

توسعه مدلی ترکیبی از شبکه‌های EfficientNet و ترنسفورمر BiFormer برای بخش‌بندی دقیق و سریع تصاویر کولونوسکوپی

محمد امین نورائی یگانه^۱، حمید سلطانیان‌زاده^{۲،۳}

چکیده

بخش‌بندی تصاویر پزشکی نقش مهمی در تشخیص دقیق و برنامه‌ریزی درمان دارد و یکی از مسائل اساسی در حوزه بینایی ماشین محسوب می‌شود. دشواری در شناسایی مرزها و تفاوت‌های بافتی به‌ویژه در شناسایی پولیپ‌ها در تصاویر کولونوسکوپی، نیاز به روش‌هایی با دقت بالا را آشکار می‌سازد. در این پژوهش، رویکردی ترکیبی ارائه شده است که در آن بخش رمزگذار مسئول استخراج ویژگی‌های چندمقیاسی و عمیق و بخش رمزگشا متکی بر سازوکار توجه، جهت بازسازی ساختارهای دقیق تصویر به‌کار رفته است. در این راستا، EfficientNetV2 برای استخراج ویژگی‌ها و BiFormer Transformer برای بازنمایی و ترسیم مرزها استفاده شده است. برای ارزیابی کارایی روش پیشنهادی، آزمایش‌ها بر روی مجموعه داده‌های معتبر کولونوسکوپی انجام شده است. یافته‌های کمی و کیفی نشان می‌دهد که روش پیشنهادی در شاخص‌های Dice و IoU، نسبت به مدل‌های مطرح و پیشرفته اخیر، بهبود ایجاد کرده است. همچنین با کاهش تعداد پارامترها و بار محاسباتی، نشان‌دهنده توازن مناسب میان دقت و پیچیدگی محاسباتی است. نتایج کیفی هم توانایی مدل در نمایش شفاف و دقیق مرز پولیپ‌ها در تصاویر کولونوسکوپی را نشان می‌دهند. بنابراین، نتایج به‌دست‌آمده نشان می‌دهد که معماری پیشنهادی از پتانسیل مناسبی برای توسعه سامانه‌های کمکی بخش‌بندی تصاویر کولونوسکوپی برخوردار است.

کلید واژه‌ها

بخش‌بندی تصاویر پزشکی، بخش‌بندی پولیپ، تصاویر کولونوسکوپی، EfficientNetV2، BiFormer، ترنسفورمرهای بینایی

۱ مقدمه

تحلیل و پردازش تصاویر پزشکی طی دهه‌های گذشته به یکی از موضوعات مهم در مهندسی پزشکی و علوم رایانه تبدیل شده است. در این میان، بخش‌بندی تصاویر پزشکی یکی از مراحل اساسی پردازش تصویر محسوب می‌شود، زیرا نقش مهمی در تشخیص بیماری، ارزیابی وضعیت بیمار و برنامه‌ریزی درمان دارد. هدف اصلی بخش‌بندی، جداسازی دقیق نواحی موردنظر، مانند بافت‌ها، اندام‌ها یا ضایعات، از سایر قسمت‌های تصویر است [۱]. در مراحل اولیه، روش‌های سنتی پردازش تصویر مانند آستانه‌گذاری، لبه‌یابی و فیلترگذاری برای تشخیص و تفکیک نواحی مختلف در تصاویر پزشکی مورد استفاده قرار می‌گرفتند. اگرچه این روش‌ها در

این مقاله در ۲۱ خرداد ۱۴۰۵ دریافت و در ۵ مهر ۱۴۰۵ پذیرفته شد.

^۱دانشگاه تهران، دانشکده‌گان فنی، دانشکده برق و کامپیوتر

amin.nooraei@ut.ac.ir

^۲دانشگاه تهران، دانشکده‌گان فنی، دانشکده برق و کامپیوتر

hszadeh@ut.ac.ir

^۳آزمایشگاه تحلیل تصاویر پزشکی، موسسه پزشکی فورد آمریکا، دیترویت،

میشیگان

hsoltan1@hfhs.org

نویسنده مسئول: حمید سلطانیان‌زاده



برخی موقعیت ها نتایج قابل قبولی ارائه می دادند، اما در مواجهه با تصاویر پیچیده تر حاصل از روش های تصویربرداری مانند MRI^۱ و CT^۲، قادر به ارائه دقت و پایداری مورد انتظار نبودند [۲].

با گسترش یادگیری ماشین، روش هایی مانند شبکه های عصبی چندلایه و ماشین های بردار پشتیبان^۳ (SVM) به عنوان جایگزین روش های کلاسیک پیشنهاد شدند. این روش ها با تکیه بر استخراج ویژگی های دستی توانستند تا حدودی عملکرد بخش بندی را بهبود دهند. با این حال، وابستگی زیاد به طراحی دستی ویژگی ها، نیاز به دانش تخصصی و صرف زمان برای استخراج و انتخاب ویژگی ها، از مهم ترین محدودیت های این روش ها بود [۳].

تحول عمده در این حوزه با معرفی شبکه های عصبی پیچشی^۴ (CNN) رخ داد. این شبکه ها با قابلیت استخراج خودکار ویژگی های محلی و سلسله مراتبی از تصاویر، عملکرد قابل توجهی در بسیاری از مسائل بینایی ماشین ارائه کردند. معماری هایی مانند U-Net نیز به طور ویژه برای بخش بندی تصاویر پزشکی طراحی شدند و به دلیل ساختار رمزگذار-رمزگشا و استفاده از اتصالات میان بر، به یکی از مدل های مرجع در این حوزه تبدیل شدند [۴]. با وجود موفقیت گسترده CNN ها، این مدل ها در مدل سازی وابستگی های بلندبرد و روابط غیرمحلی در تصاویر با محدودیت مواجه بودند.

ظهور معماری ترنسفورمر و معرفی سازوکار خودتوجهی، امکان تحلیل روابط بلندمدت در داده ها را فراهم کرد و سپس به سرعت وارد حوزه بینایی ماشین و تصویربرداری پزشکی شد. در سال های اخیر، توجه پژوهشگران به سمت مدل های مبتنی بر ترنسفورمر جلب شده است؛ مدل هایی که در ابتدا برای پردازش زبان طبیعی توسعه یافته بودند. استفاده از سازوکار خودتوجهی در ترنسفورمرها موجب شد تا مدل ها قادر به شناسایی وابستگی های گسترده و الگوهای پیچیده در تصاویر باشند [۵]. با این حال، استفاده مستقیم و مستقل از ترنسفورمرها نیز با چالش هایی مانند نیاز به حجم بالای داده و هزینه محاسباتی زیاد همراه است.

بر این اساس، طراحی مدلی که بتواند مزایای شبکه های پیچشی و ترنسفورمر را به صورت مکمل به کار گیرد، به عنوان یک نیاز اساسی مطرح شده است. در ادامه و برای روشن تر شدن روند پژوهشی این حوزه، به بررسی سیر تکامل روش های بخش بندی تصاویر پزشکی می پردازیم. ابتدا مدل های مبتنی بر CNN که تمرکز اصلی آن ها بر استخراج ویژگی های محلی و ساختارهای سلسله مراتبی بوده است، مورد بررسی قرار خواهند گرفت. سپس به مدل های مبتنی بر ترنسفورمر و به ویژه معماری های مبتنی بر ترنسفورمر های بینایی (Transformer Vision) پرداخته می شود که هدف آن ها مدل سازی وابستگی های بلندبرد و روابط غیرمحلی در تصویر است.

در نهایت، رویکردهای ترکیبی که تلاش کرده اند نقاط قوت هر دو خانواده مدل را به صورت هم افزا ادغام کنند، مورد تحلیل قرار خواهند گرفت تا مشخص شود در تحقیقات پیشین چه پیشرفت هایی حاصل شده و چالش های اصلی همچنان در کدام بخش ها باقی مانده است.

با توسعه یادگیری عمیق، شبکه های پیچشی به سرعت در بخش بندی تصاویر پزشکی مورد توجه قرار گرفتند. یکی از شاخص ترین مدل ها در این حوزه، معماری U-Net است که به طور ویژه برای بخش بندی تصاویر زیست پزشکی طراحی شد. این مدل با بهره گیری از اتصالات میان بر، میان مسیرهای پایین نمونه سازی^۵ و بالا نمونه سازی^۶، انتقال مؤثر ویژگی ها و حفظ جزئیات را ممکن می سازد و سبب افزایش دقت بخش بندی می شود. به دلیل توانایی بالا در استخراج ویژگی های عمیق و بخش بندی پیکسل به پیکسل، U-Net به یک مدل مرجع و استاندارد در حوزه بخش بندی تصاویر پزشکی تبدیل شده است [۴].

در ادامه، مدل های دیگری برای بهبود کارایی و کاهش تعداد پارامترها معرفی شدند. برای مثال، EfficientNet با استفاده از بلوک های بهینه MBConv، ضمن حفظ دقت، از نظر تعداد پارامترها سبک تر و در محاسبات کارآمدتر است. قابلیت مقیاس پذیری این مدل در ابعاد مختلف (عمق، عرض و رزولوشن) امکان انطباق آن در کاربردهای گوناگون را فراهم کرده است [۶]. همچنین، U-Lite با بهره گیری از ساختار سبک و مبتنی بر U-Net، تنها با حدود یک میلیون پارامتر قابل آموزش، توانسته است عملکرد مناسبی در بخش بندی تصاویر پزشکی ارائه دهد و به عنوان یک مدل کارآمد و کم هزینه مطرح شود [۷].

توسعه های بعدی اغلب با هدف ارتقای انتقال ویژگی و مدل سازی مؤثرتر روابط در سطوح مختلف انجام شد. مدل هایی مانند U-Net++ با افزودن اتصالات متراکم و Dense-UNet با بهره گیری از بلوک های DenseNet در بخش رمزگذار^۷، موجب بهبود همگرایی و دقت بخش بندی شده اند [۸، ۹]. در سوی دیگر، مدل DeepLab با استفاده از Atrous Convolution دامنه دید لایه های پیچشی را گسترش داد تا وابستگی های بلندمدت در تصویر بهتر شناسایی شوند [۱۰].

در سال های اخیر، با معرفی ترنسفورمرهای بینایی، امکان مدل سازی وابستگی های بلندبرد و روابط غیرمحلی در تصویر فراهم شد. مدل ViT^۸ که توسط Dosovitskiy و همکاران در سال ۲۰۲۰ معرفی گردید، نشان داد که می توان تصاویر را به صورت پچ های غیرهمپوشان به توکن تبدیل و آن ها را مانند داده های زبانی پردازش کرد [۵]. نسخه بهینه شده آن، DeiT، با بهره گیری از روش های مناسب افزایش داده، توانست با داده های آموزشی

Downsampling^۵Upsampling^۶Encoder^۷Vision Transformer^۸Magnetic Resonance Imaging^۱Computed Tomography^۲Support vector machine^۳Convolutional Neural Network^۴

کمتر، عملکرد قابل توجهی به دست آورد [۱۱].

معماری Swin Transformer نیز با معرفی پنجره‌های شیفت‌دار^۱، امکان استخراج ویژگی‌های محلی و جهانی را به صورت سلسله‌مراتبی فراهم کرد و در بخش‌بندی پزشکی، شبکه‌هایی مانند Swin-UNet بر همین اساس توسعه یافتند [۱۲، ۱۳]. مدل TransUNet نیز با ترکیب ViT در نقش رمزگذار و U-Net در نقش رمزگشا^۲، توانست ویژگی‌های جهانی و محلی را به صورت مکمل به کار گیرد و نتایج برجسته‌ای ارائه دهد [۱۴]. در ادامه، مدل‌هایی مانند BRAU-Net++ [۱۵] و MISSFormer [۱۶] با بهبود ساختار توجه، دسترسی به اطلاعات چندمقیاسی و کاهش پیچیدگی محاسباتی را هدف قرار دادند.

با توجه به مزایای CNN در استخراج ویژگی‌های محلی و مزایای ترنسفورمر در مدل‌سازی اطلاعات جهانی، ترکیب این دو دسته معماری یکی از مسیرهای اصلی پژوهش‌های اخیر بوده است. MCTrans با جایگزینی اتصالات ساده در U-Net با سازوکارهای خودتوجهی و توجه بین‌کلاسی، توانسته است وابستگی‌های معنایی را دقیق‌تر مدل‌سازی کند [۱۷]. مدل EViT-UNet نیز CNN و ViT را در قالب بلوک‌های ترکیبی برای بهره‌برداری همزمان از وابستگی‌های محلی و جهانی به کار گرفته است [۱۸]. SCUNet++ نیز با استفاده از Swin Transformer در رمزگذار و رمزگشا و CNN در بخش Bottleneck، بهبود قابل توجهی در یادگیری چندمقیاسی به دست آورده است [۱۹]. مدل HiFormer نیز دو شاخه جداگانه مبتنی بر CNN و ViT را به صورت موازی اجرا کرده و سپس آن‌ها را از طریق ماژول DLF ترکیب می‌کند، که این امر ضمن کاهش تعداد پارامترها، موجب بهبود دقت بخش‌بندی شده است [۲۰]. در نهایت، مدل MedT با استفاده از توجه محوری^۳ و طراحی دو شاخه‌ای، قابلیت پردازش مؤثر تصویر در هر دو مقیاس کلی و جزئی را فراهم کرده و عملکرد مناسبی حتی بر روی مجموعه داده‌های کوچک ارائه داده است [۲۱].

با وجود موفقیت معماری‌های ترکیبی، نحوه به کارگیری مؤلفه‌های پیچشی و ترنسفورمری در روش‌های موجود متفاوت است. برای نمونه، Polyp-PVT از یک رمزگذار مبتنی بر Pyramid Vision Transformer و ماژول‌های CFM، CIM و SAM برای استخراج و ادغام ویژگی‌های چندسطحی استفاده می‌کند [۲۲]. در Swin Transformer نیز دو شاخه مبتنی بر Swin Transformer در رمزگذار به کار گرفته شده‌اند و ویژگی‌های چندمقیاسی توسط ماژول TIF ادغام می‌شوند؛ علاوه بر این، بلوک‌های Swin Transformer در بخش رمزگشا نیز حضور دارند [۲۳]. از سوی دیگر، BRAU-Net++ سازوکار Bi-Level Routing Attention را به عنوان جزء اصلی معماری در هر دو بخش رمزگذار و رمزگشا به کار می‌گیرد و اتصالات میان‌بر را نیز با ماژول توجه

کانالی و فضایی بازطراحی می‌کند [۱۵]. بنابراین، این روش‌ها عمدتاً بر استفاده گسترده از ساختارهای ترنسفورمری در رمزگذار، رمزگشا یا هر دو بخش شبکه تمرکز دارند.

در مقابل، در این پژوهش ساختار ترکیبی^۴ متفاوتی در نظر گرفته شده است. در بخش رمزگذار، از EfficientNetV2-S پیش‌آموزش دیده به عنوان یک استخراج‌کننده کارآمد ویژگی‌های محلی و چندمقیاسی استفاده می‌شود و ساختار این معماری پایه بدون افزودن بلوک‌های ترنسفورمری حفظ می‌شود. در بخش رمزگشا، BiFormer به صورت هدفمند در انتهای مراحل اصلی بازسازی و پس از ادغام ویژگی‌های رمزگذار و رمزگشا، بازتنظیم کانالی توسط ماژول SE و لایه پیچشی قرار می‌گیرد. هدف از این طراحی، بهره‌گیری از قابلیت انتقال یادگیری و استخراج کارآمد ویژگی‌های محلی در EfficientNetV2-S و در عین حال استفاده از BiFormer برای مدل‌سازی روابط غیرمحلی در مرحله بازسازی است، بدون آنکه سازوکار توجه به تمام بخش‌های شبکه گسترش یابد. از این رو، تفاوت اصلی روش پیشنهادی با روش‌های یادشده نه صرفاً در ترکیب CNN و Transformer، بلکه در تفکیک نقش این دو مؤلفه و جایگذاری هدفمند BiFormer تنها در مسیر رمزگشا است که در کنار رمزگذار پیش‌آموزش دیده EfficientNetV2-S و ماژول SE است؛ طراحی‌ای که با هدف دستیابی به توازن مناسب میان کیفیت بخش‌بندی و پیچیدگی محاسباتی ارائه شده است.

در بخش دوم، روش پیشنهادی و جزئیات معماری شبکه ترکیبی معرفی می‌شود. بخش سوم به تشریح مجموعه داده‌ها، تنظیمات پیاده‌سازی و آموزش، معیارهای ارزیابی و ارائه نتایج اختصاص دارد. در این بخش، عملکرد مدل پیشنهادی روی دو مجموعه داده معتبر کولونوسکوپیی ارزیابی و با روش‌های موجود مقایسه می‌شود؛ همچنین ارزیابی کیفی، تحلیل عدم قطعیت و کالیبراسیون، مطالعه حذف مؤلفه‌ها، ارزیابی تعمیم‌پذیری بین مجموعه داده‌ای و بررسی تأثیر محل قرارگیری BiFormer ارائه می‌شود. در بخش چهارم، نتایج به صورت تحلیلی مورد بحث قرار می‌گیرند. بخش پنجم به محدودیت‌های پژوهش و مسیرهای پیشنهادی برای مطالعات آینده اختصاص دارد و در نهایت، در بخش ششم جمع‌بندی و نتیجه‌گیری نهایی پژوهش ارائه می‌شود.

۲ طراحی و مدل‌سازی شبکه پیشنهادی

همان‌گونه که در بخش قبل بیان شد، بخش‌بندی تصاویر پزشکی یکی از چالش‌های مهم در حوزه پردازش تصویر و یادگیری عمیق به شمار می‌رود و نیازمند روش‌هایی است که بتوانند به طور همزمان ویژگی‌های محلی و جزئی تصویر را استخراج کرده و همچنین وابستگی‌های بلندمدت و روابط پیچیده موجود در ساختار داده‌ها

^۴ Hybrid Backbone

^۱ Shifted Windows
^۲ Decoder
^۳ Axial Attention

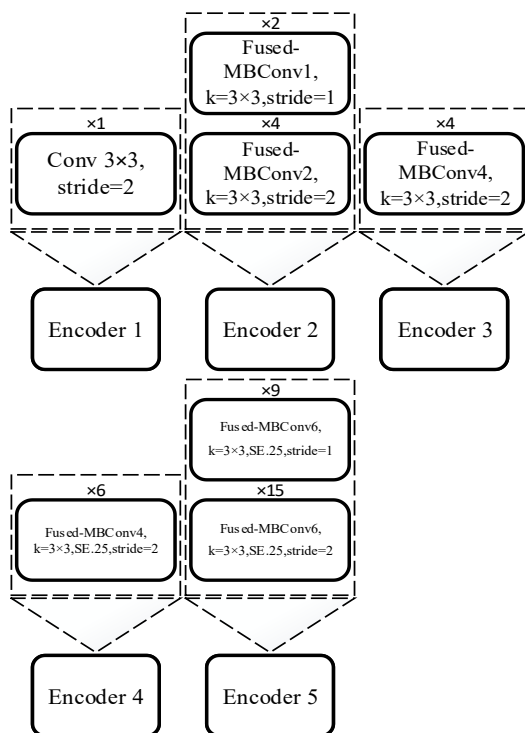
بازسازی مرحله ای در رمزگشا، معماری پیشنهادی قادر است اطلاعات مکانی و معنایی تصویر را استخراج و بازتولید کند.

۲-۲ بخش رمزگذار (استخراج ویژگی های محلی)

۲-۲-۱ ساختار کلی رمزگذار

در بخش رمزگذار مدل، از معماری EfficientNetV2-S استفاده شده است. این معماری به دلیل ساختار سبک، مقیاس پذیری و کارایی محاسباتی بالا، گزینه ای مناسب برای مسائل تصویری با حجم داده زیاد از جمله تصاویر پزشکی محسوب می شود. شکل شماره (۱) ساختار هر طبقه از انکدر را نمایش می دهد. در هر طبقه از تعدادی بلوک MBConv و Fused-MBConv استفاده شده است.

نسخه به کاررفته در پژوهش tf_efficientnetv2_s.in21k است که بر روی مجموعه داده ImageNet-21K از پیش آموزش دیده و به واسطه قابلیت انتقال یادگیری^۵، امکان دستیابی به دقت بالا حتی با داده های محدود پزشکی را فراهم می سازد.



شکل شماره (۱): نمایش ساختار هر طبقه از رمزگذار مدل پیشنهادی

۲-۲-۲ استفاده از معماری پایه EfficientNetV2

EfficientNetV2-S از چندین نوع بلوک پایه تشکیل شده است که هر کدام نقش متفاوتی در استخراج ویژگی ها دارند:

را مدل سازی کنند. بر این اساس، در این بخش روش پیشنهادی تحقیق تشریح می شود.

در این پژوهش، یک مدل ترکیبی مبتنی بر شبکه های عصبی پیچشی و ترنسفورمرهای بینایی طراحی شده است که هدف اصلی آن، افزایش دقت و پایداری در فرآیند بخش بندی تصاویر کولونوسکوپی است. بدین منظور، ابتدا ساختار کلی مدل ارائه می شود و سپس جزئیات معماری، نحوه انتخاب ماژول ها و دلیل استفاده از هر بخش به طور دقیق توضیح داده خواهد شد. در ادامه نیز نحوه ترکیب و تعامل این ماژول ها در هر مرحله از مدل مورد بررسی قرار می گیرد.

۲-۱ ساختار کلی سیستم پیشنهادی

در سیستم پیشنهادی برای بخش بندی تصاویر کولونوسکوپی، بخش رمزگذار بر پایه معماری EfficientNetV2 [۲۴] طراحی شده و بخش رمزگشا با بهره گیری از بلوک های ترکیبی شامل ترنسفورمر، لایه های پیچشی و ماژول های Squeeze-and-Excitation(SE) توسعه یافته است. انتظار می رود ترکیب این دو بخش، مدلی با دقت بهتر و پیچیدگی محاسباتی بهینه ایجاد کند.

همان گونه که در شکل شماره (۲) نمایش داده شده است، سیستم پیشنهادی از چند بخش اصلی تشکیل شده است. ابتدا تصویر ورودی با ابعاد $C \times W \times H$ وارد لایه ابتدایی^۱ می شود. این لایه وظیفه تنظیم کانال های ورودی و آماده سازی داده برای ورود به رمزگذار را بر عهده دارد. پس از آن، ویژگی ها وارد بخش رمزگذار می شوند که از پنج بلوک اصلی تشکیل شده است. این بلوک ها با استفاده از واحدهای MBConv و Fused-MBConv به تدریج ابعاد مکانی را کاهش داده و در مقابل عمق ویژگی ها را افزایش می دهند. بدین ترتیب، اطلاعات از ویژگی های سطح پایین به ویژگی های عمیق تر و انتزاعی تر منتقل می شود.

در ادامه، بخش رمزگشا قرار دارد که شامل چهار دیکدر^۲ اصلی و یک دیکدر نهایی است. هر دیکدر اصلی با دریافت ویژگی های متناظر از رمزگذار و بهره گیری از سازوکار اتصال میان گذر^۳، به بازسازی تدریجی نقشه های ویژگی می پردازد. در این بخش، علاوه بر لایه های پیچشی و کانولوشن انتقالی^۴، از ماژول BiFormer Transformer [۲۵] و ماژول SE نیز جهت تقویت وابستگی های فضایی و کانالی استفاده شده است. دیکدر نهایی (Decoder0) با استفاده از چند لایه پیچشی و انتقالی، نقشه ویژگی بازسازی شده را به ابعاد اولیه تصویر بازمی گرداند.

در پایان، یک لایه پیچشی 1×1 نقشه ویژگی نهایی را به فضای خروجی نگاشت کرده و نقشه بخش بندی با ابعاد $Classes \times W \times H$ تولید می شود. با افزایش تدریجی عمق ویژگی ها در رمزگذار و

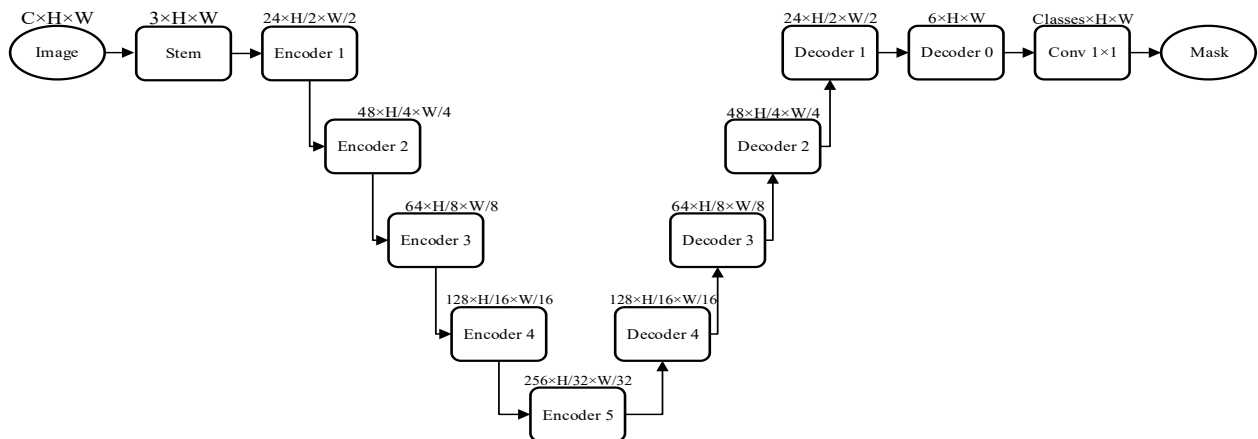
^۴ ConvTranspose2d

^۵ Transfer Learning

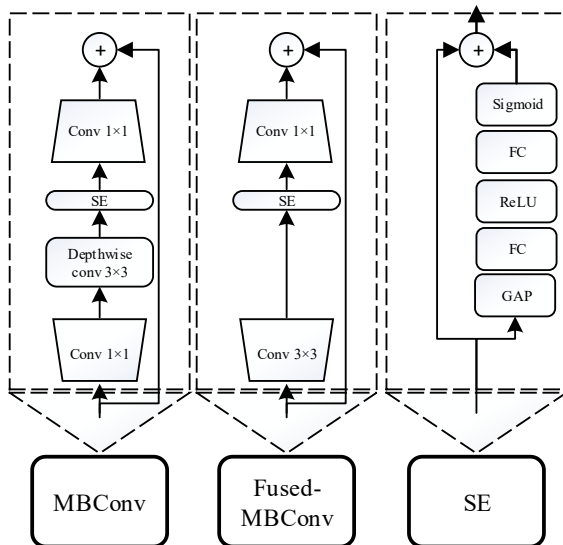
^۱ Stem

^۲ Decoder

^۳ Skip Connection



شکل شماره (۲): معماری کلی مدل پیشنهادی



شکل شماره (۳): نمایش ساختار سه بلوک اصلی تشکیل دهنده معماری EfficientNetV2 (رمزگذار)

۲-۳ بخش رمزگشا

۲-۳-۱ ساختار کلی رمزگشا

رمزگشا با الهام از ساختار U-Net طراحی شده است. این بخش به کمک مسیرهای اتصال میان‌گذر، ویژگی‌های سطح پایین و حاوی اطلاعات مکانی دقیق را از رمزگذار دریافت کرده و در فرآیند بازسازی لحاظ می‌کند. به این ترتیب، رمزگشا قادر است ترکیب متعادلی از ویژگی‌های محلی و معنایی را برای تولید نقشه بخش‌بندی نهایی فراهم سازد.

ساختار رمزگشا شامل چهار مرحله اصلی Decoder1 تا Decoder4 و یک مرحله نهایی Decoder0 است. در هر مرحله، ابتدا خروجی مرحله قبلی Decoder با ویژگی متناظر از Encoder ترکیب می‌شود. این ترکیب با عملیات Crop & Concat انجام می‌گیرد تا ابعاد تراز شده و ویژگی‌ها به صورت مؤثر ادغام شوند.

پس از الحاق ویژگی‌ها، مطابق شکل شماره (۴)، از ماژول SE جهت بازتخصیص وزن کانال‌ها و تقویت کانال‌های مهم در هر مرحله استفاده می‌شود. حضور ماژول SE در هر لایه رمزگشا باعث می‌شود مدل هنگام بازسازی، به جای

بلوک MBConv: یک بلوک پیچشی توسعه‌یافته که با استفاده از لایه پیچشی عمقی و لایه‌های بهینه‌سازی‌شده، موجب کاهش تعداد پارامترها و افزایش بازدهی محاسباتی می‌شود.

بلوک Fused-MBConv: نسخه‌ی بهبودیافته MBConv که در آن بخشی از عملیات پیچشی به صورت یکپارچه‌تر انجام شده و در نتیجه سرعت یادگیری و همگرایی مدل افزایش می‌یابد.

بلوک Squeeze-and-Excitation (SE): ماژولی که با تخصیص وزن پویا به کانال‌های ویژگی، اهمیت نسبی کانال‌ها را تنظیم کرده و سبب استخراج هدفمندتر و معنایی‌تر ویژگی‌ها می‌شود.

برای استخراج ویژگی‌ها از تصویر، ابتدا ورودی از طریق لایه‌ی Stem عبور کرده و سپس وارد زنجیره‌ای از پنج مرحله رمزگذاری می‌شود. در هر مرحله، یک یا چند بلوک MBConv یا Fused-MBConv به کار گرفته شده است. با پیشروی از رمزگذارهای ابتدایی به رمزگذارهای عمیق‌تر:

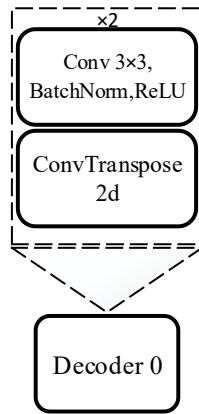
- ابعاد فضایی ویژگی‌ها ($H \times W$) به تدریج کاهش می‌یابد،
- در مقابل تعداد کانال‌ها افزایش پیدا می‌کند،
- و ویژگی‌ها از سطوح محلی و کم‌عمق به نمایش‌های معنایی و انتزاعی‌تر تبدیل می‌شوند.

شکل (۳) اجزای پایه رمزگذار EfficientNetV2-S را نشان می‌دهد. این بلوک‌ها خودِ نوآوری پژوهش حاضر نیستند؛ بلکه نقش آن‌ها فراهم کردن استخراج ویژگی‌های محلی و چندمقیاسی با هزینه محاسباتی مناسب است. نوآوری اصلی در نحوه استفاده از این ویژگی‌ها در رمزگشا و ترکیب هدفمند آن‌ها با ماژول‌های SE و BiFormer قرار دارد.

به این ترتیب، بخش رمزگذار قادر است هم ساختارهای بافتی ریز و هم الگوهای عمیق و مفهومی تصویر را به صورت سلسله‌مراتبی و مرحله‌به‌مرحله استخراج کند؛ امری که در بخش‌بندی دقیق تصاویر پزشکی نقش اساسی دارد.

کانال ها را به تعداد کلاس های بخش بندی کاهش داده و ماسک نهایی تولید می شود.

شکل (۵) مرحله نهایی بازسازی را نشان می دهد که وظیفه آن بازگرداندن نقشه ویژگی به ابعاد ورودی و تولید ماسک نهایی است. این بخش عمدتاً نقش تکمیل کننده دارد و نوآوری اصلی معماری در دیکدرهای ۱ تا ۴ و جایگذاری هدفمند BiFormer در آن ها متمرکز است.



شکل شماره (۵): ساختار اجزاء آخرین طبقه دیکدر در مدل پیشنهادی

۲-۳-۲ استفاده از ترنسفورمر بینایی BiFormer

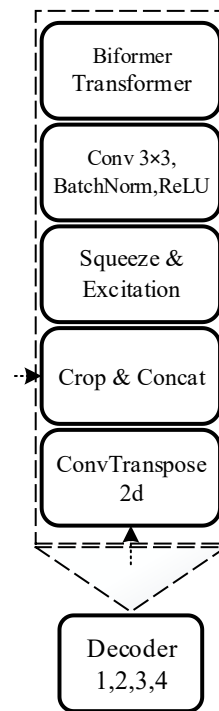
در مدل پیشنهادی، از ماژول Transformer BiFormer به عنوان بخش توجه^۲ استفاده شده است. BiFormer که اخیراً معرفی شده، سازوکاری با عنوان Bi-Level Routing Attention (BRA) ارائه می دهد [۲۵]. در این روش، ابتدا تصویر از منظر اهمیت و موقعیت به سطوح مختلف تقسیم شده و سپس ارتباط میان این سطوح به صورت تطبیقی برقرار می شود. این رویکرد موجب افزایش دقت در تشخیص نواحی کلیدی و در عین حال کاهش بار محاسباتی نسبت به روش های خود توجهی^۲ استاندارد می گردد.

نکته مهم آن است که BiFormer فقط در بخش Decoder معماری به کار گرفته شده است. دلیل این انتخاب آن است که در مرحله بازسازی، مدل نیازمند ترکیب ویژگی های محلی (استخراج شده توسط CNN) با درک روابط سراسری تصویر است. بدین ترتیب، BiFormer با تقویت ارتباط میان پیکسل های دور از یکدیگر، به بهبود دقت مرزبندی و افزایش انسجام ماسک خروجی کمک می کند. در مقابل، بخش رمزگذار که نقش اصلی آن استخراج ویژگی های محلی است، با تکیه بر ساختار EfficientNetV2 به طور کامل کفایت می کند و نیازی به استفاده از سازوکارهای توجه پیچیده در این مرحله وجود ندارد.

شکل (۶) سازوکار توجه مورد استفاده در مدل پیشنهادی را نشان می دهد. اهمیت این ماژول در پژوهش حاضر نه در تغییر ساختار داخلی BiFormer، بلکه در نحوه جایگذاری آن در انتهای مراحل اصلی رمزگشا است؛ به طوری که مدل پس از بازیابی و پالایش ویژگی های محلی، از Bi-Level Routing Attention برای

صرفاً افزایش ابعاد، بر کانال های اطلاعاتی مفید تمرکز کند. سپس بلوک Conv 3x3 به همراه Batch Normalization و تابع فعال ساز ReLU برای پالایش ویژگی ها و بهبود پایداری آموزش به کار گرفته می شود.

ماژول Transformer BiFormer جهت مدل سازی وابستگی های بلندبرد و روابط معنایی در نقشه های ویژگی در مرحله آخر در هر طبقه دیکدر استفاده می شود. این ماژول باعث می شود مدل علاوه بر حفظ جزئیات مکانی، ساختارهای گسترده تر و الگوهای تشخیصی عمیق تر را نیز شناسایی کند [۲۵].



شکل شماره (۴): اجزاء اصلی دیکدرهای چهارگانه شامل ماژول BiFormer

به این ترتیب، در هر بلوک دیکدر، ابتدا ابعاد تصویر به صورت کنترل شده افزایش یافته، سپس ویژگی های رمزگذار ادغام شده و با کمک SE و BiFormer اصلاح و تقویت می گردند. این فرایند در تمام مراحل بازسازی تکرار شده و منجر به بازتولید ساختارهای دقیق بافتی و مرزبندی های قابل اعتماد می شود.

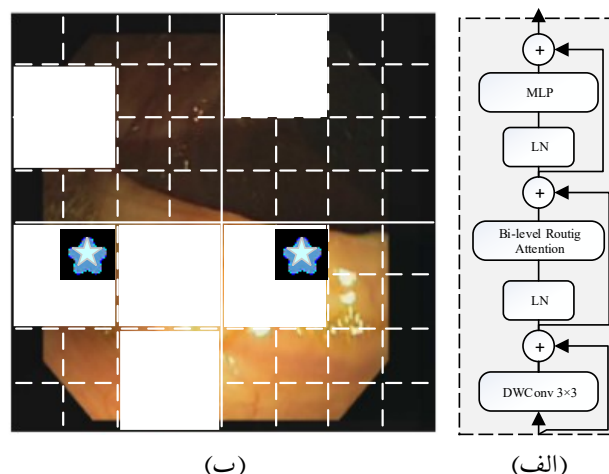
همان طور که در شکل (۴) دیده می شود، نوآوری معماری پیشنهادی در ترتیب هدفمند اجزای هر دیکدر است؛ به گونه ای که پس از افزایش ابعاد و ادغام ویژگی های Encoder و Decoder، ابتدا بازتنظیم کانالی توسط SE و پالایش پیشگی انجام شده و سپس BiFormer در انتهای هر مرحله برای مدل سازی روابط غیرمحلی به کار گرفته می شود. این ترتیب با هدف حفظ جزئیات محلی و اعمال توجه روی ویژگی های پالایش شده طراحی شده است.

در مرحله پایانی، Decoder0 با دریافت خروجی آخرین رمزگشا و اعمال Upsampling، ابعاد نقشه ویژگی را به اندازه اصلی (H×W) بازمی گرداند. در نهایت، یک لایه پیشگی ۱×۱ تعداد

^۲ Self-Attention

^۱ Attention

تقویت وابستگی‌های غیرمحلی و ارتباط میان نواحی دوردست تصویر استفاده می‌کند.



شکل شماره (۶): (الف) نمایش معماری بلوک BiFormer و (ب) سازوکار Bi-Level Routing Attention، تصویر به پنجره‌های محلی و انتخاب مسیرهای توجهی نمایش داده شده است.

از سوی دیگر، جایگاه قرارگیری BiFormer در ساختار Decoder نیز هدفمند انتخاب شده است. این ماژول در انتهای هر طبقه از Decoder قرار داده شده است. در مراحل ابتدایی بازسازی، ترکیب ویژگی‌های Encoder و Decoder و اعمال عملیات‌های پیچشی برای بازیابی جزئیات محلی کافی است و افزودن Self-Attention در این مرحله تنها موجب افزایش هزینه محاسباتی بدون بهبود مؤثر خواهد شد. اما در لایه پایانی هر Decoder، ویژگی‌های محلی تا حد زیادی بازسازی شده‌اند و نقشه ویژگی به سطحی از انتزاع رسیده است که مدلسازی وابستگی‌های غیرمحلی ارزش اطلاعاتی ایجاد می‌کند. بنابراین، قرارگیری BiFormer در این مرحله موجب ترکیب هم‌زمان اطلاعات محلی بهتر و روابط معنایی قوی‌تر می‌شود.

مزیت دیگر BiFormer نسبت به ترنسفورمرهای کلاسیک مانند ViT، کاهش پیچیدگی محاسباتی آن است. در حالی که Self-Attention استاندارد دارای پیچیدگی $O(N^2)$ نسبت به تعداد پچ‌ها است، BiFormer با استفاده از سازوکار BRA این مقدار را به $O(N \log N)$ کاهش می‌دهد. این ویژگی امکان استفاده از آن را در تصاویر با رزولوشن بالا و کاربردهای حساس بالینی فراهم می‌سازد.

در مجموع، به‌کارگیری BiFormer در انتهای هر مرحله از Decoder ترکیب بهینه‌ای از دو دیدگاه مکمل را ایجاد می‌کند: بازیابی دقیق ساختارهای محلی توسط CNN و مدلسازی روابط بلندبرد و معنایی توسط Self-Attention. این ترکیب، دقت، انسجام و قابلیت اعتماد نتایج بخش‌بندی را در تصاویر پزشکی به‌طور معناداری افزایش می‌دهد.

۳ نتایج و تحلیل

در این پژوهش، عملکرد مدل پیشنهادی بر روی دو مجموعه از مجموعه داده‌های معتبر تصاویر کولونوسکوپی و پرکاربرد در حوزه ی بخش‌بندی تصاویر پزشکی مورد ارزیابی قرار گرفته است. هدف از این ارزیابی، سنجش دقت، پایداری و قابلیت تعمیم‌پذیری مدل در مواجهه با داده‌های واقعی و متنوع است؛ داده‌هایی که در پژوهش‌های پیشین نیز به‌عنوان معیارهای استاندارد برای مقایسه‌ی روش‌های نوین مورد استفاده قرار گرفته‌اند.

افزون بر معرفی این دو مجموعه داده، شیوه‌ی آموزش مدل، تنظیمات پیاده‌سازی، معیارهای ارزیابی و نحوه‌ی آزمون بر روی هر یک از مجموعه داده‌ها تشریح می‌شود تا چارچوب ارزیابی مدل به‌طور کامل تبیین گردد. سپس نتایج به‌دست‌آمده ارائه شده و عملکرد مدل پیشنهادی با جدیدترین و برجسته‌ترین روش‌های موجود در ادبیات تحقیق مقایسه می‌شود. این مقایسه می‌تواند میزان بهبود دقت و کیفیت بخش‌بندی معماری پیشنهادی را در برابر مدل‌های CNN محور، ViT محور و مدل‌های ترکیبی نشان دهد.

۳-۱ مجموعه داده‌ها

برای آموزش و ارزیابی مدل پیشنهادی از دو مجموعه داده مرجع در حوزه‌ی تصاویر کولونوسکوپی شامل Kvasir-SEG و CVC-ClinicDB استفاده شده است. در ادامه، ابتدا معرفی مختصری از این دو مجموعه داده ارائه شده و سپس مراحل پیش‌پردازش مورد استفاده برای آماده‌سازی داده‌ها شرح داده می‌شود.

۳-۱-۱ Kvasir-SEG

مجموعه داده Kvasir-SEG شامل ۱۰۰۰ تصویر از پولیپ‌های کولونوسکوپی همراه با نقشه‌های برچسب‌گذاری دقیق^۱ است. تصاویر این مجموعه دارای تنوع بالا از نظر ابعاد (۳۳۲×۴۸۷ تا ۱۹۲۰×۱۰۷۲ پیکسل) و شکل ظاهری پولیپ‌ها هستند. این مجموعه داده به‌عنوان یکی از معیارهای اصلی در ارزیابی روش‌های بخش‌بندی ضایعات روده‌ای مورد استفاده قرار گرفته است. تصاویر توسط متخصصان از طریق سیستم آندوسکوپی جمع‌آوری شده‌اند [۲۶].

۳-۱-۲ CVC-ClinicDB

این مجموعه داده که با عنوان CVC-612 نیز شناخته می‌شود، شامل ۶۱۲ تصویر از ویدیوهای کولونوسکوپی در ابعاد ۲۸۸×۳۸۴ پیکسل است. تمامی تصاویر دارای برچسب‌های دقیق پولیپ هستند. این مجموعه داده نیز به‌صورت گسترده برای ارزیابی مدل‌های بخش‌بندی در تشخیص خودکار پولیپ‌ها به‌کار می‌رود.

تصاویر از ویدئوهای آندوسکوپی در بیمارستان بارسلونا اسپانیا تهیه شده اند [۲۷].

۳-۲ جزئیات آموزش و تنظیمات پیاده سازی مدل

پیاده سازی مدل پیشنهادی با استفاده از زبان برنامه نویسی Python و چارچوب PyTorch انجام شد. تمامی آزمایش ها بر روی دو واحد پردازشی گرافیکی NVIDIA Tesla T4، هر یک با ۱۶ گیگابایت حافظه ویدیویی، و به صورت پردازش موازی اجرا گردید تا کارایی محاسباتی افزایش یافته و زمان آموزش کاهش یابد. برای حفظ شرایط یکسان و امکان مقایسه ی منصفانه میان مجموعه داده ها، اندازه ی ورودی تصاویر در دو دیتاست Kvasir-SEG و CVC-ClinicDB برابر با 256×256 تنظیم شد.

در معماری پیشنهادی، از نسخه EfficientNetV2-S به عنوان رمزگذار استفاده شده است. پیاده سازی این بخش بر اساس مدل tf_efficientnetv2_s.in21k از کتابخانه timm انجام شده و وزن های پیش آموزش دیده روی مجموعه داده ImageNet-21K برای مقاداردهی اولیه رمزگذار مورد استفاده قرار گرفته اند. لایه طبقه بندی انتهایی مدل حذف شده و رمزگذار در حالت استخراج ویژگی های چند مرحله ای (features_only=True) به کار گرفته شده است. خروجی مراحل مختلف رمزگذار به ترتیب دارای ۲۴، ۴۸، ۶۴، ۱۶۰ و ۲۵۶ کانال است که ویژگی های سطوح مختلف آن از طریق اتصالات میان بر در اختیار مراحل متناظر رمزگشا قرار می گیرند. ساختار داخلی Stem و بلوک های Fused-MBConv و MBConv در EfficientNetV2-S تغییر داده نشده و مؤلفه BiFormer تنها در بخش رمزگشا به معماری افزوده شده است. در طول آموزش، وزن های رمزگذار نیز همراه با سایر اجزای شبکه به صورت end-to-end به روزرسانی شدند.

در دو مجموعه داده، فرآیند آموزش با استفاده از روش اعتبارسنجی متقابل پنج تایی^۱ انجام گرفت تا از ایجاد سوگیری ناشی از تقسیم بندی تصادفی جلوگیری شده و پایداری نتایج افزایش یابد. در هر فولد، ۲۰ درصد از داده ها برای آزمون و ۸۰ درصد برای آموزش و اعتبارسنجی در نظر گرفته شد؛ به گونه ای که از این ۸۰ درصد، ۹۰ درصد برای آموزش و ۱۰ درصد برای اعتبارسنجی استفاده شد. میانگین نتایج پنج فولد به عنوان عملکرد نهایی مدل گزارش شده است.

برای جلوگیری از بیش برآزش و ذخیره ی بهترین وزن های مدل، در تمامی آزمایش ها از سازوکار توقف زودهنگام^۲ همراه با سازوکار ذخیره بهترین مدل در طول آموزش^۳ بهره گرفته شد. در هر فولد، معیار انتخاب بهترین checkpoint، بیشترین مقدار Dice

روی مجموعه اعتبارسنجی بود و مدلی که بالاترین Validation Dice را کسب می کرد برای ارزیابی نهایی همان فولد مورد استفاده قرار می گرفت. همچنین از Batch Size برابر با ۳۲ برای دو دیتاست استفاده شد تا تعادلی مناسب میان سرعت آموزش و پایداری گرادین ها ایجاد شود.

بهینه ساز Adam با نرخ یادگیری اولیه 0.001 به کار گرفته شد. تابع هزینه ترکیبی شامل ۳۰ درصد Cross-Entropy Loss و ۷۰ درصد Dice Loss بود. مدت زمان آموزش در دو مجموعه داده بسته به اندازه ی مجموعه ی داده و میزان پیچیدگی تصاویر متفاوت بود. در دیتاست Kvasir-SEG مدل پس از ۸۵ اپاک و طی مدت ۱ ساعت و ۴۵ دقیقه متوقف شد. آموزش مدل بر روی دیتاست CVC-ClinicDB نیز طی ۱۵۰ اپاک حدود ۱ ساعت و ۵ دقیقه به طول انجامید.

ابزارهای اصلی آموزش شامل Batch Size، نرخ یادگیری، وزن کاهی و ضرایب تابع هزینه پیش از اجرای ارزیابی نهایی تعیین شده و سپس در تمامی فولدها ثابت نگه داشته شدند. مجموعه تنظیمات فوق در تمامی آزمایش ها به صورت یکسان اعمال شد تا شرایط آموزش و ارزیابی دو مجموعه داده تا حد امکان کنترل شده و قابل مقایسه باشد.

به منظور افزایش قابلیت بازتولید نتایج، کد پیاده سازی مدل پیشنهادی، شامل معماری شبکه، فرآیند آموزش و ارزیابی، تنظیمات اعتبارسنجی متقابل و محاسبه معیارهای عملکرد، به صورت عمومی در مخزن GitHub در دسترس قرار گرفته است.

۳-۳ معیارهای ارزیابی

برای ارزیابی عملکرد مدل پیشنهادی در بخش بندی تصاویر کولونوسکوپی، از مجموعه ای از معیارهای رایج و معتبر شامل ضریب دایس^۴، شاخص هم پوشانی^۵، دقت پیش بینی مثبت^۶، حساسیت^۷ استفاده شده است. این معیارها به طور گسترده در تحقیقات حوزه ی بخش بندی پزشکی مورد استفاده قرار می گیرند و امکان تحلیل عملکرد مدل از جنبه های مختلف دقت، هم پوشانی، توانایی تشخیص نواحی هدف و خطاهای مدل را فراهم می کنند.

ضریب Dice و IoU معیارهای اصلی در سنجش میزان هم پوشانی بین نواحی پیش بینی شده و واقعی هستند و به ویژه در کاربردهای پزشکی اهمیت ویژه ای دارند، زیرا بسیاری از ساختارهای مورد بررسی دارای ابعاد کوچک بوده و حساسیت مدل نسبت به جزئیات مرزی از اهمیت بالایی برخوردار است.

دو معیار Precision و Recall عملکرد مدل را از دو منظر متفاوت بررسی می کنند: Precision بر کاهش پیش بینی های اشتباه^۸

^۵ Intersection over Union (IoU)

^۶ Precision

^۷ Recall

^۸ False Positive

^۱ 5-Fold Cross Validation

^۲ Early Stopping

^۳ Best Checkpoint

^۴ Dice

برای مدل پیشنهادی، نتایج به صورت میانگین $\pm$ انحراف معیار پنج فولد گزارش شده‌اند. مقدار انحراف معیار Dice برابر با 0.64 واحد درصد و IoU برابر با 0.99 واحد درصد است که نشان‌دهنده نوسان محدود عملکرد مدل در تقسیم‌بندی‌های مختلف داده است.

به طور کلی، ترکیب کارایی محاسباتی و دقت بهبودیافته در بخش‌بندی پولیپ‌ها نشان می‌دهد که مدل پیشنهادی نه تنها توانسته است از روش‌های کلاسیک عبور کند، در مقایسه با نتایج گزارش شده برای بسیاری از شبکه‌های مبتنی بر Transformer، عملکرد رقابتی همراه با پیچیدگی محاسباتی پایین‌تری نشان می‌دهد.

۳-۴-۲ نتایج بخش‌بندی روی دیتاست CVC-ClinicDB

نتایج جدول شماره (۲) نشان می‌دهد که مدل پیشنهادی، با کسب ضریب Dice برابر 0.95 و IOU برابر $0.92/83$ ، در میان نتایج گزارش شده در مطالعات مقایسه شده، مقادیر بالاتری برای Dice و IoU ارائه کرده است. و از مدل‌های قدرتمندی مانند MixFormer (با دایس ثبت شده $0.94/41$)، DS-TransUNet (مقدار دایس برابر با $0.94/2$) و BRAU-Net++ (مقدار دایس $0.92/94$) نیز تا حدودی فراتر رفته است.

همچنین در معیارهای کلیدی Precision با مقدار $0.96/24$ و Recall با مقدار $0.96/15$ ، مدل پیشنهادی بالاترین مقادیر را ثبت کرده است که نشان‌دهنده توانایی آن در شناسایی دقیق لبه‌ها و جلوگیری از خطاهای نقص تشخیص^۱ و بیش تشخیص^۲ می‌باشد.

افزون بر این، مدل ما تنها $20/82$ میلیون پارامتر و $5/31$ میلیارد FLOPs دارد که به طور قابل توجهی کمتر از مدل‌های سنگینی مانند MixFormer با بیش از 123 میلیون پارامتر و 108 FLOPs است؛ بنابراین، مدل پیشنهادی در کنار عملکرد مناسب بخش‌بندی، در مقایسه با برخی مدل‌های مورد بررسی دارای پیچیدگی محاسباتی کمتری است و از نظر هزینه محاسباتی برای بررسی در سامانه‌های با محدودیت منابع گزینه قابل بررسی است، هرچند قابلیت استفاده بالینی آن نیازمند ارزیابی در شرایط عملی واقعی است.

در مجموعه داده CVC-ClinicDB نیز نتایج پنج فولد پایداری مناسبی نشان دادند؛ به طوری که انحراف معیار Dice برابر با 0.81 واحد درصد و IoU برابر با 0.90 واحد درصد به دست آمد. بنابراین، عملکرد نهایی مدل صرفاً ناشی از یک تقسیم‌بندی خاص از داده‌ها نبوده است.

لازم به ذکر است که نتایج مدل‌های مقایسه شده در جدول‌های (۱) و (۲) از مطالعات مختلف استخراج شده‌اند و تمامی روش‌ها تحت شرایط کاملاً یکسان از نظر تقسیم‌بندی داده، تنظیمات آموزش، پیش‌پردازش، اندازه ورودی و سخت‌افزار ارزیابی نشده‌اند.

متمرکز است، در حالی که Recall توانایی مدل در تشخیص کامل نواحی صحیح را می‌سنجد. این دو معیار در کاربردهای پزشکی اهمیت بالایی دارند، زیرا نوع خطا می‌تواند پیامدهای بالینی متفاوتی ایجاد کند.

ترکیب این معیارها امکان ارزیابی جامع و چندوجهی مدل را فراهم می‌سازد. در این پژوهش، ضریب دایس به عنوان معیار اصلی فرایند آموزش و اعتبارسنجی انتخاب شده و سایر معیارها برای ارزیابی نهایی مدل بر روی مجموعه‌ی آزمون گزارش شده‌اند؛ رویکردی که موجب ایجاد مقایسه‌پذیری دقیق با سایر روش‌های موجود و ارائه‌ی تحلیل کامل از عملکرد مدل می‌شود.

۳-۴-۳ مقایسه نتایج

۳-۴-۳-۱ نتایج بخش‌بندی بر روی مجموعه داده Kvasir-SEG

نتایج مقایسه‌ای در جدول شماره (۱) نشان می‌دهد که مدل پیشنهادی در مقایسه با مقادیر گزارش شده در مطالعات پیشین، عملکردی رقابتی و در برخی معیارها مقادیر بالاتری ارائه می‌دهد. این مدل با دستیابی به مقدار Dice $0.92/85$ و مقدار IOU برابر با $0.88/02$ در مقایسه با مقادیر گزارش شده برای بسیاری از روش‌های مطرح، از جمله DS-TransUNet، Swin-UNet و حتی مدل‌های سنگین سال 2025 مانند MixFormer بالاتر است. برای مثال، مقدار دایس مدل پیشنهادی نسبت به مدل MixFormer با مقدار $0.92/47$ حدود 0.38 درصد بیشتر است و نسبت به Swin-UNet نیز بیش از $1/9$ درصد بهبود نشان می‌دهد. همچنین در شاخص‌های Precision و Recall نیز نتایج این پژوهش (به ترتیب $0.93/07$ و $0.93/92$) بالاتر از مدل‌های سبک‌تر مثل U-Net و قدرتمندترهایی چون TransUNet و Polyp-PVT بوده که بیانگر توانایی بهتر مدل در شناسایی دقیق و بازخوانی نواحی هدف در تصاویر آندوسکوپی است.

در کنار این دقت بهبودیافته، مدل پیشنهادی در مقایسه با مقادیر گزارش شده برای برخی روش‌های مورد بررسی، از تعداد پارامترها و بار محاسباتی کمتری برخوردار است. تعداد پارامترهای مدل $20/82$ میلیون و میزان FLOPs برابر با $5/31$ میلیارد است.

مقایسه این اعداد با مدل‌هایی مانند DS-TransUNet با $287/75$ میلیون پارامتر و MixFormer با $108/32$ میلیارد FLOPs نشان می‌دهد که مدل پیشنهادی با حجمی بسیار کمتر و مصرف محاسباتی پایین‌تر، خروجی‌هایی هم‌تراز یا بهتر ارائه می‌کند. این ویژگی، در کنار زمان استنتاج اندازگیری شده، نشان‌دهنده پتانسیل مدل برای استفاده در سامانه‌های با محدودیت منابع است؛ با این حال، ارزیابی عملکرد در محیط‌های بالینی واقعی و سخت‌افزارهای هدف نیازمند بررسی مستقل است.

جدول ۱: نتایج متریک های مختلف روی مدل های مختلف و مدل پیشنهادی در دیتاست Kvasir-SEG (واحد پارامتر میلیون و FLOPs میلیارد است). علامت «-» نشان دهنده آن است که مقدار مربوطه در منبع اصلی مورد استناد گزارش نشده است؛ بنابراین، برای جلوگیری از برآورد یا مقایسه غیرهمسان، مقداری توسط نویسندگان جایگزین یا تخمین زده نشده است.

مدل	Dice (%)	IOU (%)	Precision (%)	Recall (%)	Params	FLOPs	سال
TransUNet[14-28]	۸۸/۲۳	۸۲/۰۸	۸۳/۵۶	۸۲/۳۱	۹۳/۲۳	-	۲۰۲۱
U-Net[4-28]	۸۸/۳۶	۸۲/۴۵	۸۱/۵۲	۸۳/۷۱	۲۴/۵	۱۰/۵	۲۰۱۵
Mas-TransUnet[29]	۹۰/۶۰	۸۴/۳	-	-	۱۱۷/۷	-	۲۰۲۵
Swin-Unet[13-28]	۹۰/۹۱	۸۵/۶۳	۸۵/۱۶	۸۷/۳۶	۹۰/۶۶	۵/۹۱	۲۰۲۱
HTC-Net[28]	۹۱/۱۸	۸۵/۹۴	۸۶/۲۲	۸۶/۷۱	۲/۰۷	-	۲۰۲۴
DS-TransUNet[23]	۹۱/۳۰	۸۵/۹۰	۹۱/۶۰	۹۳/۶۰	۲۸۷/۷۵	۵۱/۰۹	۲۰۲۲
Polyp-PVT[22]	۹۱/۷۰	۸۶/۴	-	-	۲۵/۱۱	-	۲۰۲۴
META-Unet[30]	۹۱/۸۶	۸۶/۴۲	-	-	۲۲/۲۱	۱۰۸/۳۲	۲۰۲۵
MixFormer[31]	۹۲/۴۷	۸۶/۱۵	۹۱/۸۱	۹۴/۶۳	۱۲۳/۳۵	-	۲۰۲۵
مدل پیشنهادی	۹۲/۸۵ ± ۰/۶۴	۸۸/۰۲ ± ۰/۹۹	۹۳/۰۷ ± ۰/۸۴	۹۳/۹۲ ± ۱/۱۸	۲۰/۸۲	۵/۳۱	۲۰۲۶

جدول ۲: نتایج متریک های مختلف روی مدل های مختلف و مدل پیشنهادی در دیتاست CVC-ClinicDB (واحد پارامتر میلیون و FLOPs میلیارد است). علامت «-» نشان دهنده آن است که مقدار مربوطه در منبع اصلی مورد استناد گزارش نشده است؛ بنابراین، برای جلوگیری از برآورد یا مقایسه غیرهمسان، مقداری توسط نویسندگان جایگزین یا تخمین زده نشده است.

مدل	Dice (%)	IOU	Precision	Recall	Params	FLOPs	سال
U-Net[4-31]	۹۱/۳۱	۸۵/۶۸	۹۱/۴۱	۹۱/۶۸	۲۴/۵۶	۱۰/۵	۲۰۱۵
TransUNet[14-31]	۸۹/۰۲	۸۲/۲۷	۸۷/۹۶	۹۱/۲۶	۹۳/۲۳	۲۴/۷۳	۲۰۲۱
Swin-Unet[13-31]	۹۱/۳۸	۸۵/۹۵	۹۲/۶۱	۹۲/۸۱	۹۰/۶۶	۵/۹۱	۲۰۲۱
DCSAU-Net[32-15]	۹۱/۶۷	۸۶/۱۸	۹۱/۷۲	۹۲/۰۳	۲/۵۸	۶/۲۴	۲۰۲۳
BRAU-Net++[15]	۹۲/۹۴	۸۸/۱۷	۹۳/۸۴	۹۳/۰۶	۵۰/۷۶	۲۲/۳۹	۲۰۲۴
Polyp-PVT[22]	۹۳/۷	۸۸/۹	-	-	۲۵/۱۱	-	۲۰۲۴
META-Unet[30]	۹۳/۸۵	۸۸/۹۹	-	-	۲۲/۲۱	-	۲۰۲۴
DS-TransUNet[23]	۹۴/۲۲	۸۹/۳۹	۹۳/۶۹	۹۵	۲۸۷/۷۵	۵۱/۰۹	۲۰۲۲
MixFormer[31]	۹۴/۴۱	۸۹/۲۲	۹۴/۳۷	۹۴/۵۸	۱۲۳/۳۵	۱۰۸/۳۲	۲۰۲۵
مدل پیشنهادی	۹۵/۰۲ ± ۰/۸۱	۹۲/۸۳ ± ۱/۹۰	۹۶/۲۴ ± ۱/۰۵	۹۶/۱۵ ± ۱/۶۰	۲۰/۸۲	۵/۳۱	۲۰۲۶

پولپ های کوچک، کشیده و نواحی دارای تفاوت کنتراست محدود با بافت اطراف نیز قادر به تولید ماسک های منسجم است. با وجود اختلاف های جزئی در برخی قسمت های مرزی میان ماسک پیش بینی شده و Ground Truth، شکل کلی ضایعه و موقعیت آن به خوبی بازیابی شده است. این مشاهدات کیفی با مقادیر بالای Dice و IoU گزارش شده در ارزیابی کمی سازگار بوده و نشان می دهد که مدل علاوه بر عملکرد عددی مناسب، از نظر بازسازی شکل و مرز پولپ نیز عملکرد قابل قبولی ارائه می دهد.

۳-۶ تحلیل عدم قطعیت و کالبراسیون

به منظور بررسی میزان اطمینان مدل در پیش بینی نواحی پولپ، به ویژه در نواحی مرزی و نمونه های دشوار، یک تحلیل تکمیلی عدم قطعیت بر روی فولد صفر مجموعه های آزمون Kvasir-SEG و CVC-ClinicDB انجام شد. اهمیت تحلیل عدم قطعیت و قابلیت اطمینان پیش بینی ها در سامانه های بخش بندی تصاویر پزشکی در مطالعات اخیر نیز مورد تأکید قرار گرفته است [۳۳، ۳۴]. در این ارزیابی از روش Monte Carlo Dropout استفاده

بنابراین، این مقایسه با هدف ارائه یک دید کلی از جایگاه عملکرد مدل پیشنهادی در میان روش های موجود انجام شده است و نباید به عنوان یک مقایسه کاملاً کنترل شده و مستقیم بین تمامی مدل ها تفسیر شود.

۳-۵ ارزیابی کیفی نتایج بخش بندی

به منظور ارزیابی بصری عملکرد مدل پیشنهادی، تعدادی نمونه از تصاویر ورودی، ماسک مرجع و خروجی پیش بینی شده برای دو مجموعه داده Kvasir-SEG و CVC-ClinicDB در شکل های شماره (۷) و (۸) ارائه شده است. همان گونه که مشاهده می شود، مدل در نمونه هایی با اندازه ها، شکل ها و موقعیت های متفاوت پولپ قادر است ناحیه هدف را با تطابق مناسبی با ماسک مرجع استخراج کند. در نمونه های دارای شکل نامنظم و همچنین تصاویر شامل بیش از یک ناحیه پولپ، مدل توانسته است ساختار کلی نواحی هدف و بخش عمده مرزهای آن ها را حفظ کند. همچنین نتایج بصری نشان می دهد که مدل در مواجهه با

شد؛ به گونه‌ای که برای هر تصویر، مدل ۲۰ بار در مرحله استنتاج با فعال

نگه‌داشتن لایه‌های Dropout اجرا شد.

با پیکسل‌های غیرمرزی و همچنین در پیکسل‌های دارای پیش‌بینی اشتباه با پیکسل‌های صحیح مقایسه شد. برای تسهیل تفسیر، نسبت این مقادیر در جدول (۳) گزارش شده است؛ به طوری که مقدار بزرگ‌تر از یک نشان‌دهنده افزایش عدم قطعیت در نواحی مرزی یا پیکسل‌های دارای خطای بخش‌بندی است.

مطابق جدول (۳)، در هر دو مجموعه داده میزان عدم قطعیت در نواحی دشوار افزایش یافته است. در Kvasir-SEG، میانگین Predictive Entropy در پیکسل‌های مرزی حدود $1/63$ برابر نواحی غیرمرزی و در پیکسل‌های دارای خطای بخش‌بندی حدود $1/61$ برابر پیکسل‌های صحیح بود. در CVC-ClinicDB نیز این نسبت‌ها به ترتیب $1/30$ و $1/33$ به دست آمدند. بنابراین، در هر دو مجموعه داده عدم قطعیت پیش‌بینی در اطراف مرز پولیپ و در نواحی دارای اختلاف میان پیش‌بینی مدل و ماسک مرجع بیشتر بوده است.

جدول ۳: نتایج تحلیل عدم قطعیت و کالیبراسیون مدل پیشنهادی

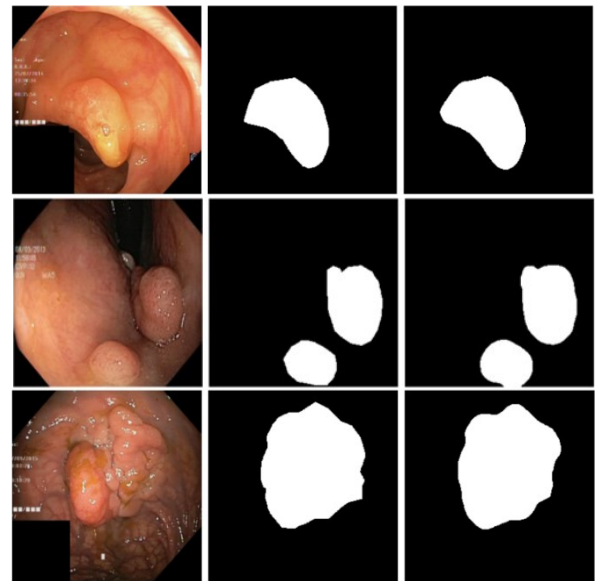
نسبت آنتروپی خطا/صحیح	ECE	Brier Score	نسبت آنتروپی مرزی/غیرمرزی	دیتاست
$1/61$	$0/1313$	$0/0487$	$1/63$	Kvasir-SEG
$1/33$	$0/1894$	$0/0491$	$1/30$	CVC-ClinicDB

از منظر کالیبراسیون، مقادیر ECE برای Kvasir-SEG و CVC-ClinicDB به ترتیب $0/1313$ و $0/1894$ و مقادیر Brier Score به ترتیب $0/0487$ و $0/0491$ به دست آمدند. این نتایج نشان می‌دهد که خروجی احتمالاتی مدل می‌تواند اطلاعات مفیدی درباره میزان اطمینان پیش‌بینی فراهم کند؛ با این حال، مقادیر ECE بیانگر آن است که کالیبراسیون مدل کاملاً ایده‌آل نیست و بهبود کالیبراسیون احتمالاتی می‌تواند در مطالعات آینده مورد توجه قرار گیرد.

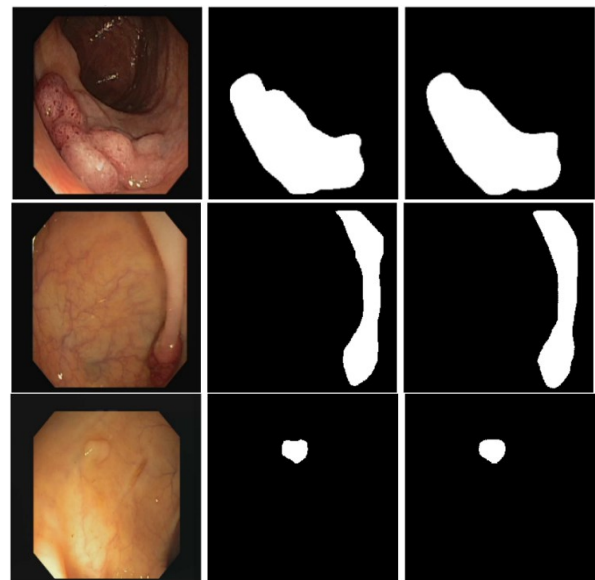
به منظور بررسی کیفی این رفتار، در شکل‌های شماره (۹) و (۱۰) نمونه‌هایی با سه سطح عملکرد شامل خوب، متوسط و چالش‌برانگیز برای دو مجموعه داده نمایش داده شده‌اند. در هر نمونه، تصویر ورودی، ماسک مرجع، ماسک پیش‌بینی شده، نقشه Predictive Entropy و نقشه عدم قطعیت اپیستمیک مبتنی بر MI ارائه شده است. مشاهده می‌شود که نواحی دارای عدم قطعیت بیشتر عمدتاً در اطراف مرزهای پولیپ و بخش‌هایی که اختلاف بیشتری میان پیش‌بینی مدل و ماسک مرجع وجود دارد متمرکز شده‌اند. این مشاهده با نتایج کمی جدول (۳) نیز سازگار است.

۷-۳ مطالعه حذف مؤلفه‌ها (Ablation Study)

به منظور بررسی سهم هر یک از مؤلفه‌های اصلی معماری پیشنهادی در عملکرد نهایی مدل، مجموعه‌ای از آزمایش‌های حذف مؤلفه (Ablation Study) بر روی دو مجموعه داده Kvasir-SEG و CVC-ClinicDB انجام شد. در این آزمایش‌ها، اثر مازول



شکل شماره (۷): نمونه‌های کیفی بخش‌بندی مدل پیشنهادی روی مجموعه داده Kvasir-SEG؛ از چپ به راست: تصویر ورودی، ماسک مرجع و ماسک پیش‌بینی شده



شکل شماره (۸): نمونه‌های کیفی بخش‌بندی مدل پیشنهادی روی مجموعه داده CVC-ClinicDB؛ از چپ به راست: تصویر ورودی، ماسک مرجع و ماسک پیش‌بینی شده.

با استفاده از توزیع پیش‌بینی‌های حاصل، Predictive Entropy به عنوان معیاری از ابهام پیش‌بینی و Mutual Information (MI) به عنوان شاخصی از عدم قطعیت اپیستمیک مدل استخراج شد. همچنین، برای بررسی میزان انطباق اطمینان پیش‌بینی مدل با صحت واقعی آن، دو معیار Expected Calibration Error (ECE) و Brier Score محاسبه شدند.

برای بررسی اینکه عدم قطعیت مدل در چه نواحی از تصویر متمرکز می‌شود، میانگین Predictive Entropy در پیکسل‌های مرزی

واحد درصد در Kvasir-SEG و ۲/۳۱ واحد درصد در CVC-ClinicDB نسبت به مدل کامل شد. همچنین، استفاده تنها از SE نسبت به مدل پایه باعث بهبود عملکرد شد، اما همچنان Dice به ترتیب ۱/۸۸ و ۱/۶۶ واحد درصد کمتر از مدل کامل باقی ماند.

جدول ۴: نتایج مطالعه Ablation روی مجموعه داده Kvasir-SEG

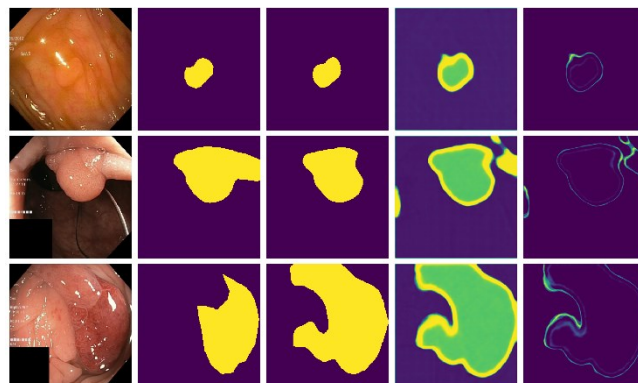
حالات	Pretrained	BiFormer	SE	Dice (%)	IoU (%)	ΔDice
رمزگشای پیشگی	✓	×	×	۸۹/۶۸	۸۴/۵۶	-۳/۱۷
پایه + SE	✓	×	✓	۹۰/۹۷	۸۵/۸۸	-۱/۸۸
پایه + BiFormer	✓	✓	×	۹۲/۱۲	۸۷/۱۹	-۰/۷۳
مدل کل پیشنهادی	×	✓	✓	۸۴/۲۴	۷۷/۱۸	-۸/۶۱
مدل کل پیشنهادی	✓	✓	✓	۹۲/۸۵	۸۸/۰۲	۰/۰۰

جدول ۵: نتایج مطالعه Ablation روی مجموعه داده CVC-ClinicDB

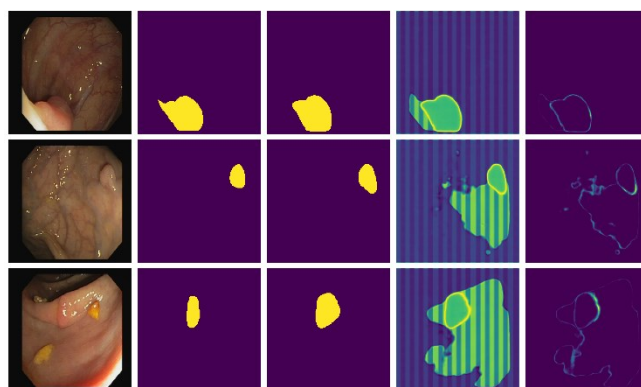
حالات	Pretrained	BiFormer	SE	Dice (%)	IoU (%)	ΔDice
رمزگشای پیشگی	✓	×	×	۹۲/۷۱	۸۸/۷۴	-۲/۳۱
پایه + SE	✓	×	✓	۹۳/۳۶	۸۹/۶۶	-۱/۶۶
پایه + BiFormer	✓	✓	×	۹۴/۴۲	۹۱/۷۲	-۰/۶۰
مدل کل پیشنهادی	×	✓	✓	۸۶/۴۸	۷۹/۸۱	-۸/۵۴
مدل کل پیشنهادی	✓	✓	✓	۹۵/۰۲	۹۲/۸۳	۰/۰۰

در مقابل، استفاده از BiFormer بدون SE عملکرد نزدیک تری به مدل کامل ایجاد کرد؛ به طوری که حذف SE از مدل دارای BiFormer موجب کاهش Dice به میزان ۰/۷۳ واحد درصد در Kvasir-SEG و ۰/۶۰ واحد درصد در CVC-ClinicDB شد. این نتیجه نشان می دهد که در میان دو مؤلفه افزوده شده به رمزگشا، BiFormer سهم بیشتری در بهبود عملکرد دارد، در حالی که SE نیز به صورت مکمل موجب افزایش بیشتر کیفیت بخش بندی می شود. بیشترین افت عملکرد زمانی مشاهده شد که وزن های پیش آموزش دیده رمزگذار حذف شدند. در این حالت، Dice در Kvasir-SEG از ۹۲/۸۵ به ۸۴/۲۴٪ و در CVC-ClinicDB از ۹۵/۰۲ به ۸۶/۴۸٪ کاهش یافت. بنابراین، استفاده از انتقال یادگیری در EfficientNetV2-S نقش مهمی در دستیابی به عملکرد نهایی دارد. در مجموع، بهترین عملکرد در هر دو مجموعه داده زمانی حاصل شد که رمزگذار پیش آموزش دیده همراه با هر دو مؤلفه BiFormer و SE مورد استفاده قرار گرفت؛ موضوعی که نقش مکمل این اجزا در معماری پیشنهادی را تأیید می کند.

BiFormer، ماژول SE و استفاده از وزن های پیش آموزش دیده در رمزگذار EfficientNetV2-S به صورت جداگانه بررسی شد. سایر تنظیمات آموزش در تمامی حالت ها ثابت نگه داشته شدند تا تغییرات عملکرد تا حد امکان به مؤلفه مورد بررسی نسبت داده شود.



شکل شماره (۹): تحلیل کیفی عدم قطعیت مدل پیشنهادی بر روی مجموعه داده Kvasir-SEG. از چپ به راست: تصویر ورودی، ماسک مرجع، ماسک پیش بینی شده، نقشه Predictive Entropy و نقشه عدم قطعیت اپیستیمیک مبتنی بر Mutual Information. ردیف ها به ترتیب نمونه های خوب، متوسط و چالش برانگیز را نشان می دهند.



شکل شماره (۱۰): تحلیل کیفی عدم قطعیت مدل پیشنهادی بر روی مجموعه داده CVC-ClinicDB. از چپ به راست: تصویر ورودی، ماسک مرجع، ماسک پیش بینی شده، نقشه Predictive Entropy و نقشه عدم قطعیت اپیستیمیک مبتنی بر Mutual Information. ردیف ها به ترتیب نمونه های خوب، متوسط و چالش برانگیز را نشان می دهند.

در حالت نخست، هر دو ماژول BiFormer و SE از مسیر رمزگشا حذف شدند. در دو حالت دیگر، به ترتیب تنها SE یا تنها BiFormer در معماری حفظ شد. همچنین برای بررسی نقش انتقال یادگیری، نسخه ای از مدل کامل بدون استفاده از وزن های پیش آموزش دیده EfficientNetV2-S مورد ارزیابی قرار گرفت. مدل کامل پیشنهادی شامل هر دو ماژول BiFormer و SE همراه با رمزگذار پیش آموزش دیده، به عنوان حالت مرجع در نظر گرفته شد. نتایج مطالعه Ablation در جداول شماره (۴) و (۵) نشان می دهد که هر سه مؤلفه مورد بررسی در عملکرد نهایی مدل نقش دارند. حذف هم زمان BiFormer و SE موجب کاهش Dice به میزان ۳/۱۷

۳-۸ ارزیابی تعمیم‌پذیری بین‌دیتاستی

به‌منظور بررسی توانایی تعمیم‌پذیری مدل پیشنهادی در مواجهه با داده‌های مستقل از منبع آموزش، دو ارزیابی بین مجموعه‌داده‌ای به‌صورت تکمیلی انجام شد. در این ارزیابی، مدل در هر مرحله تنها روی یکی از دو مجموعه‌داده Kvasir-SEG و CVC-ClinicDB آموزش داده شد و بدون هیچ‌گونه بازآموزی یا fine-tuning، مستقیماً روی مجموعه‌داده دیگر مورد آزمون قرار گرفت. بدین‌ترتیب، هدف این بخش سنجش توانایی مدل در انتقال به دامنه‌ای متفاوت از داده‌های آموزش بوده است.

جدول شماره (۶) نتایج این ارزیابی را نشان می‌دهد. هنگامی که مدل روی Kvasir-SEG آموزش داده شد و روی CVC-ClinicDB آزمون گردید، Dice برابر با $91/85\%$ و IoU برابر با $85/66\%$ به‌دست آمد. در جهت معکوس، یعنی آموزش روی CVC-ClinicDB و آزمون روی Kvasir-SEG، مقادیر Dice و IoU به‌ترتیب برابر با $86/40\%$ و $77/16\%$ حاصل شد. این نتایج نشان می‌دهد که مدل پیشنهادی از توانایی مناسبی برای تعمیم به مجموعه‌داده‌ای مستقل برخوردار است.

جدول ۶: نتایج ارزیابی تعمیم‌پذیری بین‌دیتاستی مدل پیشنهادی

دیتاست آموزش	دیتاست آزمون	Dice (%)	IoU (%)	Precision (%)	Recall (%)
Kvasir-SEG	CVC-ClinicDB	91/85	85/66	96/30	88/33
CVC-ClinicDB	Kvasir-SEG	86/40	77/16	83/92	89/67

همچنین مشاهده می‌شود که عملکرد مدل در مسیر:

Kvasir-SEG → CVC-ClinicDB

بهرتر از مسیر CVC-ClinicDB → Kvasir-SEG است. این تفاوت را می‌توان به تنوع بیشتر تصاویر و گستردگی بالاتر مجموعه‌داده Kvasir-SEG نسبت داد؛ به‌گونه‌ای که آموزش روی این مجموعه منجر به یادگیری بازنمایی‌های غنی‌تر و در نتیجه تعمیم بهتر به داده‌های CVC-ClinicDB شده است. در مقابل، محدودتر بودن تنوع تصاویر در CVC-ClinicDB می‌تواند علت افت نسبی عملکرد هنگام انتقال مدل به Kvasir-SEG باشد.

۳-۹ بررسی تأثیر محل قرارگیری BiFormer در رمزگشا

به‌منظور بررسی تأثیر محل قرارگیری ماژول BiFormer در هر مرحله از رمزگشا، چهار آرایش P2 تا P5 مورد ارزیابی قرار گرفتند. در این آرایش‌ها، موقعیت BiFormer نسبت به عملیات افزایش ابعاد، ادغام ویژگی‌های رمزگذار و رمزگشا، ماژول SE و لایه پیچشی تغییر داده شد. در P2، P3، P4 و P5، BiFormer به‌ترتیب پس از

ConvTranspose، پس از ادغام ویژگی‌ها، پس از SE و پس از لایه پیچشی قرار گرفت. حالت P5 همان آرایش مورد استفاده در مدل نهایی پیشنهادی است.

برای مقایسه این چهار حالت، علاوه بر عملکرد بخش‌بندی روی دو مجموعه‌داده Kvasir-SEG و CVC-ClinicDB، تعداد پارامترها، GFLOPs، Latency نیز بررسی شدند. اندازه‌گیری سرعت با ورودی RGB به ابعاد 256×256 ، Batch Size برابر یک و روی GPU از نوع NVIDIA Tesla T4 انجام شد. پیش از اندازه‌گیری، ۱۰۰ اجرای Warm-up انجام و سپس زمان استنتاج طی ۵۰۰ اجرای متوالی محاسبه شد. نتایج مقدار Dice نیز به صورت میانگین ۵ فولد گزارش شده است.

نتایج جدول شماره (۷) نشان می‌دهد که آرایش P5 بهترین عملکرد بخش‌بندی را در هر دو مجموعه‌داده به دست آورده است. مقدار Dice در Kvasir-SEG از $92/18\%$ در P2 به $92/85\%$ در P5 افزایش یافته و در CVC-ClinicDB نیز از $94/45\%$ به $95/02\%$ رسیده است. اگرچه اختلاف دقت میان موقعیت‌های مختلف محدود است، روند نتایج نشان می‌دهد که قرارگیری BiFormer پس از ادغام ویژگی‌های رمزگذار و رمزگشا، بازتنظیم کانالی توسط SE و پالایش پیچشی، برای مدل‌سازی روابط غیرمحلی مناسب‌تر بوده است.

جدول ۷: مقایسه عملکرد و پیچیدگی محاسباتی موقعیت‌های مختلف BiFormer

موقعیت	Dice Kvasir (%)	Dice CVC (%)	Params (M)	GFLOPs	Latency (ms/image)
P2	92/18	94/45	20/825	5/315	30/965
P3	92/31	94/61	22/954	7/446	32/275
P4	92/49	94/76	22/954	7/446	32/027
P5	92/85	95/02	20/825	5/315	31/031

از منظر هزینه محاسباتی، آرایش‌های P2 و P5 هر دو دارای $20/825$ میلیون پارامتر و $5/315$ GFLOPs هستند و سرعت استنتاج تقریباً یکسانی دارند. در مقابل، قرارگیری BiFormer در موقعیت‌های P3 و P4 باعث افزایش تعداد پارامترها به $22/954$ میلیون و GFLOPs به $7/446$ شد و هم‌زمان Latency افزایش یافت. در نتیجه، P5 نه‌تنها بالاترین Dice را در هر دو مجموعه‌داده ارائه می‌دهد، بلکه پیچیدگی محاسباتی آن نیز در سطح P2 باقی می‌ماند. بنابراین، این آرایش بهترین توازن میان دقت بخش‌بندی و هزینه محاسباتی را فراهم کرده و به‌عنوان موقعیت نهایی BiFormer در معماری پیشنهادی انتخاب شد.

۴ بحث

نتایج این پژوهش نشان می‌دهد که تفکیک نقش مؤلفه‌های پیچشی و توجهی در معماری پیشنهادی می‌تواند به ایجاد توازن مناسبی میان کیفیت بخش‌بندی و هزینه محاسباتی منجر شود.

انجام شده است؛ از این رو، عملکرد آن در شرایطی نظیر نویز شدید، کاهش کیفیت تصویر، تغییرات قابل توجه روشنایی و کنتراست، آرتیفکت های تصویری و همچنین پولیپ های بسیار کوچک یا دارای مرزهای نامشخص به صورت مستقل مورد بررسی قرار نگرفته است. بنابراین، تعمیم نتایج حاضر به تمامی شرایط بالینی باید با احتیاط صورت گیرد.

در ادامه این پژوهش، ارزیابی مقاومت مدل در برابر انواع تخریب های تصویری، بررسی عملکرد روی مجموعه داده های مستقل و بزرگتر، تحلیل اختصاصی عملکرد در پولیپ های کوچک و دشوار، و انجام ارزیابی های بین دیتاستی در نظر گرفته خواهد شد. همچنین بررسی روش های افزایش مقاومت مدل در برابر تغییرات کیفیت تصویر و کاهش هزینه محاسباتی برای استفاده در سامانه های بلادرنگ از دیگر مسیرهای توسعه این پژوهش خواهد بود.

همچنین، تحلیل عدم قطعیت و کالیبراسیون در پژوهش حاضر به صورت یک ارزیابی تکمیلی بر روی فولد صفر دو مجموعه داده انجام شده است؛ از این رو، بررسی این شاخص ها در تمامی فولدها، مجموعه داده های مستقل و همراه با روش های کالیبراسیون پس از آموزش می تواند در مطالعات آینده مورد توجه قرار گیرد.

با وجود نتایج امیدوارکننده در ارزیابی بین دیتاستی، بررسی تعمیم پذیری مدل روی مجموعه داده های بیشتر و متنوع تر، به ویژه داده های چندمرکزی و تصاویر واقعی بالینی، می تواند در مطالعات آینده انجام شود.

همچنین، بخشی از مقایسه های انجام شده با روش های پیشین بر اساس نتایج گزارش شده در مقالات اصلی آن ها انجام شده است؛ از این رو تفاوت در شرایط آموزش و ارزیابی می تواند بر مقایسه مستقیم عملکرد مدل ها اثرگذار باشد.

اگرچه زمان استنتاج مدل روی GPU نوع NVIDIA Tesla T4 اندازه گیری شده است، این ارزیابی در یک محیط آزمایشگاهی انجام شده و آزمون مستقیم سامانه در سخت افزارهای بالینی، جریان ویدئویی کولونوسکوپی و شرایط عملی بیمارستانی انجام نشده است؛ از این رو، ادعای عملکرد بلادرنگ یا آمادگی برای استفاده بالینی نیازمند ارزیابی های تکمیلی است.

اطلاعات کامل مربوط به تعداد پارامترها و FLOPs برای تمامی روش های مرجع در منابع اصلی در دسترس نبود؛ بنابراین، مقایسه کارایی محاسباتی به مدل هایی محدود شده است که این اطلاعات برای آن ها گزارش شده و نتیجه گیری در این زمینه باید با توجه به این محدودیت تفسیر شود.

۶ نتیجه گیری

در این پژوهش، یک معماری ترکیبی برای بخش بندی تصاویر کولونوسکوپی ارائه شد که در آن EfficientNetV2-S پیش آموزش دیده به عنوان رمزگذار و BiFormer به صورت هدفمند در مسیر رمزگشا به کار گرفته شد. نتایج ارزیابی روی

EfficientNetV2-S در نقش رمزگذار، ویژگی های محلی و چندمقیاسی را استخراج می کند، در حالی که BiFormer تنها در مسیر رمزگشا و پس از ادغام و پالایش ویژگی ها به کار گرفته شده است. این طراحی برخلاف برخی معماری های ترکیبی، از گسترش ساختارهای توجهی به تمام بخش های شبکه جلوگیری می کند و در نتیجه پیچیدگی محاسباتی را محدود نگه می دارد.

نتایج مطالعه Ablation نیز از این طراحی پشتیبانی می کند. حذف BiFormer افت بیشتری نسبت به حذف SE ایجاد کرد که نشان می دهد مدل سازی روابط غیرمحلی در مرحله بازسازی نقش مهمی در عملکرد نهایی دارد. در مقابل، SE نقش مکملی در بازتنظیم کانال های ویژگی ایفا می کند. همچنین افت محسوس عملکرد در حالت حذف وزن های پیش آموزش دیده نشان می دهد که انتقال یادگیری در رمزگذار EfficientNetV2-S، به ویژه با توجه به اندازه محدود مجموعه داده های پزشکی، یکی از عوامل مهم در دستیابی به عملکرد پایدار مدل است.

بررسی محل قرارگیری BiFormer نیز نشان داد که استفاده از آن پس از ادغام ویژگی های رمزگذار و رمزگشا، بازتنظیم کانالی و پالایش پیچشی، توازن مناسب تری میان دقت و هزینه محاسباتی فراهم می کند. در این حالت، BiFormer بر روی ویژگی هایی اعمال می شود که پیش از آن اطلاعات محلی و کانالی آن ها پالایش شده است. در مقابل، انتقال این ماژول به مراحل پیشین رمزگشا موجب افزایش تعداد پارامترها و GFLOPs شد، بدون آنکه افزایش متناسبی در عملکرد بخش بندی مشاهده شود.

ارزیابی بین مجموعه داده ای نیز نشان داد که مدل تا حدی قابلیت انتقال به داده هایی خارج از مجموعه آموزش را حفظ می کند، هرچند عملکرد در دو جهت انتقال یکسان نبود. عملکرد بهتر در مسیر Kvasir-SEG به CVC-ClinicDB می تواند با تنوع و تعداد بیشتر نمونه های Kvasir-SEG مرتبط باشد. این نتیجه بر اهمیت تنوع داده های آموزشی در بهبود قابلیت تعمیم مدل تأکید می کند و در عین حال نشان می دهد که ارزیابی روی داده های چندمرکزی گسترده تر همچنان ضروری است.

در مجموع، مزیت اصلی روش پیشنهادی صرفاً افزایش مقادیر معیارهای بخش بندی نیست، بلکه در نحوه ترکیب یک رمزگذار پیش آموزش دیده کارآمد با توجه غیرمحلی هدفمند در رمزگشا قرار دارد. با وجود نتایج رقابتی و پیچیدگی محاسباتی نسبتاً محدود، ارزیابی های انجام شده هنوز برای نتیجه گیری درباره آمادگی بالینی یا عملکرد بلادرنگ در شرایط عملی کافی نیستند و این موضوع نیازمند آزمون های تکمیلی در محیط های واقعی است.

۵ محدودیت ها و کارهای آینده

با وجود عملکرد مطلوب مدل پیشنهادی روی مجموعه داده های Kvasir-SEG و CVC-ClinicDB، پژوهش حاضر دارای محدودیت هایی است. ارزیابی مدل تنها بر روی دو مجموعه داده کولونوسکوپی و پس از تغییر اندازه تصاویر به ابعاد 256×256

- K. Suzuki, Ed. Rijeka, Croatia: IntechOpen, 2011, pp. 121–138.
- [4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Cham, Switzerland: Springer, 2015, pp. 234–241.
- [5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. on Learning Representations (ICLR), 2021.
- [6] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. on Machine Learning (ICML), vol. 97, pp. 6105–6114, Long Beach, CA, USA: PMLR, 2019.
- [7] B.-D. Dinh, T.-T. Nguyen, T.-T. Tran, and V.-T. Pham, "1M parameters are enough? A lightweight CNN-based model for medical image segmentation," in Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2023, pp. 1279–1284.
- [8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in Proc. International Workshop on Deep Learning in Medical Image Analysis, Cham, Switzerland: Springer, 2018, pp. 3–11.
- [9] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," Quantitative Imaging in Medicine and Surgery, vol. 10, no. 6, pp. 1275–1285, Jun. 2020.
- [10] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, Apr. 2017.
- [11] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in Proc. International Conference on Machine Learning (ICML), Jul. 2021, pp. 10347–10357.
- [12] 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/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 10012–10022.
- [13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in Proc. Eur. Conf.

مجموعه داده‌های Kvasir-SEG و CVC-ClinicDB نشان داد که مدل پیشنهادی در معیارهای اصلی بخش‌بندی عملکردی رقابتی ارائه می‌دهد و در عین حال تعداد پارامترها و هزینه محاسباتی آن در سطح نسبتاً محدودی باقی می‌ماند. نتایج مطالعه Ablation نیز نقش مکمل BiFormer، ماژول SE و انتقال یادگیری را در عملکرد نهایی مدل تأیید کرد و ارزیابی بین‌دیتاستی نشان داد که مدل تا حدی قابلیت انتقال به داده‌های مستقل از مجموعه آموزش را حفظ می‌کند.

با این حال، ارزیابی حاضر به دو مجموعه داده کولونوسکوپی محدود بوده و آزمون مستقیم مدل در داده‌های چندمرکزی، جریان ویدئویی واقعی و محیط‌های بالینی انجام نشده است. همچنین تحلیل عدم قطعیت تنها به صورت تکمیلی روی یک فولد بررسی شده است. از این رو، ارزیابی مدل بر روی مجموعه داده‌های بزرگ‌تر و متنوع‌تر، بررسی مقاومت در برابر تغییرات کیفیت تصویر، توسعه تحلیل عدم قطعیت و ارزیابی عملکرد در شرایط عملی می‌تواند از مسیرهای اصلی ادامه این پژوهش باشد.

در مجموع، نتایج این مطالعه نشان می‌دهد که ترکیب یک رمزگذار پیچشی کارآمد با توجه غیرمحلی هدفمند در رمزگشایی می‌تواند رویکردی مناسب برای دستیابی به توازن میان کیفیت بخش‌بندی و پیچیدگی محاسباتی باشد و مبنایی برای توسعه مدل‌های کارآمدتر در مطالعات آینده فراهم کند.

دسترسی به کد

کد منبع پیاده‌سازی مدل پیشنهادی، شامل معماری EfficientNetV2-S-BiFormer، فرآیند آموزش پنج‌فولدی، ارزیابی مدل و محاسبه پیچیدگی محاسباتی، به صورت عمومی در GitHub منتشر شده است:

<https://github.com/AminNoorae/EfficientNetV2-BiFormer-Polyp-Segmentation>

مراجع

- [1] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From CNN to Transformer: A Review of Medical Image Segmentation Models," Journal of Imaging Informatics in Medicine, vol. 37, no. 4, pp. 1529–1547, Aug. 2024.
- [2] A. Khan, Z. Rauf, A. R. Khan, S. H. Rathore, S. H. Khan, N. Shah, U. Farooq, H. Asif, A. Zahoor, and R. U. Khalil, "A recent survey of vision transformers for medical image segmentation," IEEE Access, Oct. 2025.
- [3] M. Jabarouti Moghaddam and H. Soltanian-Zadeh, "Medical Image Segmentation Using Artificial Neural Networks," in Artificial Neural Networks – Methodological Advances and Biomedical Applications,

- [25] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. H. Lau, "Biformer: Vision transformer with bi-level routing attention," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 10323–10333.
- [26] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A Segmented Polyp Dataset," in Proc. Int. Conf. on Multimedia Modeling (MMM), Cham, Switzerland: Springer, 2019, pp. 451–462.
- [27] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," Computerized Medical Imaging and Graphics, vol. 43, pp. 99–111, Jul. 2015.
- [28] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, and T. Tong, "HTC-Net: A hybrid CNN-transformer framework for medical image segmentation," Biomedical Signal Processing and Control, vol. 88, 105605, Feb. 2024.
- [29] A. K. Upadhyay and A. K. Bhandari, "MaS-TransUNet: A multiattention Swin Transformer U-Net for medical image segmentation," IEEE Transactions on Radiation and Plasma Medical Sciences, vol. 9, no. 5, pp. 613–626, Oct. 2024.
- [30] H. Wu, Z. Zhao, and Z. Wang, "META-Unet: Multi-scale efficient transformer attention Unet for fast and high-accuracy polyp segmentation," IEEE Transactions on Automation Science and Engineering, vol. 21, no. 3, pp. 4117–4128, Jul. 2024.
- [31] J. Liu, K. Li, C. Huang, H. Dong, Y. Song, and R. Li, "MixFormer: A mixed CNN-Transformer backbone for medical image segmentation," IEEE Transactions on Instrumentation and Measurement, vol. 74, pp. 1–12, 2025, doi: 10.1109/TIM.2024.3497060.
- [32] Q. Xu, Z. Ma, N. He, and W. Duan, "DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation," Computers in Biology and Medicine, vol. 154, p. 106626, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106626.
- [33] J. Gu, F. Tian, and I. S. Oh, "RAMIS: Increasing robustness and accuracy in medical image segmentation with hybrid CNN-transformer synergy," Neurocomputing, vol. 618, p. 129009, 2025.
- [34] X. Pan, W. Xu, and H. Xu, "DSSINet: Dynamic token selection and spatial interaction framework for medical image segmentation," Alexandria Engineering Journal, vol. 130, pp. 954–968, 2025.
- Comput. Vis. (ECCV), Cham, Switzerland: Springer Nature, Oct. 2022, pp. 205–218.
- [14] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, Feb. 2021.
- [15] L. Lan, P. Cai, L. Jiang, X. Liu, Y. Li, and Y. Zhang, "BRAU-Net++: U-shaped hybrid CNN-transformer network for medical image segmentation," arXiv preprint arXiv:2401.00722, Jan. 2024.
- [16] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer," arXiv preprint arXiv: 2109.07162, Sep. 2021.
- [17] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-compound transformer for accurate biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), Cham: Springer International Publishing, Sep. 2021, pp. 326–336.
- [18] X. Li, W. Zhu, X. Dong, O. M. Dumitrascu, and Y. Wang, "EViT-UNET: U-Net like efficient vision transformer for medical image segmentation on mobile and edge devices," in Proc. IEEE 22nd Int. Symp. Biomed. Imaging (ISBI), Apr. 2025, pp. 1–5.
- [19] Y. Chen, B. Zou, Z. Guo, Y. Huang, Y. Huang, F. Qin, Q. Li, and C. Wang, "Scunet++: Swin-UNet and CNN bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism CT image segmentation," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2024, pp. 7759–7767.
- [20] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 6202–6212.
- [21] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2021, pp. 36–46.
- [22] D. Bo, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," CAAI Artificial Intelligence Research, vol. 2, no. 3, pp. 59–0015, 2023.
- [23] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin Transformer U-Net for medical image segmentation," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–15, May 2022.
- [24] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," Proc. Int. Conf. Machine Learning (ICML), 2021, pp. 10096–10106.

تصویر، شبکه های عصبی، شناسایی الگو و تحلیل تصاویر پزشکی می باشد.



محمد امین نورائی یگانه مدرک کارشناسی خود را در رشته مهندسی برق از دانشگاه صنعتی قم در سال ۱۳۹۸ اخذ نمود. با توجه به علاقه مندی به حوزه پردازش سیگنال های دیجیتال، تا سال ۱۴۰۲ به فعالیت های پژوهشی و

تحقیقاتی در این زمینه پرداخت. در ادامه، در همان سال تحصیلات خود را در مقطع کارشناسی ارشد رشته مهندسی برق، گرایش مخابرات سیستم، در دانشگاه تهران آغاز نمود. در طول این دوره، به حوزه بینایی ماشین و بینایی رایانه ای علاقه مند شده و تمرکز خود را بر تحلیل و طراحی مدل های شبکه های عصبی معطوف ساخت. نهایتاً در سال ۱۴۰۴ با ارائه پایان نامه ای در زمینه طراحی مدل یادگیری عمیق در حوزه تصاویر پزشکی، فارغ التحصیل شد.



حمید سلطانیان زاده استاد تمام دانشکده مهندسی برق و کامپیوتر پردیس دانشکده های فنی دانشگاه تهران است. مدرک کارشناسی ارشد پیوسته مهندسی برق، گرایش الکترونیک، را در سال ۱۳۶۵ با رتبه اول از دانشکده فنی

دانشگاه تهران دریافت نمود. همچنین مدارک کارشناسی ارشد مهندسی برق، گرایش سیستم (کنترل و مخابرات)، و دکتری مهندسی برق، گرایش سیستم (پردازش سیگنال و علوم بیوالکتریک)، را به ترتیب در سال های ۱۳۶۹ و ۱۳۷۱ از دانشگاه میشیگان (آن آربر)، ایالات متحده آمریکا، اخذ کرد. ایشان به عنوان پژوهشگر ارشد با مؤسسه پزشکی فورد آمریکا و مرکز تحقیقات فیزیک نظری و ریاضیات ایران همکاری پژوهشی دارد. دکتر سلطانیان زاده سرپرستی و یا مشاوره بیش از یکصد و پنجاه پایان نامه کارشناسی ارشد و سی رساله دکتری را بر عهده داشته است. افزون بر این، بیش از هفتصد مقاله علمی در مجلات و کنفرانس های معتبر بین المللی منتشر نموده و در هیات تحریریه یا کمیته های علمی و داوری بیش از شصت مجله و کنفرانس علمی فعالیت کرده است.

ایشان عضو ارشد (Senior Member) انجمن مهندسان برق و الکترونیک (IEEE) و عضو وابسته فرهنگستان علوم جمهوری اسلامی ایران بوده و به عنوان داور طرح های پژوهشی در حوزه های بیوالکتریک و پردازش سیگنال و تصویر در سطح جهان شناخته شده است. زمینه های پژوهشی دکتر سلطانیان زاده شامل تصویرنگاری پزشکی، پردازش سیگنال و