیوستگی ویژگی‌های محلی بهره می‌گیرد. ویژگی‌های تولیدشده در سطوح

فولاد ارائه می‌دهد. نتایج تجربی حاکی از دقت خوب این روش نسبت به روش‌های مرسوم در شناسایی نقص‌های کوچک و دشوار است. این رویکرد می‌تواند به‌عنوان ابزاری مؤثر برای خودکارسازی بازرسی‌های صنعتی و پایه‌ای برای تحقیقات آتی در زمینه ارتقای تشخیص هوشمند نقص‌های سطحی مورد استفاده قرار گیرد. ساختار مقاله به‌صورت زیر تنظیم شده است: در بخش ۲ مروری بر روش‌های پیشین، مطالعات مرتبط و چالش‌های موجود در تشخیص نقص‌های سطحی فولاد بررسی شده‌اند. بخش ۳ روش پیشنهادی به معرفی تفصیلی چارچوب SSLA-Net و اجزای اصلی آن شامل مدل پایه تشخیص، استخراج ویژگی‌های معنایی، ماژول تلفیقی و توابع ضرر اختصاص دارد. در بخش ۴ نتایج و بحث عملکرد مدل پیشنهادی ارزیابی شده و تحلیل نتایج ارائه می‌شود. در بخش ۵ نتیجه‌گیری جمع‌بندی پژوهش ارائه شده است.

۲- مروری بر روش‌های پیشین

در سال‌های اخیر، استفاده از روش‌های نوین پردازش تصویر در تشخیص نقص‌های سطحی فولاد، پیشرفت‌های قابل‌توجهی را به همراه داشته است. معماری‌های پیشرفته باتکیه بر استخراج ویژگی‌های دقیق از تصاویر، توانسته‌اند عملکرد مناسبی در شناسایی و طبقه‌بندی انواع نقص‌ها ارائه دهند. با این حال، چالش‌هایی نظیر دشواری در تشخیص نقص‌های ریز و پیچیده، شباهت ظاهری میان کلاس‌های مختلف، و ضعف در نمایش معنایی نواحی تصویری، همچنان مانعی برای دستیابی به دقت بالا در شرایط صنعتی واقعی محسوب می‌شوند. علاوه بر این، بسیاری از روش‌های موجود در حفظ تعادل میان دقت و سرعت پردازش با محدودیت‌هایی مواجه‌اند. در پاسخ به این مسائل، رویکردهای جدید با تمرکز بر طراحی معماری‌های ترکیبی و چندمرحله‌ای، تلاش کرده‌اند تا با تلفیق اطلاعات ناحیه محور، ویژگی‌های معنایی و سازوکارهای توجه، دقت و پایداری تشخیص را بهبود بخشند. روش پیشنهادی این پژوهش نیز در همین راستا، باهدف ارتقا عملکرد در محیط‌های صنعتی و غلبه بر محدودیت‌های موجود طراحی شده است.

² Window-based Self-Attention / Residual Structure

¹ Principal Component Analysis

استفاده از YOLOv4 محل نقص‌ها را شناسایی کرده و سپس تصاویر برش خورده حاصل از مرحله اول توسط مدل U-Net مورد تقسیم‌بندی دقیق قرار گرفتند. این ترکیب هوشمندانه از تشخیص و تفکیک، به‌ویژه در شناسایی نقص‌های کوچک و پراکنده، موجب افزایش چشمگیر دقت گردید. با وجود این مزایا، عملکرد مدل در دسته‌هایی با نمونه‌های اندک، تحت تأثیر عدم تعادل داده‌ها قرار گرفت؛ موضوعی که لزوم بهره‌گیری از راهکارهای بهبود تعادل کلاس‌ها را برجسته می‌سازد.

یانگ و لیو روشی بهینه‌شده مبتنی بر معماری RetinaNet [۲۱] ارائه کرده‌اند که در این روش، با بهره‌گیری از کانولوشن‌های تغییرپذیر [۲۲] در ساختار اصلی شبکه ResNet طراحی شده است که امکان تنظیم پویا و تطبیقی شکل کرنل‌های کانولوشنی را فراهم می‌سازد [۸]. این ویژگی موجب استخراج دقیق‌تر ویژگی‌های نقص‌هایی با اشکال نامنظم و متنوع می‌گردد. علاوه بر آن، ماژول CA-BiFPN [۲۳] برای تلفیق ویژگی‌های چند مقیاسی معرفی شده است که با استفاده از مکانیسم توجه، تمرکز شبکه را بر نواحی مرتبط با نقص‌ها افزایش داده و در نتیجه دقت تشخیص را بهبود می‌بخشد. همچنین، تابع زیان IA-BCELoss [۲۴] با ترکیب مؤلفه‌های طبقه‌بندی و رگرسیون، موجب ارتقا کیفیت باکس‌های مرزی و بهبود عملکرد نهایی مدل می‌شود. این روش دقت بالایی دارد، با این حال، پیچیدگی محاسباتی ناشی از استفاده از کانولوشن‌های تغییرپذیر ممکن است در سناریوهای زمان واقعی با منابع محدود، چالش‌هایی ایجاد کند.

در حالت کلی بررسی مطالعات اخیر در حوزه تشخیص نقص‌های سطحی فولاد نشان‌دهنده پیشرفت چشمگیر در بهره‌گیری از الگوریتم‌های یادگیری عمیق است. مدل‌های مختلف با استفاده از تکنیک‌های تقسیم‌بندی معنایی، تشخیص شیء و مکانیسم‌های توجه، توانسته‌اند بخشی از چالش‌های موجود مانند تنوع شکل نقص‌ها، شباهت به پس‌زمینه و ابعاد کوچک آن‌ها را برطرف کنند. با این حال، محدودیت‌هایی نظیر دقت پایین در تشخیص نقص‌های ریز، پیچیدگی محاسباتی در کاربردهای زمان واقعی، عدم تعادل داده‌ها و ضعف در تعمیم‌پذیری در شرایط صنعتی متنوع همچنان

مختلف به صورت مقیاس متقاطع به انکودر منتقل می‌شوند و تعامل درون مقیاس و بین مقیاس به بهبود دقت در شناسایی اهداف کوچک و متنوع کمک می‌کند. همچنین، مکانیزم Soft IoU-Aware برای هم‌ترازی دقیق امتیازات طبقه‌بندی به کار گرفته شده است. نتایج تجربی نشان می‌دهند که این روش در مقایسه با مدل‌های پیشرفته موجود عملکرد بهتری دارد، هرچند پیچیدگی محاسباتی بالا، کاربرد آن را در شرایط زمان واقعی و تصاویر با وضوح بالا محدود می‌سازد.

روش RSTD-YOLOv7 با هدف ارتقا عملکرد مدل YOLOv7 [۱۲] در تشخیص نقص‌های سطحی فولاد طراحی شده است [۷]. در این معماری، ماژول RFBVGG^۱ و مکانیزم توجه SimAM [۱۳] در شبکه پس‌زمینه^۲ به کار گرفته شده‌اند تا از افت اطلاعات بافتی جلوگیری شود و انسجام ویژگی‌های محلی حفظ گردد. همچنین، در شبکه میانی^۳، ماژول STRVGG مبتنی بر ساختار ترنسفورمر Swin [۱۴] معرفی شده است که نقش مهمی در استخراج دقیق‌تر ویژگی‌های عمیق و کاهش ازدست‌رفتن اطلاعات دارد. در ادامه، سر^۴ تشخیص بهبودیافته DSDH نیز برای افزایش دقت طبقه‌بندی و تسریع همگرایی مدل مورد استفاده قرار گرفته است. این ترکیب ماژول‌ها موجب بهبود قابل توجه در شناسایی نقص‌های کوچک و با ویژگی‌های نامشخص شده و قابلیت کاربرد صنعتی این روش را افزایش داده است. با این حال، وابستگی به اجزای خاص و پیچیدگی محاسباتی بالا، ممکن است در برخی سناریوهای زمان واقعی و تصاویر با وضوح بالا محدودیت‌هایی ایجاد کند و نیاز به بهینه‌سازی بیشتر برای تعمیم‌پذیری در محیط‌های صنعتی متنوع‌تر احساس می‌شود.

در سال ۲۰۲۵، پژوهشی توسط اشرفی و همکاران ارائه شد که از مدل‌های پیشرفته نظیر U-Net [۱۵]، FCN-8 [۱۶] و FPN [۱۷] بهره گرفته شد که با شبکه‌های پس‌زمینه از پیش آموزش‌دیده‌ای مانند EfficientNet-B0 [۱۰] و ResNet34 [۱۸] ترکیب شده بودند. همچنین، مدل YOLOv4 [۱۹] برای تشخیص اولیه نقص‌ها بر روی مجموعه داده SEVERSTAL [۲۰] به کار گرفته شد [۵]. در ادامه، ساختاری دومرحله‌ای طراحی گردید که ابتدا با

³ Neck

⁴ Head

¹ Receptive Field Block with VGG-style backbone

² Backbone

۳-۱ مدل پایه تشخیص نقص

در این پژوهش، برای شناسایی نقص‌های سطح ورق‌های فولادی از معماری Faster R-CNN [۲۵] استفاده شده است. این معماری شامل دو بخش اصلی است: (۱) استخراج ویژگی‌ها، (۲) تولید نواحی پیشنهادی و سپس طبقه‌بندی دقیق این نواحی و تصحیح مختصات جعبه‌های محدودکننده^۱. این ساختار به افزایش دقت در شناسایی نواحی کوچک، نامنظم و بافت‌محور که در تصاویر صنعتی فولاد رایج است کمک می‌کند.

مرحله اول شامل استفاده از شبکه عمیق ResNet50 [۱۸] به‌عنوان شبکه پس‌زمینه است که پیش‌تر روی مجموعه‌داده ImageNet آموزش دیده است. ساختار این شبکه باعث استخراج ویژگی‌های غنی و چندسطحی می‌گردد که قادر به مدل‌سازی تغییرات ظریف مانند خطوط نازک و بافت‌های با وضوح کم است. لایه‌های ابتدایی این شبکه در فرایند آموزش ثابت نگه داشته شدند تا دانش عمومی حفظ شود و لایه‌های انتهایی برای تطبیق با داده‌های خاص نقص‌های سطح فولاد مجدداً آموزش داده شده‌اند که منجر به تسریع آموزش و افزایش دقت نهایی مدل می‌شود. انتخاب ResNet50 به‌عنوان شبکه پس‌زمینه کمک می‌کند تا مدل ضمن حفظ دقت بالا، توانایی استخراج ویژگی‌های پیچیده و خاص نقص‌های سطح فولاد را داشته باشد.

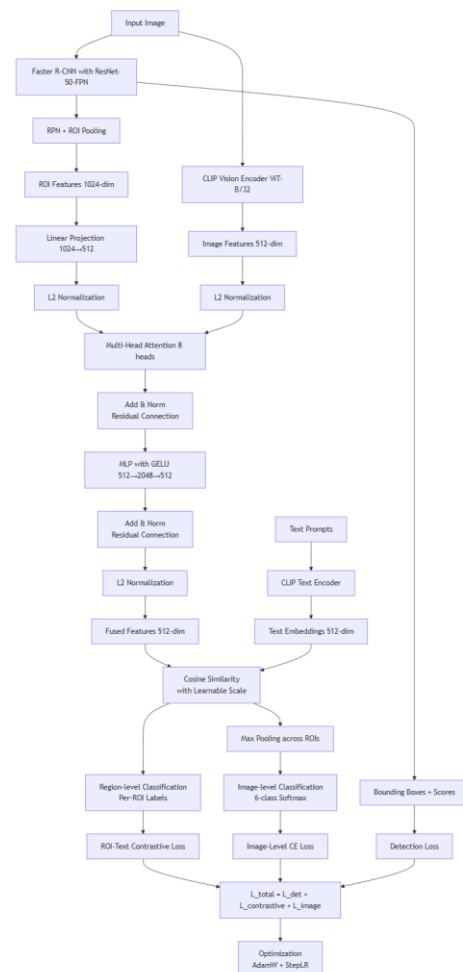
در مرحله دوم، شبکه پیشنهاد ناحیه^۲ (RPN) با الگوریتم‌های مبتنی بر لنگر نواحی مستعد وجود نقص را تولید کرده و سپس با استفاده از روش حذف بیشینه تداخل غیر حداکثری^۳ (NMS) جعبه‌های تکراری حذف می‌گردند تا تنها نامزدهای مهم حفظ شوند. شبکه پیشنهاد ناحیه از نه لنگر اولیه با ترکیب سه مقیاس و سه نسبت ابعاد بهره می‌گیرد و در هر تصویر حدود ۳۰۰ ناحیه پیشنهادی تولید می‌کند.

عملیات تجمیع ناحیه موردعلاقه بر روی این نواحی انجام شده و ویژگی‌های هر ناحیه استخراج می‌شود. طبقه‌بندی نهایی از طریق دولایه کاملاً متصل با ۱۰۲۴ و ۵۱۲ نورون و تابع فعال‌ساز ReLU انجام می‌شود. این رویکرد امکان بررسی دقیق و مستقل هر ناحیه

پابرجاست. این موارد، نیاز به توسعه راهکارهای نوآورانه برای دستیابی به تعادل میان دقت، سرعت و بهره‌وری محاسباتی را برجسته می‌سازد.

۳-۲ روش پیشنهادی

در این پژوهش، یک چارچوب ترکیبی برای تشخیص و طبقه‌بندی نقص‌های سطحی فولاد ارائه شده است که مبتنی بر تلفیق مدل‌های تشخیص ناحیه، و نمایش‌های معنایی است. مراحل روش پیشنهادی در شکل (۱) نشان‌داده شده است. این روش شامل مراحل زیر است:



شکل (۱): معماری روش پیشنهادی

³ Non-Maximum Suppression

¹ Bounding box

² Region Proposal Network

نهایت، برجستگی که بیشترین شباهت را دارد به آن ناحیه اختصاص داده می‌شود. به این ترتیب، به جای یک طبقه‌بندی صریح سستی، تخصیص برجسب بر اساس تطابق معنایی در فضای مشترک تصویر-متن انجام می‌شود. این قابلیت موجب می‌شود مدل فراتر از ویژگی‌های صرفاً بصری عمل کرده و درک مفهومی عمیق‌تری از نقص‌ها به دست آورد.

۳-۳ طراحی ماژول تلفیقی

برای دستیابی به نمایشی ترکیبی از نقص‌های سطحی فولاد، یک ماژول تلفیقی مبتنی بر مکانیزم توجه چندسری طراحی شده است. هدف این ماژول، ادغام اطلاعات مکانی حاصل از خروجی‌های ناحیه‌ای مدل Faster R-CNN با ویژگی‌های معنایی استخراج شده از CLIP است تا مدل بتواند تصمیم‌گیری دقیق‌تری بر اساس هر دو نوع داده انجام دهد.

در این ساختار، ویژگی‌های ناحیه‌ای با ابعاد ۱۰۲۴ ابتدا از طریق یک لایه خطی به فضای برداری ۵۱۲ نگاشته شده و سپس نرمال‌سازی می‌شوند. بردارهای تعبیه شده CLIP نیز به صورت جداگانه نرمال‌سازی می‌شوند. مکانیزم توجه چندسری با ۸ سر و بردارهای تعبیه شده با ابعاد ۵۱۲، برای مدل‌سازی روابط میان نواحی تصویری و ویژگی‌های معنایی به کار گرفته می‌شود.

خروجی توجه از طریق اتصالات باقیمانده با ورودی ترکیب شده و سپس توسط یک شبکه MLP پردازش می‌شود. این شبکه شامل دو لایه کاملاً متصل است. در گام نخست، ابعاد به ۲۰۴۸ افزایش یافته و سپس به ۵۱۲ کاهش می‌یابد. در این فرایند از تابع فعال‌ساز GELU و نرمال‌سازی لایه‌ای استفاده شده است. خروجی نهایی نیز بار دیگر تحت نرمال‌سازی قرار گرفته و با استفاده از یک پارامتر مقیاس‌پذیر قابل‌یادگیری تنظیم می‌شود. این طراحی منجر به تولید نمایش‌های تلفیق شده‌ای می‌شود که نه تنها موقعیت فضایی نقص را در نظر می‌گیرند، بلکه معنای مفهومی آن را نیز بازنمایی می‌کنند. به طور مشخص، مکانیزم توجه چندسری با مدل‌سازی هم‌زمان روابط میان ویژگی‌های ناحیه‌ای و تعبیه‌های معنایی، امکان هم‌ترازسازی دقیق‌تر بین نواحی تصویری و توصیف‌های متنی را فراهم می‌سازد.

را فراهم می‌آورد و دقت قابل‌توجهی نسبت به مدل‌های تک‌مرحله‌ای مانند YOLO و SSD در شناسایی نقص‌های ریز، پراکنده و پیچیده - به‌ویژه در شرایط نویز صنعتی و تنوع بافت بالا - ارائه می‌دهد.

۲-۳ استخراج ویژگی‌های معنایی

در این پژوهش برای ارتقا درک معنایی مدل و افزایش دقت در طبقه‌بندی نقص‌های سطحی، از مدل CLIP^۱ [۲۶] استفاده شده است. CLIP که توسط OpenAI معرفی شده، با بهره‌گیری از یادگیری میان تصویر و متن، قادر است نمایش‌های برداری مشترکی از داده‌های تصویری و زبانی تولید کند.

در این چارچوب، برای هر کلاس نقص سطحی فولاد، یک جمله توصیفی متنی طراحی می‌شود. به عنوان مثال، برای کلاس crazing جمله‌ای مانند "a close-up photo of crazing defect" در نظر گرفته شده است. این جملات توسط بخش متنی CLIP به بردارهای تعبیه‌شده زبانی تبدیل می‌شوند که نمایانگر معنای مفهومی هر کلاس هستند. به طور هم‌زمان، نواحی تصویری استخراج‌شده از مرحله تشخیص نیز توسط بخش تصویری CLIP پردازش شده و به بردارهای تصویری متناظر تبدیل می‌گردند.

مدل CLIP مورد استفاده در این پژوهش بر پایه معماری ترنسفورمر ViT-B/32 [۲۷] طراحی شده است. در این ساختار، تصویر ورودی به قطعات با ابعاد ۳۲×۳۲ تقسیم می‌شود و سپس با بهره‌گیری از مکانیزم توجه، نمایش‌های معنایی سطح بالا استخراج می‌گردد. برخلاف شبکه‌های کانولوشنی که عمدتاً بر ویژگی‌های محلی تمرکز دارند، معماری‌های ترنسفورمر مانند ViT قادرند ارتباطات بلند برد میان بخش‌های مختلف تصویر را مدل‌سازی کنند. قابلیت‌هایی که در شناسایی دقیق‌تر عیوب سطحی پیچیده و پراکنده نقش کلیدی ایفا می‌کند.

ویژگی منحصربه‌فرد CLIP، ایجاد یک فضای مشترک معنایی میان تصویر و متن است. در این فضا، هر ناحیه پیشنهادی از تصویر به صورت یک بردار تعبیه‌شده بازنمایی می‌شود و با تمام بردارهای متنی مربوط به کلاس‌ها مقایسه می‌گردد. شباهت محاسبه‌شده نشان می‌دهد که هر ناحیه به کدام کلاس متنی نزدیک‌تر است و در

¹ Contrastive Language-Image Pretraining

ترکیب وزن‌دار این توابع، مدل را به سمت یادگیری چندوجهی هدایت می‌کند و موجب افزایش دقت، کاهش ابهام در طبقه‌بندی، و بهبود عملکرد در شرایط صنعتی می‌شود. توابع ضرر به صورت زیر هستند:

تابع ضرر تشخیص شیء:

$$\ell_{\text{det}} = \ell_{\text{cls}} + \lambda_1 \ell_{\text{reg}} \quad (1)$$

$$\ell_{\text{cls}} = -\sum_i y_i \log(p_i)$$

$$\ell_{\text{reg}} = -\sum_i \text{Smooth}_{L1}(t_i - t_i^*)$$

تابع ضرر ناحیه - متن:

$$\ell_{\text{contrast}} = -\log \frac{\exp(\text{sim}(v_i, t_i) / \tau)}{\sum_j \exp(\text{sim}(v_i, t_j) / \tau)} \quad (2)$$

تابع ضرر سطح تصویر:

$$\ell_{\text{img}} = -\sum_k y_k \log(p_k) \quad (3)$$

تابع ضرر نهایی ترکیبی:

$$\ell_{\text{total}} = \alpha \ell_{\text{det}} + \beta \ell_{\text{contrast}} + \gamma \ell_{\text{img}} \quad (4)$$

در نهایت روش پیشنهادی با تلفیق مدل‌های تشخیص ناحیه، نمایش‌های معنایی و ماژول توجه، چارچوبی چندلایه برای شناسایی دقیق نقص‌های سطحی فولاد ارائه می‌دهد. طراحی توابع ضرر ترکیبی، یادگیری هم‌زمان ویژگی‌های مکانی، مفهومی و سطح تصویر را ممکن ساخته است. این ساختار، پایه‌ای منسجم برای ارزیابی عملکرد مدل در شرایط صنعتی واقعی فراهم می‌کند.

۴- نتایج و بحث

در این بخش، عملکرد معماری پیشنهادی در شناسایی نقص‌های سطحی فولاد مورد ارزیابی قرار گرفته است. ارزیابی‌ها بر اساس مجموعه‌داده صنعتی NEU-DET [۳] انجام شده و شامل تحلیل نتایج در دو سطح تشخیص ناحیه‌ای و طبقه‌بندی سطح تصویر است.

۴-۱ معرفی مجموعه‌داده

برای ارزیابی عملکرد معماری پیشنهادی، از مجموعه‌داده استاندارد NEU-DET استفاده شده است. این مجموعه‌داده شامل تصاویر خاکستری از سطح فولاد بوده و به طور خاص برای تشخیص نقص‌های سطحی صنعتی طراحی شده است. NEU-DET شامل

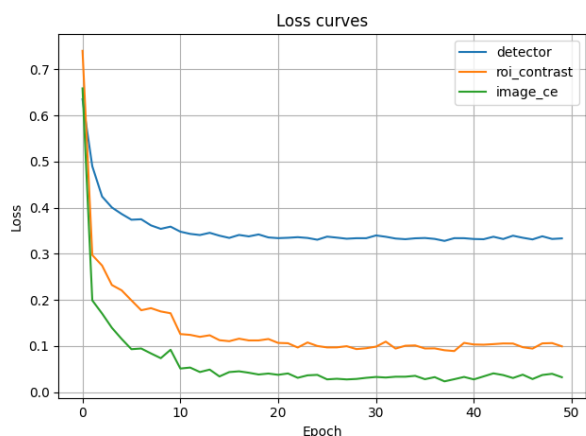
نحوه اتصال دقیق خروجی‌های مدل Faster R-CNN و ورودی‌های CLIP در شکل (۱) نمایش داده شده است. در این ساختار، ویژگی‌های استخراج شده از نواحی با ابعاد ۱۰۲۴ ابتدا از طریق یک لایه نگاشت خطی به فضای ۵۱۲-بعدی تبدیل و نرمال‌سازی می‌شوند. بردارهای متنی حاصل از CLIP نیز در همان فضای ۵۱۲-بعدی نرمال‌سازی شده و به ماژول توجه چندسری شامل ۸ سر توجه وارد می‌شوند. سپس خروجی توجه از طریق اتصالات باقیمانده و شبکه MLP با ساختار ۵۱۲→۲۰۴۸→۵۱۲ پردازش شده و ویژگی‌های ادغام شده نهایی تولید می‌گردند. بدین ترتیب، ارتباط میان نمایش‌های بصری و معنایی به صورت دقیق و هم‌تراز مدل‌سازی می‌شود.

در فرایند آموزش، وزن‌های مدل از پیش آموزش دیده CLIP در هر دو بخش تصویری و متنی ثابت نگه داشته شده‌اند تا از بروز ناپایداری گرادیان و انتقال بیش از حد جلوگیری شود. تنها پارامترهای لایه فرافکنی خطی، ماژول توجه چندسری و شبکه MLP به صورت قابل یادگیری تنظیم گردیده‌اند. ماژول توجه چندسری با ۸ سر و ابعاد برداری ۵۱۲ پیاده‌سازی شده است و شبکه MLP شامل دو لایه کاملاً متصل با ابعاد ۵۱۲→۲۰۴۸→۵۱۲ است. برای به‌روزرسانی پارامترها از بهینه‌ساز AdamW و کاهنده نرخ یادگیری StepLR استفاده شده است. همچنین، وزن‌های ترکیبی توابع زیان به صورت تجربی تعیین شده‌اند تا توازن مناسبی میان جنبه‌های مکانی، معنایی و کلی تصویر برقرار گردد.

۴-۳ طراحی توابع ضرر ترکیبی

برای آموزش دقیق معماری پیشنهادی، سه تابع ضرر طراحی شده‌اند که به صورت هم‌زمان جنبه‌های مکانی، معنایی و کلی تصویر را پوشش می‌دهند. نخست، تابع استاندارد مدل Faster R-CNN برای تشخیص و موقعیت‌یابی نقص‌ها حفظ شده است. سپس، یک تابع میان ویژگی‌های ناحیه‌ای و بردارهای تعبیه شده متنی CLIP تعریف شده تا مدل بتواند نواحی تصویری را با توصیف‌های معنایی هم‌راستا کند. در نهایت، از یک تابع آنتروپی متقاطع در سطح تصویر استفاده شده تا مدل کلاس غالب نقص را در کل تصویر تشخیص دهد.

آموزش برابر با ۴ و در مرحله اعتبارسنجی برابر با ۲ تنظیم شد. برای بررسی روند همگرایی مدل پیشنهادی، نمودار تغییرات سه تابع ضرر اصلی در طول مراحل آموزش رسم شده است. همان‌طور که در شکل (۳) مشاهده می‌شود، هر سه تابع ضرر در طول آموزش کاهش یکنواخت و پایداری داشته‌اند. تابع ضرر تشخیص ناحیه‌ای که با رنگ آبی نمایش داده شده، از مقدار اولیه حدود ۰/۶۳ به حدود ۰/۳۳ همگرا شده است، تابع تطابق ناحیه-متن با رنگ نارنجی از مقدار اولیه ۰/۷ به حدود ۰/۰۹ کاهش یافته، و تابع طبقه‌بندی سطح تصویر با رنگ سبز از ۰/۶۵ به حدود ۰/۰۳ رسیده است. این روند نشان‌دهنده بهینه‌سازی مؤثر هر بخش از مدل و تعامل مناسب میان وظایف چندگانه در معماری پیشنهادی است، به‌گونه‌ای که کاهش هم‌زمان و پایداری هر سه تابع ضرر بیانگر یادگیری هماهنگ و هم‌افزای مدل در سطوح مختلف تشخیص و طبقه‌بندی است.

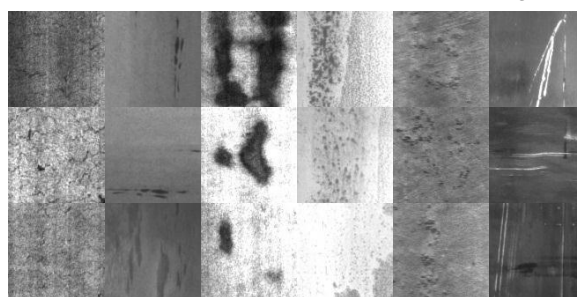


شکل (۳): نمودار تغییرات سه تابع ضرر اصلی مدل پیشنهادی

۴-۳ نتایج تشخیص ناحیه‌ای

برای ارزیابی عملکرد مدل در سطح تشخیص ناحیه‌ای، از معیار میانگین دقت متوسط در آستانه ۵۰ درصد هم‌پوشانی (mAP@50) استفاده شده است. این معیار نشان می‌دهد که مدل تا چه حد قادر است نواحی دارای نقص را به‌درستی شناسایی و طبقه‌بندی کند. این امر با در نظر گرفتن تطابق هندسی میان جعبه‌های

۶ کلاس نقص رایج در فولاد صنعتی است. این کلاس‌ها شامل پوسته‌های نورد شده^۱ (RS)، لکه‌ها^۲ (Pa)، ترک‌های سطحی^۳ (Cr)، سطح حفره‌دار^۴ (PS)، ناخالصی‌ها^۵ (In) و خراش‌ها^۶ (Sc) هستند. این مجموعه داده در مجموع شامل ۱۸۰۰ تصویر است که به‌صورت پیش‌فرض به ۱،۴۴۰ تصویر برای آموزش (Train) و ۳۶۰ تصویر برای تست (Test) تقسیم‌بندی شده‌اند. این تقسیم‌بندی توسط سازندگان مجموعه داده انجام شده و تنوع بالایی در ظاهر نقص‌ها را پوشش می‌دهد. در شکل (۲)، نمونه‌هایی از هر کلاس نشان داده شده است.



شکل (۲): تعدادی از تصاویر دیتاست NEU-DET در شش کلاس مختلف

۴-۲ آماده‌سازی داده‌ها، پیش‌پردازش و آموزش

برای افزایش تنوع داده‌ها و ارتقای توانایی مدل در تعمیم به شرایط واقعی، مجموعه‌ای از تبدیلات تصویری بر روی داده‌ها اعمال شده است. این تبدیلات شامل چرخش‌های تصادفی، تغییرات در روشنایی و کنتراست، اعمال نویز گوسی، و برش‌های موضعی بوده‌اند. هدف از این فرایند، شبیه‌سازی شرایط متغیر محیط صنعتی و افزایش مقاومت مدل در برابر تغییرات ظاهری بوده است، بدون آن که ساختار نقص‌ها دچار اختلال شود. روش پیشنهادی در ۵۰ ایپوک آموزش داده شده است. فرایند آموزش با استفاده از بهینه‌ساز AdamW انجام شد که برای بخش Faster R-CNN نرخ یادگیری برابر با $1e-5$ و برای بخش ماژول تلفیقی نرخ یادگیری برابر با $1e-4$ در نظر گرفته شد. مقدار ضرایب توابع ضرر یکسان در نظر گرفته شده است. مقدار اندازه دسته در مرحله

⁴ Pitted Surface

⁵ Inclusion

⁶ Scratches

¹ Rolled-in Scale

² Patches

³ Cracking

برتر قرار می‌دهد. نکته قابل توجه فاصله کمتر از ۱٪ (۰/۹٪) با روش برتر RSTD-YOLOv7 با مقدار ۷۹/۳٪ است که نشان‌دهنده عملکرد بسیار رقابتی روش پیشنهادی است. علاوه بر معیار $mAP@50$ ، به منظور ارزیابی دقت مدل در همپوشانی‌های سخت‌گیرانه‌تر، مقدار $mAP@75$ نیز محاسبه شد که مقدار ۴۸/۰۳٪ به دست آمد. این نتیجه نشان می‌دهد مدل توانایی حفظ دقت مکانی را حتی در شرایط همپوشانی بالا دارد که برای کاربردهای صنعتی اهمیت ویژه‌ای دارد.

برخلاف برخی روش‌ها که تنها در شناسایی دسته‌های خاص موفق عمل می‌کنند، روش پیشنهادی توازن دقت قابل قبولی از خود نشان می‌دهد. این سطح از پایداری و تعادل در دقت، پتانسیل بالای روش پیشنهادی را برای کاربرد در محیط‌های صنعتی واقعی - که تنوع و پیچیدگی نقص‌ها یک چالش اساسی است - به‌خوبی نشان می‌دهد.

جدول (۱): نتایج حاصل از تشخیص ناحیه‌ای در روش‌های متفاوت

روش	cr	in	pa	ps	rs	sc	mAP (%)
[۴]	۷/۴۱	۷۶/۳	۸۶/۳	۸۵/۱	۵۸/۱	۸۵/۶	۷۲/۴
[۲۸]	۸۷/۵	۷۵/۴	۷۶/۲	۸۱/۷	۶۵/۳	۷۲/۲	۷۶/۴
LAACNet [۲۹]	۴۷/۱	۸۸/۴	۹۲/۸	۷۶/۵	۵۹/۳	۹۲/۲	۷۶/۰
[vRSTD-YOLOv7]	۵۱/۲	۷۸/۱	۹۲/۷	۹۲/۴	۶۷/۲	۹۴/۰	۷۹/۳
[۳۰]	۵۹/۱	۸۶/۷	۹۵/۱	۸۵/۹	۶۵/۶	۷۲/۶	۷۷/۵
[۳۲]	۵۵/۱	۸۳/۰	۹۳/۶	۸۶/۱	۵۹/۷	۸۴/۲	۷۷/۰
[۳۳]	۵۱/۸	۸۱/۵	۹۰/۵	۹۰/۰	۵۹/۴	۸۶/۵	۷۶/۶
[۳۴]	-	-	-	-	-	-	۷۳/۸
[۳۶]	۶۵/۲	۸۲/۰	۸۷/۲	۷۴/۰	۷۶/۸	۷۴/۴	۷۶/۶
[۳۸]	-	-	-	-	-	-	۷۸/۲
روش پیشنهادی	۵۸/۹	۸۲/۴	۹۳/۸	۸۵/۲	۵۹/۳	۹۰/۳	۷۸/۴

مدل استخراج شده و با فرمول زیر محاسبه می‌شود:

$$confidence = \max_i P(class_i | box) \quad (6)$$

مطابق با شکل (۴) در تصویر اول، مدل با درجه اعتماد بالا (۰/۹۴) موفق به تشخیص دقیق یک ناحیه در کلاس in شد که تطابق هندسی مطلوبی با داده واقعی داشت. تصویر دوم شامل سه ناحیه pa بود که دو مورد با درجه اعتماد بالا (۰/۹۷ و ۰/۸۵) به‌درستی شناسایی شدند، درحالی‌که مورد سوم با درجه اعتماد پایین‌تر (۰/۶۱) نشان‌دهنده ابهام یا ضعف در تشخیص است. در تصویر

محدودکننده پیش‌بینی شده و جعبه‌های محدودکننده واقعی صورت می‌گیرد. این معیار با استفاده از معادله (۵) محاسبه می‌شود:

$$AP_c = \int_0^1 P_c(r) dr \quad (5)$$

$$mAP_{50} = \frac{1}{C} \sum_{c=1}^C AP_c$$

$$IoU = \frac{|B_{pred} \cap B_{gt}|}{|B_{pred} \cup B_{gt}|}$$

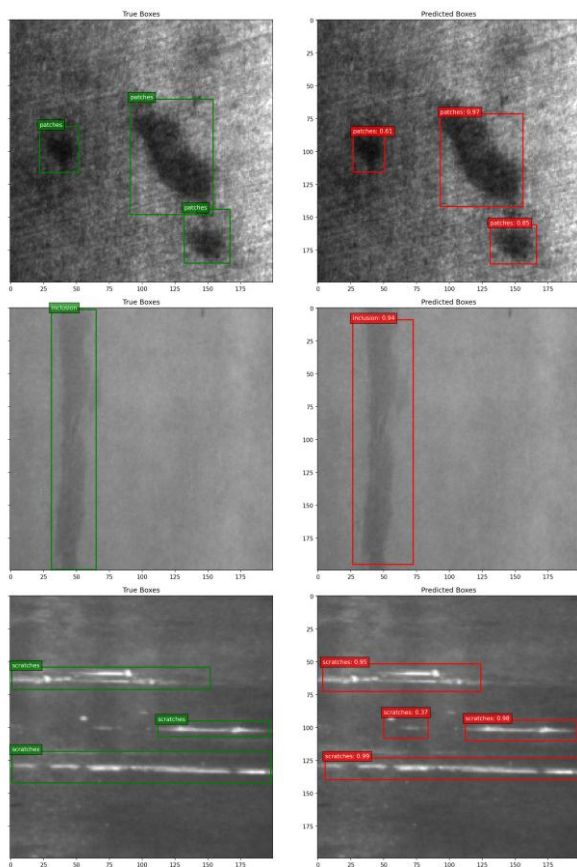
نتایج حاصل از ارزیابی روش پیشنهادی در مقایسه با چندین روش مطرح در جدول (۱) ارائه شده است. در این جدول، مقادیر $mAP@50$ برای هر یک از کلاس‌های شش‌گانه نقص و همچنین میانگین کلی آن گزارش شده است. مقدار $mAP@50$ کلی روش پیشنهادی ۷۸/۴٪ به دست آمده که آن را در جایگاه دومین روش

برای ارزیابی کیفی عملکرد مدل، مجموعه‌ای از تصاویر خروجی شامل جعبه‌های محدودکننده واقعی و پیش‌بینی شده در شکل (۴) نشان‌دهنده شده است. در این تصاویر، جعبه‌های قرمز نشان‌دهنده موقعیت واقعی نقص‌ها بوده و جعبه‌های سبز موقعیت‌های پیش‌بینی شده توسط مدل را نمایش می‌دهند. عددی که روی هر جعبه پیش‌بینی شده نوشته شده است، درجه اعتماد مدل به پیش‌بینی آن ناحیه است که میزان اطمینان مدل نسبت به تعلق ناحیه به یک کلاس خاص را نشان می‌دهد. این مقدار از خروجی لایه طبقه‌بندی

تعبیه‌سازی‌های معنایی CLIP تفاوت‌های مفهومی میان دسته‌های عیوب فولاد را به شکل مؤثری برجسته کرده و موجب افزایش فاصله کلاسی در فضای ویژگی شده‌اند. (۲) بهبود تشخیص نواحی کوچک و بافت‌محور:

ماژول توجه چندساز با مدل‌سازی روابط متقابل میان ROIها و متن، حساسیت مدل به جزئیات ناحیه‌ای را افزایش داده است. (۳) هم‌افزایی میان اطلاعات ناحیه‌ای و تصویر کامل:

استفاده از شباهت کسینوسی در سطح ROI و تجمیع پایدار آن در سطح تصویر، ریسک خطاهای محلی را کاهش داده و تصمیم نهایی را دقیق‌تر کرده است. این نتایج نشان می‌دهد SSLA-Net توانسته است چالش‌هایی مانند نقص‌های کوچک، بافت‌های شبیه به هم و تغییرات نوری را به‌صورت کارآمد مدیریت کند.



شکل (۴): مقایسه جعبه‌های محدودکننده واقعی (سبز) و

پیش‌بینی شده (قرمز) در تشخیص نقص سطحی

ماتریس برخورد مدل پیشنهادی که در شکل (۵) نشان داده شده، توزیع پیش‌بینی‌های مدل را در هر کلاس نقص نمایش می‌دهد. در

سوم، مدل چهار ناحیه SC را شناسایی کرده است که سه مورد با درجه اعتماد بسیار بالا (تا ۰/۹۹) تطابق خوبی با جعبه‌های واقعی داشتند، اما یک مورد با درجه اعتماد پایین (۰/۳۷) یک پیش‌بینی اشتباه است. این تصاویر نشان می‌دهند که مدل در تشخیص دقیق و طبقه‌بندی نواحی دارای نقص عملکردی قابل‌اتکا دارد، هرچند تنظیم آستانه می‌تواند به کاهش خطاهای جزئی کمک کند.

۴-۴ نتایج طبقه‌بندی تصویر

در این بخش، عملکرد مدل پیشنهادی در طبقه‌بندی انواع نقص‌های سطحی فولاد با معیارهای معتبر و استاندارد در حوزه تشخیص خودکار عیوب صنعتی مورد بررسی و مقایسه قرار گرفته است. معیارهای متداول ارزیابی شامل دقت (Accuracy)، صحت (Precision)، بازیابی (Recall) و معیار ترکیبی F1-score است که همه به‌صورت جامع در ارزیابی عملکرد مدل‌ها لحاظ گردیده‌اند. نتایج مقایسه‌ای استخراج شده از مقالات مرجع و معتبر در این حوزه، در قالب جدول ۲ ارائه شده‌اند تا امکان تحلیل و بررسی مقاومتی میان روش‌ها فراهم شود.

بر اساس نتایج ارائه شده در جدول، مدل پیشنهادی در تمامی معیارهای ارزیابی عملکرد به مقدار کامل (۱۰۰٪) دست یافته است. این امر گویای توانمندی بالای مدل در شناسایی دقیق و کامل شش نوع نقص سطحی استاندارد موجود در مجموعه داده معروف NEU-DET بوده و نشان از همگامی کامل پیش‌بینی‌های مدل با داده‌های واقعی دارد. دومین روش برتر، روش ارائه شده در مرجع [۳۵] است که دقت و عملکرد بسیار نزدیک به مدل پیشنهادی دارد. روش‌های قدیمی‌تر علی‌رغم پیشرفت‌های قبلی، عملکرد مناسبی نداشته و در معیارهای ذکر شده فاصله قابل‌ملاحظه‌ای با مدل‌های نوین دارند. ضروری است ذکر شود که برخی از روش‌های مقایسه‌ای تنها دقت کلی را گزارش کرده‌اند، بدون ارائه جزئیات کامل معیارهای طبقه‌بندی که این امر مقایسه دقیق و عمیق را دشوار ساخته است.

دستیابی به دقت ۱۰۰٪ در طبقه‌بندی سطح تصویر را می‌توان به سه عامل کلیدی نسبت داد:

(۱) تقویت قابلیت تفکیک مفهومی کلاس‌ها:

معماری **Faster R-CNN**، نمایش‌های معنایی مدل **CLIP** و ماژول توجه چندسری، عملکردی دقیق و چندلایه در شرایط صنعتی واقعی ارائه می‌دهد. استفاده هم‌زمان از اطلاعات مکانی و مفهومی، موجب افزایش دقت در شناسایی نواحی کوچک، پراکنده و بافت‌محور شده و امکان طبقه‌بندی معنایی دقیق‌تر را فراهم ساخته است. طراحی توابع ضرر ترکیبی نیز نقش مؤثری در هدایت مدل به سمت یادگیری هم‌زمان ویژگی‌های ناحیه‌ای، معنایی و کلی تصویر ایفا کرده است.

نتایج تجربی نشان داد که روش پیشنهادی در تشخیص ناحیه‌ای با دستیابی به مقدار $mAP@50$ برابر با $78/4\%$ عملکردی رقابتی و متوازن نسبت به روش‌های موجود دارد. همچنین، در طبقه‌بندی سطح تصویر، مدل موفق به دستیابی به دقت 100% شده است که بیانگر توانمندی آن در تشخیص کلاس غالب نقص در شرایط عملیاتی است. دقت بالای مدل به‌ویژه در طبقه‌بندی سطح تصویر، ناشی از هم‌افزایی میان نمایش‌های معنایی **CLIP** و ویژگی‌های مکانی استخراج‌شده از **Faster R-CNN** است. این ترکیب موجب افزایش تفکیک‌پذیری کلاس‌ها، بهبود تشخیص نواحی کوچک و کاهش خطاهای موضعی در تصمیم‌گیری نهایی شده است. این نتایج، همراه با ساختار چندلایه و قابلیت تعمیم‌پذیری مدل، نشان‌دهنده پتانسیل بالای روش پیشنهادی برای به‌کارگیری در سامانه‌های نظارت صنعتی و توسعه‌های آتی در حوزه تشخیص نقص‌های سطحی است.

درعین‌حال، لازم به ذکر است که مدل پیشنهادی دارای برخی محدودیت‌ها نیز هست. به دلیل استفاده از معماری‌های عمیق مانند **CLIP** و ماژول توجه چندسری، هزینه محاسباتی مدل نسبتاً بالا است و اجرای بلادرنگ آن در محیط‌های صنعتی نیازمند سخت‌افزارهای گرافیکی پیشرفته است. به‌عنوان مسیرهای آتی، می‌توان به بهینه‌سازی شبکه برای کاهش پیچیدگی محاسباتی و افزایش سرعت اجرا، و ارزیابی عملکرد مدل بر روی داده‌های صنعتی واقعی با شرایط نوری و بافتی متنوع‌تر اشاره کرد. این اقدامات می‌توانند موجب ارتقای کاربردپذیری و قابلیت تعمیم مدل در خطوط تولید واقعی شوند.

این ماتریس، سطرها نمایانگر کلاس‌های واقعی و ستون‌ها نمایانگر کلاس‌های پیش‌بینی شده هستند. نکته برجسته در ماتریس مدل پیشنهادی، صفر بودن کلیه مقادیر خارج از قطر اصلی است؛ این موضوع ضرورتاً بیانگر عدم بروز هیچ خطای طبقه‌بندی و پیش‌بینی دقیق و کامل تمامی نمونه‌ها است. اگرچه **SSLA-Net** عملکرد قدرتمندی نشان داده است؛ اما پیاده‌سازی ماژول توجه چندسری و پردازش ویژگی‌های **ROI** هزینه محاسباتی بیشتری نسبت به تشخیص‌گرهای تک‌مرحله‌ای دارد.

جدول (۲): نتایج حاصل از طبقه‌بندی تصویر در روش‌های متفاوت

روش	Precision	Recall	F1-score	Acc
[۲]	۰/۸۸	۰/۸۷	۰/۸۷	۰/۸۷
[۳۵]	۹۹/۳۳	۹۹/۳۳	۹۹/۱۷	۹۹/۴۴
[۳۵]	-	-	-	۸۹/۶۰
[۳۹]	-	-	-	۹۹/۰۵
[۴۰]	-	-	-	۹۵/۰۰
[۴۱]	۹۷/۸۱	۹۷/۷۷	۹۷/۷۸	۹۷/۷۷
[۴۲]	-	-	-	۹۷/۲۲
InceptionResNet-V2 [۴۳]	۱۰۰	۱۰۰	۱۰۰	۱۰۰
روش پیشنهادی	۱۰۰	۱۰۰	۱۰۰	۱۰۰

True	crazing	60	0	0	0	0	0
	inclusion	0	60	0	0	0	0
	patches	0	0	60	0	0	0
	pitted_surface	0	0	0	60	0	0
	rolled-in_scale	0	0	0	0	60	0
	scratches	0	0	0	0	0	60
		crazing	inclusion	patches	pitted_surface	rolled-in_scale	scratches
		Predicted					

شکل (۵): ماتریس برخورد مربوط به روش پیشنهادی

۵- نتیجه‌گیری

در این پژوهش، یک چارچوب ترکیبی برای تشخیص و طبقه‌بندی نقص‌های سطحی فولاد طراحی و ارزیابی شده است که با تلفیق

References

- [1] X. Wen, J. Shan, Y. He, and K. Song, "Steel Surface Defect Recognition: A Survey," *Coatings*, vol. 13, no. 1, art. no. 17, 2022, doi: 10.3390/coatings13010017.
- [2] O. C. Han and U. Kutbay, "Detection of Defects on Metal Surfaces Based on Deep Learning," *Applied Sciences*, vol. 15, no. 3, art. no. 1406, 2025, doi: 10.3390/app15031406.
- [3] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493–1504, Apr. 2020, doi: 10.1109/TIM.2019.2913672.
- [4] X. Lv, F. Duan, J. J. Jiang, X. Fu, and L. Gan, "Deep metallic surface defect detection: The new benchmark and detection network," *Sensors*, vol. 20, no. 6, p. 1562, Mar. 2020, doi: 10.3390/s20061562.
- [5] S. Ashrafi, M. Rahim, H. Li, and A. Zhang, "Steel surface defect detection and segmentation using deep neural networks," *Results in Engineering*, vol. 25, art. no. 103972, 2025, doi: 10.1016/j.rineng.2025.103972.
- [6] X. Chen, L. Wang, Y. Zhao, and J. Liu, "HCT-Det: A High-Accuracy End-to-End Model for Steel Defect Detection Based on Hierarchical CNN-Transformer Features," *Sensors*, vol. 25, no. 5, art. no. 1333, 2025, doi: 10.3390/s25051333.
- [7] H. Song, "RSTD-YOLOv7: A steel surface defect detection based on improved YOLOv7," *Scientific Reports*, vol. 15, art. no. 19649, 2025, doi: 10.1038/s41598-025-04811-w.
- [8] Z. Yang and Y. Liu, "A steel surface defect detection method based on improved RetinaNet," *Scientific Reports*, vol. 15, art. no. 6045, 2025, doi: 10.1038/s41598-025-88527-x.
- [9] I. U. Khan, N. Aslam, M. Aboulmour, A. Bashamakh, F. Alghool, N. Alsawayan, R. Alturaif, H. Gull, S. Z. Iqbal, and T. Hussain, "Deep Learning-Based Surface Defect Detection in Steel Products Using Convolutional Neural Networks ", *Mathematical Modelling of Engineering Problems*, vol. 11, no. 11, pp. 3006–3014, Nov. 2024, doi: 10.18280/mmep.111113.
- [10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.
- [11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.
- [12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, Jun. 2023, pp. 7464–7475. doi: 10.1109/CVPR52729.2023.00721.
- [13] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 11863–11874. [Online]. Available: <https://doi.org/10.48550/arXiv.2103.06264>
- [14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022. doi: 10.1109/ICCV48922.2021.00986
- [15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28
- [16] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440. [Online]. Available: <https://doi.org/10.48550/arXiv.1411.4038>
- [17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125. [Online]. Available: <https://doi.org/10.48550/arXiv.1612.03144>
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016, pp. 770–778. [Online]. Available: <https://arxiv.org/abs/1512.03385>
- [19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv:2004.10934*, 2020. [Online]. Available: <https://doi.org/10.48550/arXiv.2004.10934>
- [20] A. Grishin, B. Bardintsev, and Severstal, "Severstal: Steel Defect Detection," *Kaggle Competition Dataset*, 2019. [Online]. Available: <https://www.kaggle.com/c/severstal-steel-defect-detection>

- [21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)**, 2017, pp. 2980–2988. [Online]. Available: <https://doi.org/10.48550/arXiv.1708.02002>
- [22] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 764–773. [Online]. Available: <https://doi.org/10.48550/arXiv.1703.06211>
- [23] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13713–13722. [Online]. Available: <https://doi.org/10.48550/arXiv.2103.02907>
- [24] Q. Li, X. Jia, J. Zhou, L. Shen, and J. Duan, "Rediscovering BCE Loss for Uniform Classification," *arXiv preprint arXiv:2403.07289*, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2403.07289>
- [25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. [Online]. Available: <https://doi.org/10.1109/TPAMI.2016.2577031>
- [26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 8748–8763. [Online]. Available: <https://arxiv.org/abs/2103.00020>
- [27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [28] C. Zhang, X. Bai, and X. Zhang, "Metal plate surface defect detection method based on attention mechanism," *Comput. Integr. Manuf. Syst.*, vol. 29, no. 3, pp. xx–xx, 2023.
- [29] Z. L. Lv et al., "Lightweight adaptive activation convolution network based defect detection on polished metal surfaces," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108482, 2024. doi: 10.1016/j.engappai.2024.108482
- [30] M. Siddiqui, Y. Muhammad, and H. Ahn, "Faster metallic surface defect detection using deep learning with channel shuffling," *arXiv preprint, arXiv:2406.14582*, 2024.
- [31] B. Li et al., "NGD-YOLO: An improved real-time steel surface defect detection algorithm," *Electronics*, vol. 14, no. 14, p. 2859, 2025.
- [32] C. Li et al., "Steel surface defect detection method based on improved YOLOX," *IEEE Access*, vol. 12, pp. 37643–37652, 2024.
- [33] S. Zhou, Y. Cai, Z. Zhang, and J. Yin, "MESCD-DETR: An improved RT-DETR algorithm for steel surface defect detection," *Electronics*, vol. 14, no. 11, p. 2232, 2025.
- [34] Z. Li, X. Wei, M. Hassaballah, Y. Li, and X. Jiang, "A deep learning model for steel surface defect detection," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 885–897, 2024.
- [35] A. A. M. Ibrahim and J. R. Tapamo, "Transfer learning-based approach using new convolutional neural network classifier for steel surface defects classification," *Scientific African*, vol. 23, p. e02066, 2024.
- [36] T. Wi, M. Yang, S. Park, and J. Jeong, "D²-SPDM: Faster R-CNN-based defect detection and surface pixel defect mapping with label enhancement in steel manufacturing processes," **Applied Sciences**, vol. 14, no. 21, p. 9836, 2024.
- [37] S. Tian, P. Huang, H. Ma, J. Wang, X. Zhou, S. Zhang, J. Zhou, R. Huang, and Y. Li, "CASDD: Automatic surface defect detection using a complementary adversarial network," **IEEE Sensors Journal**, vol. 22, no. 20, pp. 19583–19595, Oct. 2022. doi: 10.1109/JSEN.2022.3202179
- [38] C. Chen, H. Lee, and M. Chen, "Steel surface defect detection method based on improved YOLOv9," *Scientific Reports*, vol. 15, no. 1, p. 25098, 2025. DOI: 10.1038/s41598-025-10647-1
- [39] L. Yi, G. Li, and M. Jiang, "An end-to-end steel strip surface defects recognition system based on convolutional neural networks," **Steel Research International**, vol. 88, no. 2, p. 1600068, Feb. 2017. doi: 10.1002/srin.201600068
- [40] A. A. M. Ibrahim and J. R. Tapamo, "Steel surface defect detection and classification using bag of visual words with BRISK," in **Proc. Congress on Smart Computing Technologies**, Singapore, Dec. 2022, pp. 235–246. doi: 10.1007/978-981-99-2468-4_18
- [41] M. Jeong, M. Yang, and J. Jeong, "Hybrid-DC: A hybrid framework using ResNet-50 and Vision Transformer for steel surface defect classification in the rolling process," *Electronics*, vol. 13, no. 22, p. 4467, Nov. 2024. DOI: 10.3390/electronics13224467.



- [42] S. T. K. Ayon, F. M. Siraj, and J. Uddin, "Steel Surface Defect Detection Using Learnable Memory Vision Transformer," *Computers, Materials & Continua*, vol. 82, no. 1, 2025.
- [43] A. Bouguettaya and H. Zarzour, "CNN-based hot-rolled steel strip surface defects classification: a comparative study between different pre-trained CNN models", Int. J. Adv. Manuf. Technol., vol. 132, no. 1, pp. 399–419, 2024.

SSLA-Net: Semantic–Spatial Learning Architecture for Robust Steel Surface Defect Detection

Masoumeh Rezaei^{1*}, Mansoureh Rezaei², Mehri Rajaei³

¹Assistant Professor, Department of Computer Engineering, University of Sistan and Baluchestan, Zahedan, Iran

²PhD in Artificial Intelligence, Computer Engineering Department, Yazd University, Yazd, Iran

³ Assistant Professor, Department of Information Technology Engineering, University of Sistan and Baluchestan, Zahedan, Iran

Article Information

Original Research Paper

Received:

2025 September 13

Accepted:

2025 November 03

Keywords:

Multimodal Learning, Surface Defect Detection, Multi-Head Attention, Semantic–Spatial Representation, Industrial Vision

Corresponding Author*:

mrezaei@ece.usb.ac.ir

Abstract

In the steel industry, accurate and rapid detection of surface defects plays a vital role in ensuring product quality and reducing waste. However, the visual similarity between texture patterns and the complexity of defective regions limit the effectiveness of conventional machine vision methods. In this study, a multimodal hybrid framework is proposed for the detection and classification of steel surface defects. The proposed model integrates the region-based detection capability of Faster R-CNN with the semantic representations of CLIP, connected through a multi-head attention module that jointly models the relationship between visual and conceptual features. This synergy enhances the accuracy in detecting small and scattered defects and improves class separability in the feature space. The architecture employs three complementary loss functions — region detection loss, region–text alignment loss, and image-level loss — to enable simultaneous learning of spatial and semantic information. Experimental results on the NEU-DET benchmark dataset demonstrate that the proposed model achieves 100% accuracy in image-level classification and 78.04% mAP@50 in region-level detection. These promising results indicate that CLIP-based multimodal learning can significantly enhance the accuracy and robustness of defect detection in real industrial environments.



: 10.22034/ABMIR.2025.23624.1166

E-ISSN: [2821-2037](#)

/© 2025. Published by Yazd University This is an open access article under the CC BY 4.0 License (<https://creativecommons.org/licenses/by/4.0/>).

