نشریه علمی مهندسی فناوری اطلاعات مکانی

سال دهم، شماره چهارم، زمستان ۱۴۰۱

Vol.10, No.4, Winter 2023

۸۷ -۱۰۸

مقاله پژوهشی

DOR: 20.1001.1.20089635.1401.10.4.5.5

بکارگیری چارچوب یادگیری انتقال برای قطعه‌بندی معنایی تصاویر با وضوح بالا پهپاد-مبنا در مناطق شهری

عباس مجیدی‌زاده [۱]، حدیثه سادات حسنی [۲*] ، مرضیه جعفری [۳]

۱- دانشجوی کارشناسی ارشد فتوگرامتری - گروه ژئودزی و مهندسی نقشه‌برداری - دانشگاه تفرش

۲- استادیار گروه ژئودزی و مهندسی نقشه‌برداری - دانشگاه تفرش

تاریخ دریافت مقاله: ۱۴۰۱/۱۰/۰۷ تاریخ پذیرش مقاله: ۱۴۰۲/۰۲/۲۳

چکیده

قطعه‌بندی معنایی برای پردازش داده‌های وسایل نقلیه هوایی بدون سرنشین (*UAV*) یکی از تحقیقات پیشرو در زمینه فتوگرامتری، سنجش-ازدور و بینایی‌کامپیوتر در سال‌های اخیر بوده است. این روش توجه فزاینده‌ای را از طرف صنعت و دانشگاه به خود جلب کرده است. بسیاری از کاربردها، از جمله نقشه‌برداری هوایی صحنه‌های شهری، تعیین موقعیت اشیا در تصاویر هوایی، استخراج خودکار ساختمان‌ها از تصاویر سنجش‌ازدور یا هوایی با وضوح بالا و غیره، نیازمند الگوریتم‌های قطعه‌بندی دقیق و کارآمد هستند. با این حال، قطعه‌بندی معنایی مناسب و دقیق با استفاده از رویکرد یادگیری عمیق (آموزش کلی یک شبکه عصبی عمیق با وزن‌دهی تصادفی) به حجم زیادی از تصاویر آموزشی و برچسب‌گذاری‌شده نیاز دارد. با توجه به اینکه در حوزه تصاویر هوایی شهری با چالش کمبود داده‌های برچسب‌گذاری‌شده مواجه هستیم، در این مقاله از رویکرد یادگیری انتقال برای قطعه‌بندی معنایی تصاویر پهپاد-مبنا نواحی شهری استفاده کرده‌ایم. روش پیشنهادی، یک چارچوب یادگیری انتقال مبتنی بر معماری رمزگذار-رمزگشا پیچشی *DeepLabV۳Plus* را با مدل از قبل آموزش دیده *ResNet-۵۰* در مجموعه *ImageNet* برای قطعه‌بندی معنایی صحنه‌های شهری پیاده‌سازی می‌کند. مجموعه داده مورد مطالعه در این تحقیق، مجموعه *UAVid۲۰۲۰*، یک مجموعه داده قطعه‌بندی معنایی پهپاد-مبنا در منطقه شهری از انجمن بین المللی فتوگرامتری و سنجش‌ازدور *(ISPRS)* است. برای ارزیابی عملکرد قطعه‌بندی معنایی روش پیشنهادی، از شبکه‌های عصبی *U-Net* و *Seg-Net* استفاده کردیم. نتایج بدست آمده از قطعه‌بندی معنایی تصاویر پهپاد-مبنا، اثربخشی چارچوب یادگیری انتقال پیشنهادی را نسبت به مدل‌های یادگیری عمیق نشان می‌دهد. از نظر معیار دقت کلی، معماری *DeepLabV۳Plus-ResNet۵۰* بهترین نتیجه را با ۸۱٫۹۳٪ دقت در مقایسه با شبکه‌های عصبی *U-Net* و *Seg-Net* با دقت-های ۷۴٫۳۵٪ و ۷۹٫۱۵٪ و کسب کرد.

کلید واژه‌ها : قطعه‌بندی معنایی، وسیله نقلیه هوایی بدون سرنشین، یادگیری انتقال، شبکه عصبی عمیق رمزگذار-رمزگشا پیچشی، *DeepLabV۳Plus*.

[*] نویسنده مکاتبه کننده: استان مرکزی، تفرش، کیلومتر ۵ جاده تهران، دانشگاه تفرش.

تلفن: ۰۸۶۲۲۳۶۷۴۳۰

Email: h.hasani@tafreshu.ac.ir

۱- مقدمه

در دهه اخیر به‌واسطه پیشرفت روش‌های بینایی-کامپیوتر[۱]، تجزیه و تحلیل داده‌های جمع‌آوری‌شده از سنجنده‌های هوابرد مانند دوربین‌های تصویربرداری/ویدئویی هوایی در بسیاری از رویکردهای پردازشی درک صحنه، نظیر طبقه‌بندی، قطعه‌بندی و قطعه‌بندی معنایی به موضوعی مهم و چالش‌برانگیز در انجمن فتوگرامتری و سنجش‌ازدور تبدیل شده است [۱]. بیش از دو دهه از معرفی شاخه قطعه‌بندی معنایی به عنوان یکی از مراحل اساسی پردازشی برای کاربرد-های مبتنی‌بر بینایی‌کامپیوتر می‌گذرد. یادگیری عمیق بخشی از رویکرد یادگیری ماشین است که در سال‌های اخیر توجه زیادی را از طرف محققین به خود جلب کرده است [۲]. قطعه‌بندی معنایی به عنوان بخشی از درک صحنه، با تقسیم تصاویر به اشیاء معنی‌دار، وظیفه برچسب‌گذاری هر قطعه از تصویر را به یک دسته از یک مجموعه دسته‌های از پیش تعریف شده دارد [۳]. براین اساس قطعه‌بندی به طور فزاینده‌ای در طیف وسیعی از کاربردها از جمله برنامه‌ریزی و توسعه شهری، پایش کاربری اراضی، تشخیص و مکان‌یابی اشیاء در تصاویر هوایی با وضوح بالا، استخراج خودکار ساختمان و جاده از تصاویر سنجش‌ازدور و یا هوایی، رانندگی خودکار وسایل نقلیه خودران، تشخیص بیابان-زایی، مدیریت منابع زمین و غیره کاربرد دارد [۴]. در سال‌های اخیر، پهپادها به طور گسترده برای کاربردهای مختلف پردازشی نظیر طبقه‌بندی و قطعه‌بندی معنایی در زمینه فتوگرامتری و سنجش‌ازدور مورد استفاده قرار گرفته است. پهپادها با برنامه‌های پروازی انعطاف‌پذیر و اخذ تصاویر با وضوح مکانی بالا، تصاویر را از ارتفاع به مراتب پایین اخذ کرده و همچنین امکان دسترسی، نظارت و تجزیه و تحلیل اطلاعات مفید را در مکان‌ها و زمان خاص برای کاربردهای حیاتی (بلایای طبیعی، برنامه‌ریزی و به‌روزرسانی اطلاعات شهری، نقشه‌برداری هوایی و غیره) به سرعت فراهم می‌آورند [۵].

توسعه شبکه‌های یادگیری عمیق و بهبود قابل توجه عملکرد آنها در سال‌های اخیر، قطعه‌بندی معنایی در درک صحنه‌های هوایی شهری را به موضوعی ضروری و در عین حال کاربردی تبدیل کرده است [۶]. قطعه-بندی می‌تواند به صورت نظارت‌شده یا بدون‌نظارت انجام شود. طیف گسترده‌ای از مطالعات در قطعه‌بندی معنایی و طبقه‌بندی داده‌های هوایی و سنجش‌ازدور منتشر شده است. این رویکردها را می‌توان بر اساس نوع داده‌های به کار رفته، روش مورد استفاده برای طبقه‌بندی قطعه‌ها، نوع قطعه‌بندی و روش تشخیص و ردیابی شی طبقه‌بندی کرد. در این بخش، مطالعات انجام‌شده در حوزه قطعه‌بندی معنایی صحنه‌های هوایی با استفاده از الگوریتم‌های سنتی یادگیری ماشین و دو روش یادگیری عمیق و یادگیری انتقال بررسی شده‌اند.

رویکردهای سنتی با توجه به داده ورودی طبقه‌بندی به دو دسته پیکسل-مبنا و شیء-مبنا تقسیم می‌شوند که در دسته اول مبنای دسته‌بندی پیکسل‌های تصویر و در دسته دوم، قطعات استخراج شده از تصویر هستند. در رویکردهای سنتی عملیات پردازش برروی داده‌های ورودی در دو مرحله مستقل استخراج ویژگی‌ها و طبقه‌بندی انجام می‌شود. برای این منظور، ابتدا ویژگی-هایی نظیر بافت از داده‌های ورودی استخراج شده و سپس فرآیند طبقه‌بندی توسط یکی از طبقه‌بندی-کننده‌های معمول انجام می‌شود. اگرچه این روش‌ها می‌توانند تصویر را به بخش‌های جداگانه تقسیم کنند، اما در نتایج آنها تعلق معنایی دقیقی به نواحی قطعه-بندی شده وجود ندارد [۷]. در این حوزه رویکردهای مختلفی برای استخراج ویژگی‌های مبتنی‌بر پیکسل توسعه داده شده است. به عنوان رویکردهای پیشین برای استخراج ویژگی‌های مکانی و بافتی از تصاویر، توصیف‌کننده‌های ریاضی متعددی نظیر تبدیل مستقل

[۱] *Computer vision*

از مقیاس ویژگی[۱] [۸]، ماتریس هم‌رویداد سطوح خاکستری[۲] [۹]، هیستوگرام گرادیان‌های شیب‌گرا[۳] [۱۰] و غیره مورد استفاده قرار گرفتند. همین‌طور رویکردهایی که طی دو دهه اخیر برای طبقه‌بندی ویژگی‌های استخراج‌شده ارائه شده‌اند، شامل طبقه‌بندی‌کننده‌هایی مانند ماشین بردار پشتیبان[۴] [۱۱] و جنگل‌های تصادفی[۵] [۱۲] و غیره بودند که با استفاده از ارتباطات بین دسته‌ها و ایجاد پیوند بین پیکسل‌های هم‌جوار قطعه‌بندی را انجام می‌دادند.

استرجس[۶] و همکاران (۲۰۰۹) رویکردی را براساس چارچوب مدل احتمالی میدان تصادفی شرطی[۷] *(CRF)* به منظور ادغام ویژگی‌های مبتنی بر حرکت و ظاهر (رنگ، مکان و توصیفگرهای هیستوگرام گرادیان شیب-گرا) برای قطعه‌بندی پیکسل-مبنای جاده ارائه کردند. آنها برای تمایز ویژگی‌های هر دسته از یک رویکرد تقویت‌کننده با هدف تعیین مرزهای دقیق جاده در نقشه قطعه‌بندی استفاده کردند [۱۳]. لالیبرت[۸] و همکاران (۲۰۰۹) از رویکرد شیء-مبنا برای تحلیل تصاویر با حدتفکیک بالای پهپاد استفاده کردند. با افزایش حدتفکیک مکانی، حدتفکیک طیفی کاهش یافته و به منظور جبران آن در این مطالعه از آنالیز بافت استفاده شده است. در هر مقیاس قطعه‌بندی، ویژگی‌های بافت بهینه برمبنای درخت تصمیم‌گیری انتخاب شدند *[۱۴]*. ژانگ[۹] و همکاران (۲۰۱۰) در چارچوبی برپایه نقشه‌های عمق متراکم به تجزیه و تحلیل صحنه‌های معنایی و تشخیص شیء پرداختند. آنها پنج ویژگی سه‌بعدی مستقل از دید (عمود سطح، ارتفاع از سطح زمین، مسطح بودن سطح محلی، مسطح بودن سطح همسایه و فاصله تا مسیر دوربین) که با دسته شی متفاوت هستند را از نقشه‌های عمق متراکم استخراج کرده و با استفاده از روش جنگل تصادفی، ویژگی‌ها استخراجی را طبقه‌بندی کردند [۱۵]. موراندوتزو[۱۰] و ملگانی[۱۱] (۲۰۱۴) روشی را با استفاده از جستجوی پنجره لغزان در تصاویر پهپاد برای تشخیص وسایل نقلیه پیشنهاد کردند. این روش برمبنای استخراج ویژگی‌های هیستوگرام شیب‌های جهت‌دار با استفاده از عملیات فیلتر کردن در جهت افقی و عمودی بود و خودروها پس از محاسبه میزان شباهت برمبنای مشخصه خودروهای مرجع، استخراج شدند [۱۶].

از مزایای روش‌های سنتی می‌توان به استخراج ویژگی-های معنادار توسط کاربر و استفاده از طبقه‌بندی کننده‌های گوناگون اشاره کرد. هر چند در این رویکردها، مرحله استخراج ویژگی با چالش‌هایی نظیر ایجاد فضای ویژگی با ابعاد بالا، وجود ویژگی های اضافی و وابسته مواجه خواهد بود. همچنین این روش‌ها با توجه به استخراج دستی ویژگی‌ها، فقدان اطلاعات سلسله مراتبی و زمان‌بر بودن فرآیند پردازش، نمی‌توانند نتیجه رضایت‌بخشی به‌دست آورند که منجر به راهکاری کم‌دقت، ناکارآمد، کند و انسجام پایین در بین مراحل مختلف می‌شوند [۱۷]. به منظور حل این مسائل، روش‌های یادگیری عمیق که جزء روش‌های نظارت‌شده در قطعه‌بندی معنایی هستند، توجه محققان را به خود جلب کردند. با توسعه حوزه یادگیری عمیق، همواره تلاش بر این بوده که چالش-های موجود را برطرف کرده و استخراج ویژگی‌ها به صورت خودکار در حین عملیات آموزش توسط

---

۱ *Scale Invariant Feature Transform (SIFT)*

۲ *Gray Level Co-ocurrance Matrix (GLCM)*

۳ *Histigram of Gradient (HoG)*

۴ *Support Vector Machines (SVMs)*

۵ *Random Forests (RF)*

۶ *Sturgess*

۷ *Conditional Random Field (CRF)*

۸ *Laliberte*

۹ *Zhang*

۱۰ *Moranduzzo*

۱۱ *Melgani*

الگوریتم یادگیری عمیق، در یک فرآیند پایان‌به‌پایان[۱] صورت گیرد [۱۸]. این رویکردها در مقایسه با روش‌های سنتی، به دلیل استخراج اطلاعات معنایی، روشی مؤثر برای قطعه‌بندی ویژگی‌ها هستند. اما یک چالش مهم و تاثیرگذار در رویکردهای مبتنی‌بر یادگیری عمیق، محدودیت دسترسی به مجموعه داده‌های آموزشی و برچسب‌گذاری‌شده متناظر است. یک مدل یادگیری عمیق نظارت‌شده معمولاً به تعداد زیادی نمونه‌های آموزشی برچسب‌گذاری‌شده نیاز دارد. در زمینه فتوگرامتری و سنجش از دور، برچسب‌زدن به داده‌های استخراج شده در تهیه نمونه‌های آموزشی برای هر دسته بسیار وقت‌گیر، پرهزینه و مستعد خطا است، و اغلب به مهارت بالای متخصص انسانی و بازدیدهای میدانی نیاز دارند [۱۹]. آموزش شبکه‌های عصبی عمیق به منابع محاسباتی عظیم، مقدار قابل توجهی از داده‌های برچسب‌گذاری‌شده و زمان آموزش زیاد برای همگرایی نیاز دارد که در عمل دستیابی به آن‌ها نیازمند سخت‌افزارهای پردازشی قدرتمند است. یک روش کاربردی در حل این چالش، استفاده از روش یادگیری انتقال[۲] است [۲۰]. هدف یادگیری انتقال، کاهش زمان آموزش، ارائه عملکردی بهتر برای آموزش شبکه‌های عصبی عمیق و کاهش نیاز به تعداد زیادی از مجموعه داده‌های برچسب‌گذاری شده با استفاده از انتقال وزن‌های یک شبکه از قبل‌آموزش‌دیده و تنظیم دقیق آن برای کار موردنظر در حوزه هدف به واسطه مجموعه داده‌های جدید است[۲۱]. تنظیم و مقداردهی اولیه وزن‌های شبکه عصبی با استفاده از یک مدل از قبل‌آموزش‌دیده به جای آموزش شبکه عصبی از ابتدا با وزن‌دهی اولیه تصادفی می‌تواند شروع مطلوب و در عین حال سریع‌تری را برای شبکه عصبی عمیق فراهم کند و نرخ همگرایی شبکه عصبی را بهبود ببخشد. در این حوزه، به دلیل ظرفیت بالای یادگیری ویژگی‌ها، مدل‌های قطعه‌بندی معنایی مبتنی‌بر شبکه‌های عصبی از قبل آموزش‌دیده، نظیر شبکه‌های عصبی پیچشی عمیق، معماری‌های رمزگذار-رمزگشا، شبکه‌های عصبی خودرمزگذار و غیره کارایی بالایی را در سال‌های اخیر کسب کرده‌اند [۲۲]. در ادامه، مروری از تحقیقات انجام شده در زمینه قطعه‌بندی معنایی نواحی شهری به‌واسطه رویکردهای یادگیری عمیق ارائه می‌شود.

زنجانی و ون‌گرون[۳] (۲۰۱۶) مطالعه‌ای را با هدف درنظر گرفتن اطلاعات پویای صحنه‌های شهری به منظور افزایش دقت در قطعه‌بندی معنایی تصاویر انجام دادند. آنها برای استخراج اطلاعات پویا از جریان نوری متراکم صحنه استفاده کردند. در رویکرد پیشنهادی این مطالعه از شبکه عصبی پیچشی *DeepLab* به عنوان طبقه‌بندی کننده برچسب پیکسل‌ها و به دنبال آن از مدل میدان تصادفی شرطی کاملاً متصل به عنوان مرحله پس‌پردازشی (اعمال جریان نوری صحنه) استفاده شده است [۲۳]. وی[۴] و همکاران (۲۰۱۸) به منظور برچسب گذاری پیکسل-مبنا معنایی تصاویر ماهواره‌ای یک مدل رمزگذار-رمزگشا پیچشی عمیق پیشنهاد کردند. در معماری پیشنهادی از واحد پیچشی افزایش‌یافته چند نرخی[۵] برای به دست آوردن ویژگی اشیاء چند مقیاسه استفاده شد. پیچش‌های افزایش‌یافته[۶] بدون نیاز به افزایش حافظه محاسباتی و زمان، به طور مؤثر دامنه پذیرش فیلتر لایه‌های پیچشی را برای ادغام بیشتر اطلاعات زمینه‌ای افزایش می‌دهند. همین‌طور برای بهبود برچسب‌گذاری (به‌ویژه پیکسل‌های اطراف مرز طبقه‌بندی) از مدل میدان تصادفی شرطی به عنوان

[۱] *End to end*

[۲] *Transfer Learning*

[۳] *van Gerven*

[۴] *Wei*

[۵] *Multi-rate dilated convolution unit (MRDC)*

[۶] *Dilated Convolution*

لایه کاملاً متصل در شبکه استفاده شد [۲۴]. لی[۱] و همکاران (۲۰۱۹) در چارچوب یک مدل یادگیری عمیق پیچشی با استفاده از ترکیب شبکه رمزگذار-رمزگشا *U*-شکل و ساختار ادغام هرم فضایی به‌عنوان اتصالی بین مسیر رمزگذار و رمزگشا مدلی را تحت عنوان مدل ادغام هرم فضایی *U*-شکل[۲] برای قطعه-بندی معنایی ساختمان در تصاویر با وضوح بالا ماهواره-ای ارائه کردند. مدل پیشنهادی از بخش ادغام هرم فضایی برای استخراج اطلاعات زمینه‌ای چند مقیاسه و از شبکه رمزگذار-رمزگشا برای بازیابی اطلاعات از دست رفته استفاده می‌کند [۲۵]. گیریشا[۳] و همکاران (۲۰۱۹) به موضوع قطعه‌بندی معنایی فریم‌های ویدئویی پهپاد با استفاده از شبکه‌های عصبی پیچشی[۴] پرداختند. روش پیشنهاد شده، با هدف ارزیابی عملکرد الگوریتم‌های قطعه‌بندی معنایی مبتنی‌بر شبکه‌های عصبی پیچشی (شبکه عصبی کاملا پیچشی[۵] *(FCN)* و معماری *U-Net*) انجام شد [۱]. لی[۶] و همکاران (۲۰۲۰) در پژوهشی با طراحی یک شبکه افزایش چند مقیاسه، قطعه‌بندی معنایی تصاویر پهپاد-مبنا شهری را با استفاده از استخراج ویژگی‌های چند مقیاسه انجام دادند. در معماری پیشنهادی از روش بهینه‌سازی فضای ویژگی[۷] برای استخراج اطلاعات توالی داده‌ها و از مدل کاملا متصل میدان تصادفی شرطی سه بعدی [۸] برای پیش‌بینی برچسب نهایی استفاده کردند. این روش هر دو حوزه‌های مکانی و زمانی را با بهره‌گیری از اطلاعات جریان‌نوری و ردیابی‌ها در قطعه‌بندی معنایی تصاویر پوشش می‌دهد. مدل پیشنهادی عملکرد مناسبی در حل چالش اثرات تغییر مقیاس در تصاویر هوایی مایل نسبت به شبکه‌های عصبی رایجی همچون *Dilation Net*، *U-Net* و *FCN-8s* ارائه داد [۲۶].

قطعه‌بندی معنایی تصاویر هوایی نواحی شهری در چارچوب یادگیری عمیق نسبت به روش‌های سنتی یادگیری ماشین از دقت بالاتری برخودار است، اما فرآیند آموزش مدل‌های عصبی عمیق در حوزه تصاویر هوایی با چالش‌هایی نظیر محدودیت در منابع محاسباتی و همین‌طور محدودیت در داده‌های برچسب-دار آموزشی برای قطعه‌بندی مناسب و دقیق تصاویر روبرو است. بنابراین، آموزش شبکه‌های عصبی پیچشی عمیق برمبنای وزن‌دهی اولیه تصادفی، علاوه بر اینکه زمان‌بر است، برای همگرایی و عملکردی مطلوب به نمونه‌های آموزشی برچسب‌دار زیادی نیاز دارد [۲۷]. از این‌رو در دسته سوم مطالعات، رویکردهای یادگیری انتقال ارائه شده‌اند. هدف یادگیری انتقال، کاهش نیاز به حجم بالای داده‌های برچسب‌گذاری‌شده برمبنای انتقال وزن یک شبکه از قبل آموزش‌دیده و تنظیم دقیق آن برای کار جدید است. تنظیم دقیق کل شبکه با یادگیری انتقال عموماً بسیار سریع‌تر و ساده‌تر از آموزش شبکه با وزن‌های اولیه تصادفی است [۲۸]. یادگیری انتقال زمانی مفید است که برای آموزش یک شبکه عصبی عمیق داده‌های کافی در حوزه هدف موجود نباشد و برای کار مشابه دیگر حجم زیادی از داده‌ها با قابلیت انتقال به مسئله هدف، موجود باشد؛ یا زمانی‌که مدلی وجود داشته که قبلاً بر روی داده‌های مشابه حوزه هدف آموزش‌دیده باشد. با این حال، حتی اگر داده‌های کافی در حوزه منبع برای آموزش یک مدل از ابتدا وجود داشته باشد و وظایف حوزه منبع و هدف به هم مرتبط نباشند، مقداردهی اولیه پارامترها با استفاده از یک مدل از قبل آموزش داده شده، نقطه شروع بهتری نسبت به مقداردهی اولیه تصادفی برای همگرایی نهایی حوزه هدف است [۲۹]. در ادامه مروری

---

[۱] *Liu*

[۲] *U-Shaped Spatial Pyramid Pooling (USPP)*

[۳] *Girisha*

[۴] *Convolutional Neural Networks (CNNs)*

[۵] *Fully Convolutional Network (FCN)*

[۶] *Lyu*

[۷] *Feature Space Optimization (FSO)*

[۸] *۳D Conditional Random Field (۳D CRF)*

بر برخی از مطالعات انجام شده در این حوزه برای قطعه‌بندی معنایی تصاویر هوایی نواحی شهری شده است.

یو[۱] و همکاران (۲۰۱۸) یک چارچوب پایان‌به‌پایان برای قطعه‌بندی معنایی تصاویر هوایی با وضوح مکانی بالا نواحی شهری پیشنهاد کردند. معماری پیشنهادی، ترکیبی از چارچوب یادگیری باقیمانده و ساختار ادغام هرم برای آموزش ساده‌تر مدل و استخراج ویژگی‌های چندمقیاسه است. چارچوب یادگیری باقیمانده، یک شبکه پیچشی از قبل آموزش‌دیده مبتنی‌بر معماری *ResNet۱۰۱-v۲* است که برروی پایگاه داده *ImageNet* آموزش داده شده است. ساختار ادغام هرم، یک نسخه بهبودیافته از معماری *PSPNet*، برای استخراج ویژگی‌ها در مقیاس‌های مختلف است. نتایج قطعه‌بندی نهایی رویکرد پیشنهادی، با استفاده از یک عملیات پیچشی بر روی نقشه‌های ویژگی به‌دست می‌آید [۳۰]. ژانگ[۲] و همکاران (۲۰۱۹) در مطالعه‌ای با طراحی یک شبکه رمزگشا پیچشی چندمقیاسه به مسئله قطعه‌بندی پیکسل-مبنا معنایی تصاویر هوایی پرداختند. معماری پیشنهادی از یک رمزگذار (شبکه از قبل آموزش دیده *VGG۱۶*) برای استخراج ویژگی‌های معنایی خام تصاویر و یک رمزگشا نامتقارن مبتنی‌بر شبکه کاملاً پیچشی تشکیل شده است. بخش رمزگشا از ترکیب سه جزء پیچش‌های افزایش‌یافته، پیچش‌های وارون[۳] و لایه حداقل ادغام [۴]تشکیل شده است. شبکه طراحی شده در این کار، با توجه به رویکرد انتقال دانش و معماری گسترده‌ای که داشت در مقایسه با شبکه‌های *FCN*، *SegNet*، و *U-net*، عملکرد بهتری در قطعه‌بندی تصاویر هوایی ارائه کرد [۳۱].

پانبونیئن[۵] و همکاران (۲۰۱۹) در چارچوب یادگیری انتقال با ارائه یک شبکه پیچشی سراسری[۶] به مسئله قطعه‌بندی معنایی چند شیئی نواحی شهری در تصاویر ماهواره‌ای *Vaihingen* از پایگاه داده *ISPRS* پرداختند. روش پیشنهادی متشکل از دو بخش بود، نخست طراحی یک شبکه پیچشی سراسری با مقداردهی وزن-های اولیه شبکه از طریق مدل‌های از قبل آموزش‌دیده (*ResNet-۵۰*، *ResNet-۱۰۱* و *ResNet-۱۵۲*) و افزودن تعداد لایه‌های بیشتر برای استخراج ویژگی‌های چند مقیاسه با وضوح‌های مختلف و سپس استفاده از رویکرد "کانال توجه[۷]" برای تعیین وزن‌های مناسب هر ویژگی استخراج شده در مراحل مختلف معماری پیشنهادی است. در نهایت، نتایج نشان داد که روش پیشنهادی نسبت به شبکه‌های معمول از جمله رمزگذار-رمزگشا پیچشی عمیق از عملکرد بهتری در قطعه‌بندی معنایی برخوردار است [۳۲].

لی[۸] و همکاران (۲۰۲۰) یک روش یادگیری انتقال عمیق را بر اساس شبکه عصبی کاملاً پیچشی (*FCN*) برای قطعه‌بندی معنایی نواحی شهری با استفاده از مجموعه تصاویر ماهواره‌ای *Vaihingen* پیشنهاد کردند. در این مطالعه، پارامترهای شبکه عصبی *PSPNet-s* با تصاویر پهپاد آموزش دیده و همین‌طور مدل رقومی متناظر سطح[۹] (*DSM*) را برای قطعه‌بندی تصاویر سنجش از دور هوابرد انتقال دادند. روش پیشنهادی از دقت کلی بالاتر و زمان اجرای کمتری نسبت به مدلی که مستقیماً توسط تصاویر سنجش‌ازدور آموزش داده شده، برخوردار است [۳۳]. ژانگ و همکاران (۲۰۲۲) با استفاده از یادگیری انتقال و مدل *FT-ResNet۵۰* به

---

۱ *Yu*

۲ *Zhang*

۳ *Transposed Convolution*

۴ *Unpooling*

۵ *Panboonyuen*

۶ *Global Convolutional Network (GCN)*

۷ *Channel Attention*

۸ *Liu*

۹ *Digital Surface Model (DSM)*

شناسایی آتش در مناطق جنگلی با استفاده از تصاویر پهپاد پرداختند. نتایج بدست آمده بیانگر دقت بالاتر روش پیشنهادی نسبت به شبکه *ResNet*۵۰ و *VGG*۱۶ است [۳۴].

در یک دهه اخیر، به دلیل ظهور مجموعه‌هایی بزرگ-مقیاس از داده‌های هوایی با محدودیت در برچسب-گذاری، روش‌های سنتی یادگیری ماشین و مدل‌های یادگیری عمیق تحت تأثیر چالش‌های فضای استخراج ویژگی و کمبود داده‌های برچسب‌دار قرار گرفتند و در عین حال سرعت و دقت محاسباتی پایینی را برای همگرایی نمونه‌های آموزشی ارائه دادند. این رویکردها برای قطعه‌بندی معنایی دقیق به حجم زیادی از داده-های آموزشی برچسب‌دار و همچنین دستگاهی با منابع محاسباتی بالا نیاز دارند که دستیابی به آنها سخت، زمان‌بر، نیازمند نیروهای انسانی متخصص و از نظر اقتصادی مقرون به صرفه نیست. در سال‌های اخیر، به واسطه توسعه یادگیری انتقال و کارایی قابل توجه آن در انتقال دانش از حوزه منبع به حوزه هدف، در مقایسه با روش‌های سنتی یادگیری به یکی از موضوعات مورد علاقه محققان در زمینه های مختلف علمی تبدیل شد.

در این مطالعه، هدف قطعه‌بندی معنایی تصاویر مایل پهپاد صحنه‌های شهری از دیدگاه رویکرد یادگیری انتقال، به‌واسطه انتقال وزن‌های یک شبکه ازقبل آموزش‌دیده *(ResNet-۵۰)* و استفاده از معماری رمزگذار-رمزگشا پیچشی *DeepLabV۳Plus* است. *DeepLabV۳Plus* شبکه‌ای پیچشی با دو بخش رمزگذار-رمزگشا است که برای قطعه‌بندی معنایی تصاویر، ابتدا ویژگی‌های تصویر توسط بخش رمزگذار استخراج و کدگذاری می‌شوند. سپس، توسط بخش رمزگشا ویژگی‌های استخراج شده رمزگشایی و با نقشه-های ویژگی کم‌بعد الحاق می‌شوند. در نهایت، بردارهای پیش‌بینی هر دسته از طریق یک لایه کاملاً متصل طبقه‌بندی ارائه می‌شوند. معماری پیشنهادی با توجه به تعداد بالای تصاویر پهپاد در پوشش مناطق وسیع و محدودیت در دسترسی به نمونه‌های برچسب‌دار آموزشی زیاد در نظر گرفته شده است.

۲- روش پیشنهادی

روش پیشنهادی متشکل از چهار مرحله اصلی (آماده-سازی مجموعه داده ورودی، پیش‌پردازش، طراحی و آموزش شبکه عصبی و ارزیابی شبکه آموزش دیده برروی داده آزمایشی) است. در فلوچارت شکل (۱) فرآیند پیاده‌سازی هر مرحله با جزئیات نشان داده شده است.

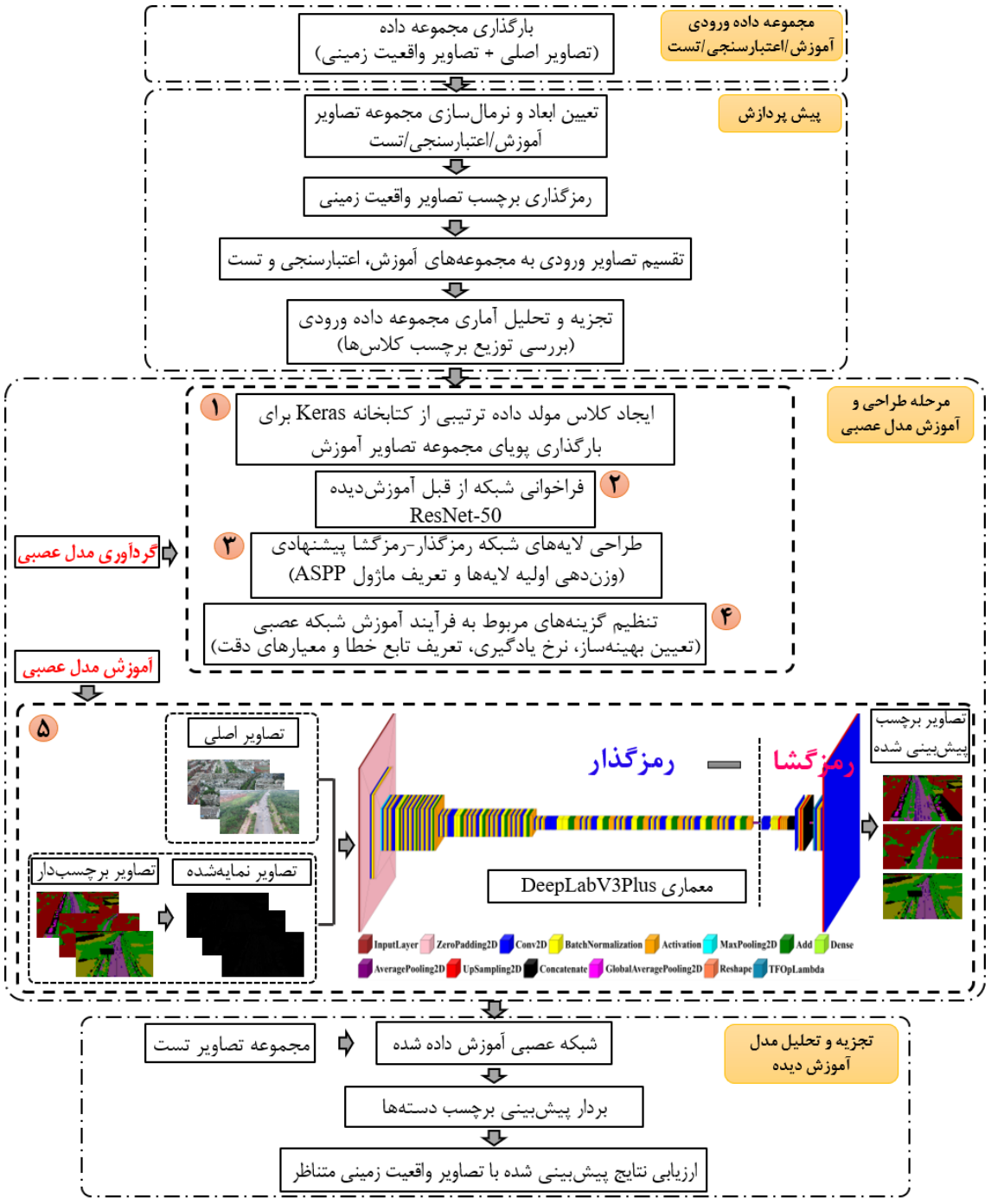
پیاده‌سازی رویکرد پیشنهادی در چهار مرحله به شرح زیر است:

(الف) آماده‌سازی مجموعه داده ورودی (آموزش/اعتبارسنجی/آزمایشی): متشکل از تصاویر اصلی و واقعیت زمینی متناظر از مجموعه داده قطعه-بندی معنایی پهپاد-مبنا نواحی شهری *UAVid*۲۰۲۰[۱] است [۲۶].

(ب) پیش‌پردازش: انجام تنظیمات و پیش‌پردازش‌های اولیه برروی مجموعه تصاویر ورودی به منظور آماده-سازی آنها برای آموزش شبکه عصبی از قبیل تعیین اندازه تصاویر ورودی، نرمال‌سازی تصاویر، رمزگذاری برچسب تصاویر واقعیت زمینی، تقسیم داده‌ها به آموزش/آزمایشی/ اعتبارسنجی و تجزیه و تحلیل آماری توزیع برچسب دسته‌ها.

مرحله (ج) و (د) پس از شکل(۱) آورده شده است.

[۱] *https://uavid.nl/*



شکل ۱: فلوچارت روش پیشنهادی

(ج) طراحی و آموزش شبکه عصبی: این مرحله، مربوط به تعیین و تنظیم پارامترهای آموزش شبکه عصبی، طراحی و وزن‌دهی اولیه لایه‌ها و عملیات آموزش شبکه است که طی آن ویژگی‌های مکانی مجموعه تصاویر ورودی استخراج و درنهایت به یک طبقه‌بندی‌کننده داده می‌شوند. برای این منظور از شبکه عصبی باقیمانده ۵۰-*ResNet* که از قبل برروی مجموعه *ImageNet* آموزش‌داده شده است، برای انتقال دانش و

وزن‌دهی اولیه لایه‌های شبکه عصبی پیشنهادی استفاده می‌کنیم. در نهایت فرآیند آموزش مدل عصبی برای قطعه‌بندی معنایی تصاویر، به‌واسطه مراحل استخراج و طبقه‌بندی ویژگی‌ها در حین عملیات آموزش توسط الگوریتم یادگیری انتقال به صورت پایان‌به‌پایان انجام می‌شود.

(د) ارزیابی عملکرد مدل پیشنهادی: در این مرحله مجموعه تصاویر آزمایشی به شبکه عصبی آموزش‌دیده داده می‌شوند و در خروجی، بردار پیش‌بینی برچسب دسته‌ها بدست می‌آیند. نتایج پیش‌بینی شده حاصل از قطعه‌بندی معنایی تصاویر آزمایشی را با نقشه‌های واقعیت زمینی متناظر مقایسه و معیارهای ارزیابی قطعه‌بندی (دقت[۱]، صحت[۲]، بازیابی[۳]، امتیاز F_1[۴]، نسبت اشتراک به اجتماع[۵] *(IOU)* هردسته، میانگین معیار *IOU* (*MeanIOU*) و میانگین *BFScore*[۶] *(MeanBFScore)*) محاسبه می‌شود.

در این مطالعه، یک چارچوب برچسب‌گذاری معنایی برای قطعه‌بندی صحنه‌های شهری از تصاویر هوایی با وضوح بالا مبتنی‌بر پهپاد پیشنهاد شده است. مدل پیشنهادی یک شبکه عصبی پایان‌به‌پایان با ساختار یادگیری انتقال مبتنی‌بر معماری رمزگذار-رمزگشا پیچشی *DeepLabV۳Plus* است. با توجه به موفقیت شبکه‌های رمزگذار-رمزگشا در کاربردهای سنجش از دور و همچنین ارزیابی‌های مختلف انجام شده، این شبکه به عنوان روش پیشنهادی در نظر گرفته شد. یک نمای کلی از معماری طراحی شده در شکل (۲) نشان داده شده است.

به طورکلی *DeepLabV۳Plus* متشکل از دو زیرشبکه رمزگذار-رمزگشا به همراه بخش ادغام هرم فضایی متخلخل[۷] است. بخش ادغام هرم فضایی در زیرشبکه رمزگذار وظیفه اخذ اطلاعات زمینه‌ای چند-مقیاس را با ادغام ویژگی‌ها در وضوح‌های مختلف برعهده دارد. در عین حال ساختار رمزگشا عملکرد قطعه‌بندی را با تمرکز برروی اخذ دقیق‌تر مرزهای اشیاء از تصاویر ورودی بهبود می‌بخشد. زیرشبکه رمزگذار پیشنهادی، ترکیبی از معماری پیچشی ازقبل آموزش‌دیده *ResNet*-۵۰ با ساختار باقیمانده و بخش ادغام هرم فضایی متخلخل با نرخ افزایش متغیر است که امکان کدگذاری اطلاعات چند-مقیاسه و افزایش عملکرد شبکه را فراهم می‌کند. شبکه پیچشی از قبل آموزش‌دیده در این معماری با هدف انتقال یادگیری و وزن‌دهی اولیه لایه‌ها برای بهبود فرآیند آموزش شبکه عصبی معرفی می‌شود. به دنبال آن بخش ادغام هرم فضایی متخلخل ویژگی‌های چند-مقیاسه را به واسطه مجموعه‌ای از لایه‌های پیچشی متخلخل موازی شامل یک پیچش (۱×۱)، سه پیچش (۳×۳) با نرخ‌های افزایش ۶، ۱۲، ۱۸ و یک شاخه ادغام میانگین استخراج می‌کند. برای نرمال‌سازی داده‌ها، پس از هر عملیات ادغام از لایه‌های نرمال‌ساز دسته‌ای[۸] و فعال‌ساز یکسوسازی‌شده خطی[۹] *(Relu)* استفاده شده است. سپس، نقشه‌های ویژگی استخراج شده توسط بخش ادغام هرم فضایی متخلخل، به هم الحاق شده، و به یک لایه پیچشی با ابعاد (۱×۱) داده می‌شوند. در نهایت نقشه ویژگی چند-مقیاسه پیچشی به عنوان خروجی رمزگذار ارائه می‌شود.

از سوی دیگر، زیرشبکه رمزگشا با ترکیبی از ویژگی-های سطح پایین و سطح بالا پیاده‌سازی شده است.

---

[۱] *Accuracy*

[۲] *Precision*

[۳] *Recall*

[۴] F_1*Score*

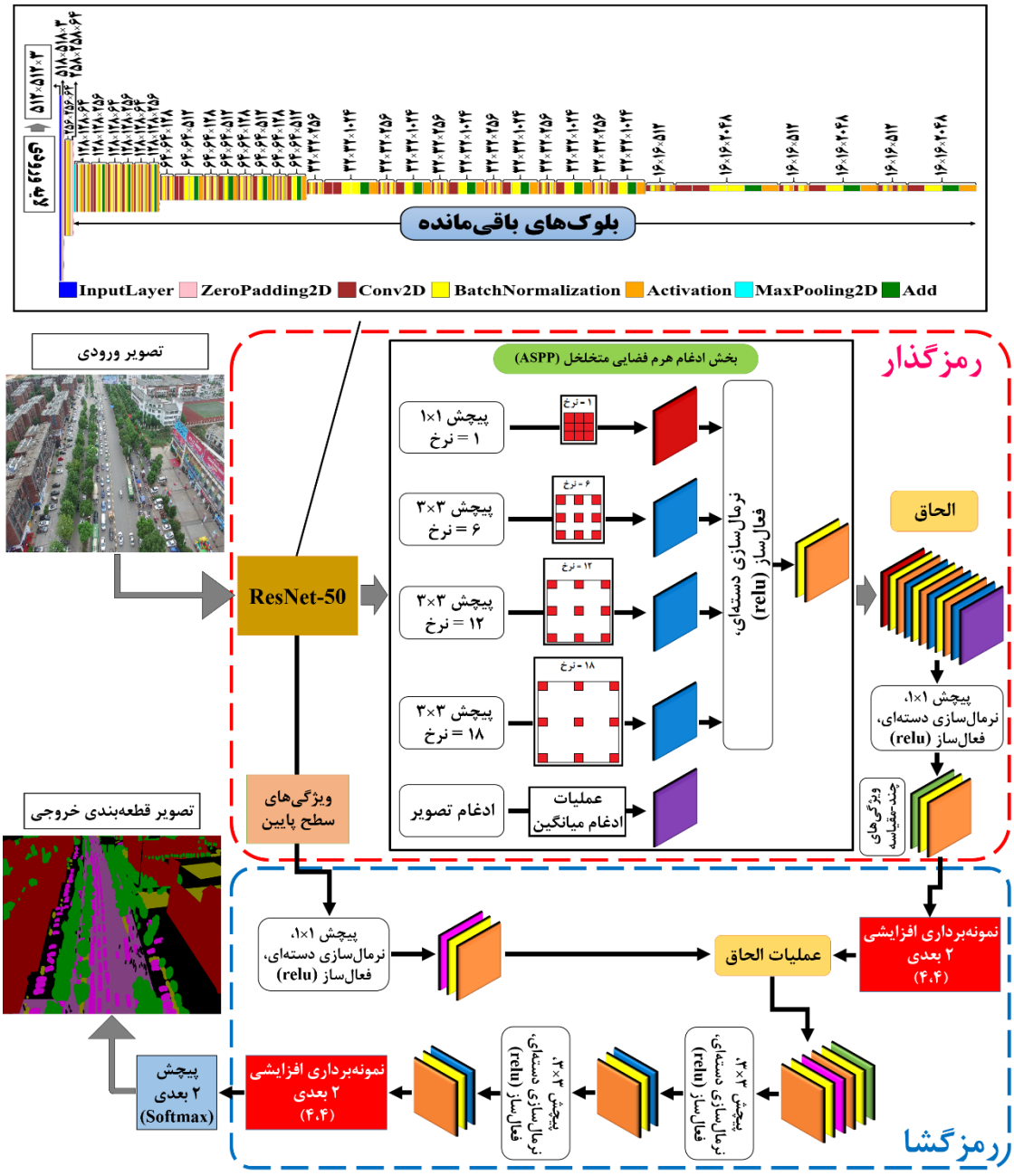
[۵] *Intersection-over-union (IOU)*

[۶] *Mean Boundary* F_1*Score*

[۷] *Atrus sence pyramid pooling (ASPP) module*

[۸] *Batch normalization (BN)*

[۹] *Rectified linear unit (Relu)*



شکل ۲: معماری *DeepLabV۳Plus* با شبکه از قبل آموزش‌دیده *ResNet-۵۰*

در بخش رمزگشا، ابتدا ویژگی‌های چند-مقیاسه حاصل از بخش رمزگذار به واسطه یک درونیابی دوخطی با ضریب چهار، نمونه‌برداری افزایشی می‌شوند. سپس با ویژگی‌های سطح پایین استخراج‌شده از شبکه عصبی *ResNet-۵۰*، به هم الحاق می‌شوند. در این مرحله به منظور متعادل کردن اهمیت بین ویژگی‌های سطح پایین و ویژگی‌های معنایی سطح بالا چند-مقیاسه حاصل از رمزگذار، از یک لایه پیچشی (۱×۱)، نرمال-ساز دسته‌ای و فعال‌ساز *Relu* برای کاهش تعداد کانال-ها استفاده شده است. پس از عملیات الحاق، از لایه‌های

پیچشی (۳×۳)، نرمال‌ساز دسته‌ای و تابع فعال‌ساز *Relu* برای اصلاح و رمزگشایی ویژگی‌ها، و درنهایت از یک لایه نمونه‌برداری افزایشی دوخطی با ضریب چهار و به دنبال آن یک لایه پیچشی سافت-مکس[۱] *(SoftMax)* با تعداد کانال شش برای دستیابی به نقشه قطعه‌بندی با ابعاد تصویر اصلی ورودی، استفاده می‌شود.

جزئیات مربوط به شبکه رمزگذار *ResNet*-۵۰، بخش ادغام هرم فضایی متخلخل و زیرشبکه رمزگشا معماری پیشنهادی در بخش (۲-۱) ارائه شده است.

۲-۱- معماری رمزگذار-رمزگشا پیچشی *DeepLabV۳Plus*

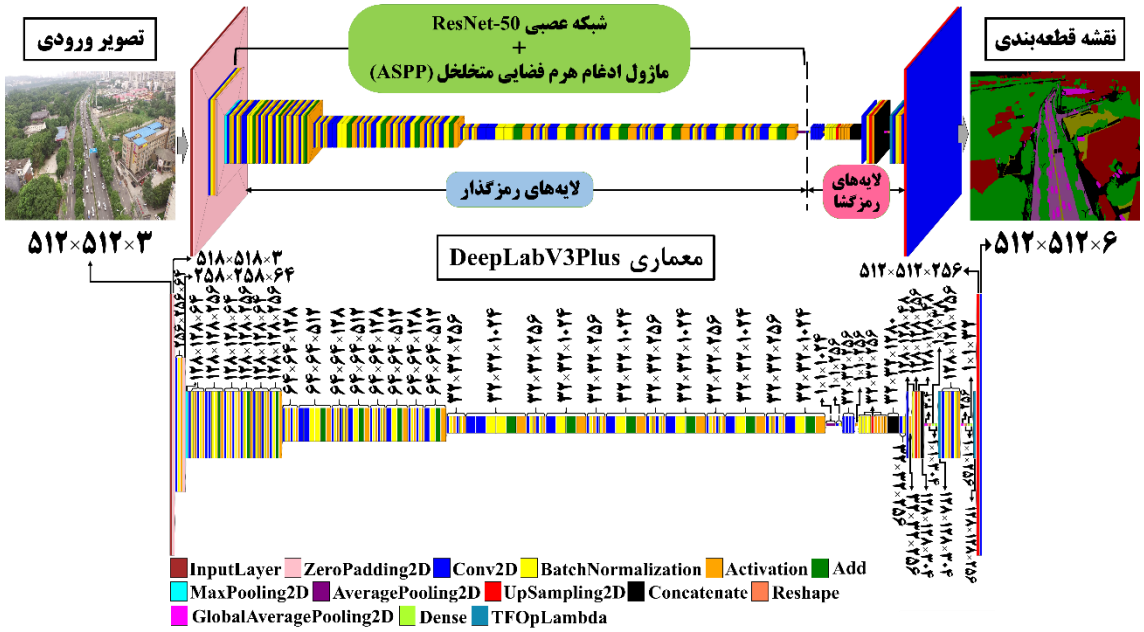
در این بخش، معماری شبکه پیشنهادی برای قطعه-بندی معنایی شش دسته شی مختلف از تصاویر هوایی شهری با وضوح بالا ارائه داده می‌شود. همانطور که در شکل (۲) نشان داده شده است، ما از معماری *DeepLabV۳Plus* به عنوان شبکه اصلی و از انتقال-دهنده *ResNet*-۵۰ آموزش‌دیده برروی پایگاه داده *ImageNet* به عنوان مدل از قبل آموزش‌دیده استفاده کردیم. *DeepLabV۳Plus* یک شبکه پیچشی مبتنی بر معماری رمزگذار-رمزگشا برای قطعه‌بندی معنایی است. زیرشبکه رمزگذار به‌واسطه معماری پیچشی و بخش ادغام هرم فضایی متخلخل، ویژگی‌های مکانی چند-مقیاسه را از داده‌های آموزشی ورودی استخراج می‌کند و زیرشبکه رمزگشا بردار برچسب کلاس‌های داده را از نقشه‌های ویژگی رمزگذاری شده به‌واسطه نمونه‌برداری افزایشی بازیابی می‌کند. همانطورکه در شکل (۳) نشان داده شده است، ساختار اصلی شبکه *DeepLabv۳Plus* از سه بخش تشکیل شده است. بخش اول مربوط به معماری رمزگذار پیچشی از قبل آموزش دیده *ResNet*-۵۰ و وزن‌دهی اولیه لایه‌های شبکه عصبی پیشنهادی است. در این معماری آخرین لایه کاملاً متصل را حذف کرده و بقیه لایه‌های *ResNet*-۵۰ به عنوان یک استخراج‌کننده ویژگی چند-مقیاسه تلقی می‌شود. بخش دوم مبتنی‌بر ساختار ادغام هرم فضایی متخلخل است که متشکل از چهار لایه پیچشی متخلخل با نرخ-های افزایش متغیر برای استخراج اطلاعات زمینه‌ای چند-مقیاسه است. بخش آخر مربوط به ساختار رمزگشا برای بازیابی ویژگی‌های سطح پایین و سطح بالا رمزگذاری‌شده است که به‌واسطه آن ویژگی‌های هر شاخه از طریق عملیات نمونه‌برداری افزایشی در یک نقشه ویژگی واحد رمزگشایی می‌شود.

این معماری شامل لایه ورودی با ابعاد (۳×۵۱۲×۵۱۲) و یک لایه‌گذاری صفر با حفظ ابعاد تصویر ورودی است. اولین لایه پیچشی رمزگذار، متشکل از یک عملیات پیچشی با ۶۴ هسته (۷×۷) با گام دو و به دنبال آن لایه نرمال‌سازی دسته‌ای، تابع فعال‌ساز *Relu* و یک حاشیه‌گذاری صفر[۲] است. پس از آن، یک لایه حداکثر ادغام دوبعدی[۳] با اندازه هسته (۳×۳)، گام دو اعمال شده است. بعد از عملیات حداکثر ادغام، تعریف لایه-های رمزگذار *ResNet*-۵۰ که از سه بخش تشکیل شده، وجود دارد. بخش اول شامل شش بلوک باقی-مانده، بخش دوم شامل هشت بلوک باقی‌مانده و بخش سوم شامل ۱۲ بلوک باقی‌مانده است. هر بلوک باقی-مانده از دو عملیات پیچشی متوالی به همراه لایه‌های نرمال‌ساز دسته‌ای و فعال‌ساز *Relu* تشکیل شده است.

[۱] *SoftMax*

[۲] *Zero Padding*

[۳] *MaxPooling۲D*



شکل ۳: معماری *DeepLabV۳Plus* طراحی‌شده (ابعاد تصاویر ورودی ۵۱۲×۵۱۲×۳ و ابعاد نقشه قطعه‌بندی خروجی ۵۱۲×۵۱۲×۷).

ابعاد ورودی در اولین بلوک‌های هر بخش از نظر (عرض و ارتفاع) با گام دو کاهش می‌یابد. تمام بلوک‌های باقی‌مانده شامل لایه‌های نرمال‌ساز دسته‌ای و اتصالات اضافی پرش بین رمزگذار و رمزگشا است که مشکل محوشدگی گرادیان و پدیده تنزل در فرآیند آموزش را تا حدود زیادی حل و از بیش‌برازش[۱] شبکه در مجموعه داده‌های آموزش و اعتبارسنجی جلوگیری می‌کند. در کنار تعریف بلوک‌های باقی‌مانده، پیاده‌سازی بخش ادغام هرم فضایی متخلخل را در زیرشبکه سوم از رمزگذار *ResNet*-۵۰ برای بازیابی اطلاعات در مقیاس‌های مختلف، با استفاده از اعمال چندین لایه پیچشی با نرخ‌های افزایش متغیر برروی تصاویر ورودی وجود دارد. بخش ادغام هرم فضایی متخلخل متشکل از پنج شاخه با ابعاد فیلتر۲۵۶ است. این بخش از یک لایه پیچشی با اندازه هسته = (۱×۱) و نرخ افزایش یک، سه لایه پیچشی با اندازه هسته = (۳×۳) و نرخ‌های افزایش به‌ترتیب ۶، ۱۲، ۱۸ و یک لایه ادغام میانگین (۱×۱) به منظور ادغام ویژگی‌های چند-مقیاسه تشکیل شده است. همان‌طور که در شکل (۳) نشان داده شده است، در نهایت ویژگی‌های حاصل از پنج شاخه به هم متصل شده و از یک لایه پیچشی (۱×۱) عبور داده می‌شوند. زیر شبکه رمزگشا متشکل از دو عملیات نمونه‌برداری افزایشی با بهبود چهار برابری اندازه نقشه ویژگی در هر عملیات است. در رمزگشا برای بازیابی اطلاعات مکانی از اتصالات پرش بین رمزگذار و رمزگشا استفاده می‌شود. ورودی بخش رمزگشا متشکل از نقشه‌های ویژگی چند-مقیاسه (۳۲×۳۲) با ۲۵۶ کانال حاصل از خروجی زیرشبکه رمزگذار است. نقشه‌های ویژگی چند-مقیاسه ورودی، ابتدا به‌واسطه درون‌یابی دوخطی با ضریب چهار در ابعاد (۱۲۸،۱۲۸) نمونه‌برداری افزایشی می‌شود، سپس با ویژگی‌های سطح پایین با وضوح مکانی یکسان از شبکه عصبی *ResNet*-۵۰ الحاق می‌شوند. قبل از عملیات الحاق، به منظور جلوگیری از اثرگذاری زیاد تعداد کانال‌های با ویژگی‌های سطح پایین بر اطلاعات معنایی در ویژگی‌های سطح بالا، تعداد کانال‌های حاصل از ویژگی‌های سطح پایین با اعمال لایه‌های پیچشی (۱×۱)،

[۱] *Over-fitting*

نرمال‌سازی دسته‌ای و فعال‌ساز *Relu* کاهش داده می‌شوند. در گام بعد، ویژگی‌های الحاقی از دو بلوک متوالی متشکل از لایه‌های (پیچشی (۳×۳)، نرمال‌سازی دسته‌ای و فعال‌ساز *Relu*) با ابعاد (۱۲۸×۱۲۸) و تعداد کانال‌های ۲۵۶ به منظور رمزگشایی ویژگی‌ها عبور داده می‌شوند. سپس به‌واسطه عملیات ادغام میانگین سراسری و تعریف لایه‌های متراکم دو بعدی، ادغام نقشه‌های ویژگی برای طبقه‌بندی دسته‌های مربوطه انجام می‌شود. این لایه‌ها به جای لایه‌های کاملاً متصل در شبکه عصبی طراحی شده‌اند و عملکردی به نسبت بهتر برای طبقه‌بندی نقشه‌های ویژگی پیش‌بینی شده را دارا هستند. در نهایت پس از عملیات ادغام میانگین سراسری بر روی هر نقشه ویژگی، بردارهای پیش‌بینی با استفاده از درون‌یابی دوخطی با ابعاد (۴×۴) برای بازگرداندن به اندازه اصلی نمونه‌برداری افزایشی می‌شوند و نتایج قطعه‌بندی معنایی به واسطه یک لایه پیچشی دو بعدی سافت-مکس با ابعاد (۵۱۲×۵۱۲) و شش کانال در خروجی ارائه داده می‌شوند.

۲-۲- معیار ارزیابی

یک موضوع اساسی بعد از فرآیند آموزش مدل قطعه‌بندی معنایی، ارزیابی و ارائه گزارشی کامل از عملکرد روش پیشنهادی است. رویکرد متداول برای تجزیه و تحلیل مدل‌های آموزش‌دیده، بررسی عناصر قطری و غیرقطری ماتریس درهم‌ریختگی (*matrix Confusion*) است. شکل (۴) یک ماتریس درهم‌ریختگی را برای شش دسته برچسب شی نشان می‌دهد.

شکل ۴: ماتریس درهم‌ریختگی

به منظور ارزیابی عملکرد کمی قطعه‌بندی معنایی مدل پیشنهادی، معیار دقت کلی، *IOU* برای هر دسته، *MeanIOU*، و *MeanBFScore* را به ترتیب طبق روابط(۱) تا (۴) محاسبه شد. نسبت اشتراک به اجتماع *(IOU)* یکی از معیارهای محبوب در سنجش عملکرد قطعه‌بندی است که همپوشانی دو شیء را با محاسبه نسبت سطح اشتراک به اجتماع کمی‌سازی می‌کند. در واقع معیار همبستگی بین برچسب‌های پیش‌بینی و هدف است که برای هر دسته جداگانه محاسبه می‌شود. میانگین نسبت اشتراک به اجتماع *(MeanIOU)*، برابر با محاسبه میانگین معیار *IOU* برای تمام دسته‌ها است.

میانگین *BFScore*، میزان تطبیق مرز پیش‌بینی‌شده اشیاء را با مرز واقعیت زمینی در کل مجموعه داده اندازه‌گیری می‌کند. مقادیر این معیارها در محدوده صفر تا یک است و مقادیر بالاتر نشان‌دهنده عملکرد بهتر قطعه‌بندی است.

$$\text{OverallAccuracy} = \frac{\sum_{i=1}^{N_{cls}} C_{ii}}{\sum_{i=1}^{N_{cls}} \sum_{j=1}^{N_{cls}} C_{ij}} \quad \text{رابطه(۱)}$$

رابطه(۲)

$$IOU_i = \frac{\text{target} \cap \text{prediction}}{\text{target} \cup \text{prediction}} = \frac{C_{ii}}{\sum_{j=1}^{N_{cls}} C_{ij} + \sum_{\substack{j=1 \\ i \neq j}}^{N_{cls}} C_{ji}}$$

$$\text{Mean}_{\text{IOU}} = \frac{1}{\text{N}_{\text{cls}}} \sum_{\text{i=1}}^{\text{N}_{\text{cls}}} \text{IOU}_{\text{i}}$$ رابطه(۳)

رابطه(۴)

$$\text{MeanBFScore} = \frac{1}{\text{N}_{\text{cls}}} \sum_{\text{i=1}}^{\text{N}_{\text{cls}}} \text{avg}(\text{F}_{\text{i}}\text{Score}_{\text{i}})$$

در این روابط، N_{cls}، تعداد دسته‌ها، C_{ii} تعداد پیکسل‌های متعلق به دسته i که به درستی به عنوان دسته i برچسب‌گذاری شده و C_{ij} تعداد پیکسل‌های متعلق به دسته i که طبقه‌بندی کننده آنها را به عنوان دسته j برچسب‌گذاری کرده است. همچنین امتیاز F_1 طبق رابطه(۵) محاسبه می‌شود. در رابطه(۵)، مقادیر صحت و بازیابی بر مبنای روابط(۶) و (۷) تعیین می‌شوند.

$$\text{F}_{\text{i}}\text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$ رابطه(۵)

$$\text{Precision}_{\text{i}} = \frac{\text{C}_{\text{ii}}}{\sum_{\text{j=1}}^{\text{N}_{\text{cls}}} \text{C}_{\text{ij}}}$$ رابطه(۶)

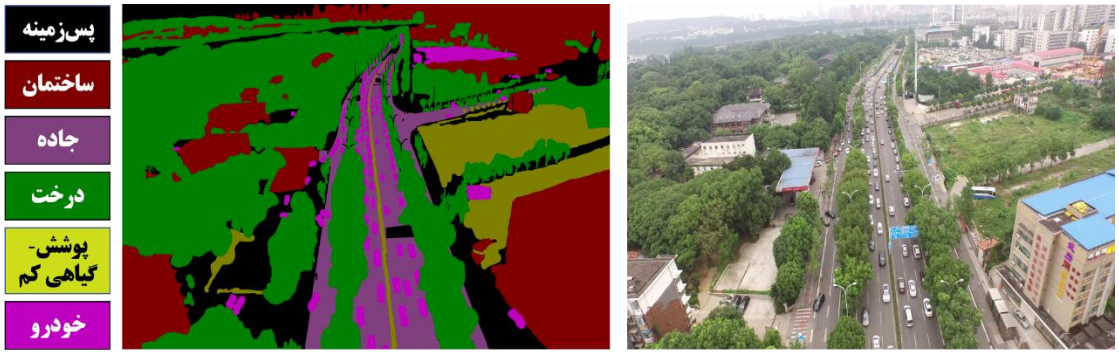
$$\text{Recall}_{\text{i}} = \frac{\text{C}_{\text{ii}}}{\sum_{\text{j=1}}^{\text{N}_{\text{cls}}} \text{C}_{\text{ji}}}$$ رابطه(۷)

۳- آزمایش‌ها و ارزیابی‌ها

در این بخش، تنظیمات مربوط به آزمایش‌های انجام شده، جزئیات پیاده‌سازی شبکه‌های عصبی و ارزیابی نتایج بدست آمده ارائه می‌شود. ابتدا مجموعه داده توضیح داده شده است. سپس، تنظیمات مربوط به فرآیند پیاده‌سازی معماری پیشنهادی مبتنی‌بر یادگیری انتقال *DeepLabV۳Plus* و الگوریتم‌های یادگیری ماشین و عمیق مقایسه‌ای را مورد بررسی قرار می‌دهیم. در نهایت، جزئیات مربوط به معیارهای ارزیابی عملکرد قطعه‌بندی معنایی مورد استفاده در این پژوهش آورده شده است.

۳-۱- مجموعه داده

مجموعه داده مورد استفاده در این کار مجموعه داده *UAVid۲۰۲۰* از پایگاه داده *ISPRS* بوده که در سال ۲۰۲۰ ارائه شده است. *UAVid۲۰۲۰* یک مجموعه داده قطعه‌بندی معنایی پهپاد-مبنا با وضوح بالا (تصاویر با وضوح بالا ۴*K* در نماهای مایل) با تمرکز بر صحنه‌های شهری است. تصاویر واقعیت زمینی در این مجموعه داده در هشت دسته (ساختمان، جاده، درخت، پوشش گیاهی کم، خودرو متحرک، خودرو ساکن، پس‌زمینه و انسان) برچسب‌گذاری شده‌اند. اندازه تصاویر اصلی ۲۱۶۰×۳۸۴۰ و ۲۱۶۰×۴۰۹۶ پیکسل است [۲۶]. وضوح بسیار بالای مجموعه تصاویر پهپاد، منابع محاسباتی/سخت افزاری بالایی را برای ذخیره ویژگی‌های حاصل از آموزش شبکه عصبی می‌طلبد. در این کار با توجه به محدودیت در منابع محاسباتی و ابعاد بسیار بالا مجموعه تصاویر، اندازه تصاویر اصلی را به ۵۱۲×۵۱۲ پیکسل تغییر می‌دهیم. در عین حال با کاهش ابعاد تصاویر، تعداد پیکسل‌های آموزشی مربوط به دسته انسان در هر تصویر به میزان قابل توجهی کاهش می‌یابد. این امر باعث می‌شود که دسته انسان در فرآیند آموزش قابل تشخیص نباشد و اشتباهاً به جای یک دسته دیگر به شبکه معرفی شود. به همین خاطر، برای بهبود عملکرد قطعه‌بندی معنایی از لحاظ کمی و کیفی، دسته انسان را از فرآیند آموزش حذف شد. همچنین از آنجائی‌که در این مطالعه از تصاویر برای قطعه‌بندی معنایی استفاده شده است، امکان تفکیک خودروی ساکن و متحرک وجود ندارد و درکل از شش دسته (ساختمان، جاده، خودرو ، درخت، پوشش گیاهی کم و پس‌زمینه) برای آموزش و ارزیابی مدل آموزش دیده در قطعه‌بندی معنایی تصاویر پهپاد شهری استفاده شده است. در شکل (۵) نمونه‌ای از یک تصویر ورودی و واقعیت زمینی متناظر از مجموعه داده شهری *UAVid۲۰۲۰* نشان داده شده است.



(ب) (الف)

شکل ۵: نمونه‌ای از از مجموعه داده *UAVid*۲۰۲۰ (الف) تصویر اصلی (ب) واقعیت زمینی .

۳-۲- تنظیم پارامترها

آزمایش‌های مربوط به پیاده سازی مدل *DeepLabV۳Plus* پیشنهادی در نرم‌افزار پایتون نسخه ۳٫۶ با استفاده از کتابخانه یادگیری عمیق کراس[۱] بر روی پلتفرم تنسورفلو[۲] انجام شده است. در این مطالعه از پردازنده گرافیکی تسلا *T۴*[۳] با ۱۳ گیگابایت حافظه و پهنای باند بالای *GDDR۶* استفاده شده است. برای ارزیابی عملکرد قطعه‌بندی معنایی در حین فرآیند آموزش، به حداقل رساندن خطای آموزش و به روزرسانی پارامترهای مدل عصبی طراحی شده، از تابع خطا *Dice*، بهینه‌ساز تخمین گشتاور تطبیقی[۴] *(Adam)*، معیار میانگین *IOU* و دقت آموزش مدل استفاده کردیم. لایه‌های رمزگذار (بلوک‌های مربوط به شبکه *ResNet-۵۰*) را با وزن‌های از قبل آموزش‌دیده برروی پایگاه داده *ImageNet* مقداردهی اولیه شده و وزن‌های لایه‌های دیگر مدل نیز با یک مقدار اولیه یکنواخت، مقداردهی اولیه می‌شوند. آموزش مدل در ۳۰ تکرار با نرخ یادگیری ۰٫۰۰۰۵ و اندازه دسته چهار برای مجموعه تصاویر ورودی در هر تکرار انجام می‌شود.

همچنین طرح توقف اولیه برای جلوگیری از بیش‌برازش داده‌های اعتبارسنجی در مدل، برروی کنترل معیار دقت اعتبارسنجی برمبنای ۲۰ تکرار تنظیم شده است.

۳-۳- نتایج

در گام نخست، آموزش شبکه در ۳۰ تکرار انجام شد که در شکل (۶) نحوه تغییرات دقت و خطا آموزش مدل قطعه‌بندی معنایی *DeepLabV۳Plus* روی مجموعه داده‌های آموزش و اعتبارسنجی *UAVid*۲۰۲۰ در تکرارهای مختلف نشان داده شده است.

همان‌طورکه در شکل (۶) مشخص است، خطا آموزش مدل به تدریج کاهش و دقت افزایش یافته است و در سطح بالا با افزایش تکرارها این مقادیر پایدار باقی می‌مانند. مدل پیشنهادی به سرعت همگرا و میانگین دقت آموزش مدل قطعه‌بندی پیشنهادی در ۳۰ تکرار به بیش از ۹۰ درصد رسیده است. علاوه‌بر این سیر کاهش خطای آموزش مدل پیشنهادی به طور قابل توجهی سریع است. در گام نخست ارزیابی، عملکرد روش پیشنهادی با استفاده از دقت کاربر و دقت تولیدکننده به دست آمده از ماتریس درهم‌ریختگی مدل‌های قطعه‌بندی مورد بررسی قرار گرفته است.

در جدول (۱) نتایج بدست آمده از روش پیشنهادی با دو مدل *Seg-Net* و *U-Net* به عنوان دو شبکه عصبی عمیق قدرتمند مقایسه شده است.

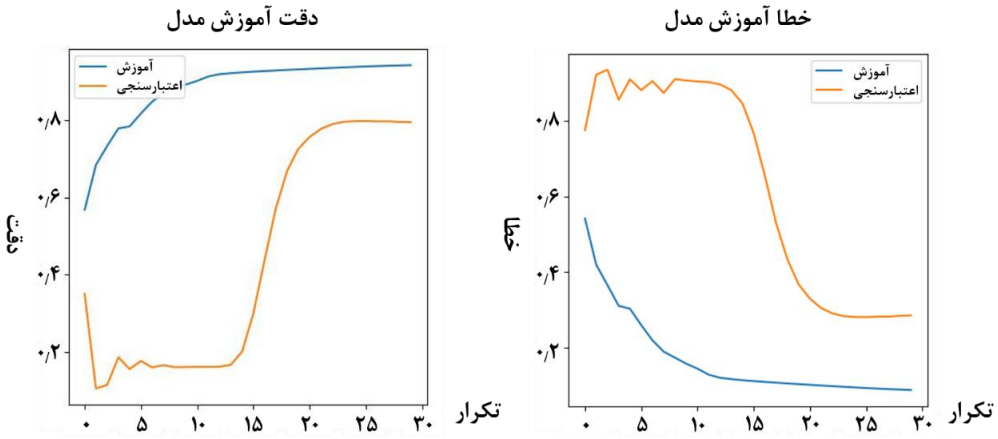
---

[۱] *Keras*

[۲] *TensorFlow*

[۳] *TESLA T۴*

[۴] *Adaptive Moment Estimation (Adam)*



شکل ۶: نمودار دقت و خطا آموزش مدل پیشنهادی در مجموعه داده *UAVid*۲۰۲۰. (الف) سیر دقت روی داده‌های آموزش و اعتبارسنجی، (ب) خطا آموزش مدل برروی داده‌های آموزش و اعتبارسنجی.

جدول ۱: دقت کاربر و تولید کننده روش پیشنهادی در مقایسه با *Seg-Net* و *U-Net*.

روش / دسته	دقت کاربر (٪)			دقت تولید کننده (٪)		
	DeepLabV3 Plus	Seg-Net	U-Net	DeepLabV3 Plus	Seg-Net	U-Net
پس‌زمینه	۶۹٫۹۵	**۷۰٫۴**	۶۰٫۳	**۶۴٫۹**	۶۰٫۸۱	۵۷٫۳
ساختمان	**۸۸٫۹۲**	۸۸٫۳	۸۶٫۱۳	**۹۴٫۱۴**	۹۲٫۸۴	۹۳٫۳۶
جاده	۷۹٫۲	**۷۹٫۲۵**	۷۸٫۶۵	۸۱٫۰۲	**۸۱٫۶۲**	۷۳
درخت	**۷۷٫۸۲**	۷۲٫۷	۷۱٫۱	۸۳٫۴۳	**۸۶٫۱۱**	۶۹٫۴
پوشش گیاهی کم	**۷۹**	۷۷٫۵۴	۶۰٫۵۴	**۶۵٫۸**	۵۵٫۲۱	۵۹٫۲
خودرو	**۷۹٫۲۱**	۷۳٫۲۲	۷۸٫۹	۶۱٫۷۲	**۶۸٫۱**	۵۹٫۸۱

از آنجائی‌که که در تصویربرداری مایل پهپاد تغییرات مقیاس بالایی بین اشیا یکسان در فواصل یا دسته‌های مختلف وجود دارد و عوارض یکسان در مقیاس‌های متفاوت ظاهر می‌شوند، تفکیک دسته‌ها در یک منطقه شهری با پیچیدگی‌های زیادی مواجه است. نتایج بدست آمده در جدول (۱) نشان می‌دهد که روش پیشنهادی در تفکیک اکثر دسته‌ها (ساختمان، درخت، پوشش گیاهی کم و خودرو)دقتی بالاتر و در برخی دسته‌ها (جاده و پس‌زمینه) دقتی نزدیک به دو روش دیگر دارد. معماری *DeepLabV۳Plus* عملکرد بالایی را در تفکیک دسته ساختمان به عنوان یکی لز مهم‌ترین دسته‌ها در مناطق شهری با دقت کاربر ۸۸٫۹۲ درصد نسبت به سایر دسته‌ها ارائه کرده است که بیانگر قابلیت بالای روش پیشنهادی در منطقه پیچیده شهری در تصاویر مایل پهپاد می‌باشد. مدل پیشنهادی در تفکیک دسته‌های جاده و پس‌زمینه با اختلافی بسیار کم (کمتر از یک درصد) عملکردی نزدیک نسبت به معماری *Seg-Net* در طبقه‌بندی دسته‌ها داشته است. در جدول (۲) سه معیار صحت، بازیابی و امتیاز F_1 برای هر دسته آورده شده است.

جدول ۲: معیارهای صحت، بازیابی و امتیاز F_1 برای هردسته.

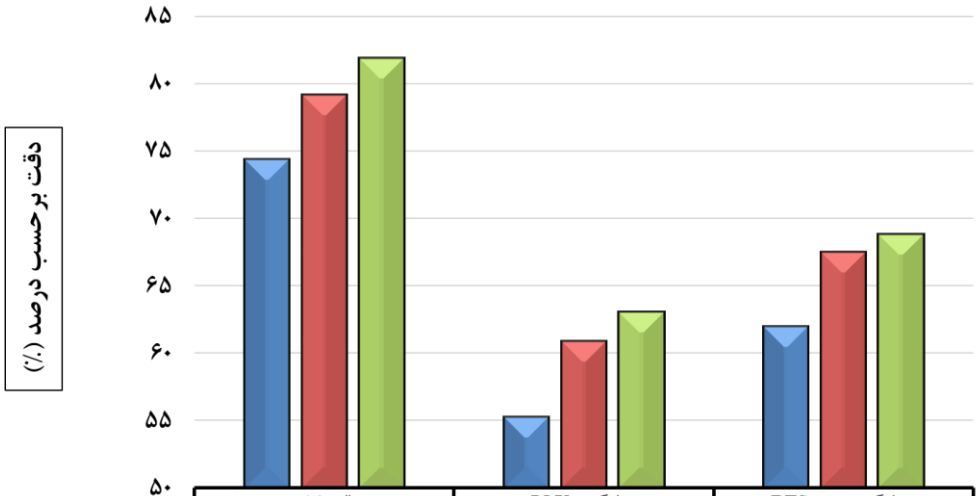
معیار	صحت			بازیابی			امتیاز F_1		
روش / دسته	DeepLabV3 Plus	Seg-Net	U-Net	DeepLabV3 Plus	Seg-Net	U-Net	DeepLabV3 Plus	Seg-Net	U-Net
پس‌زمینه	**۰٫۷**	**۰٫۷**	۰٫۶	**۰٫۶۵**	۰٫۶۱	۰٫۵۷	**۰٫۶۷**	۰٫۶۵	۰٫۵۹
ساختمان	**۰٫۸۹**	۰٫۸۸	۰٫۸۶	**۰٫۹۴**	۰٫۹۳	۰٫۹۳	**۰٫۹۱**	۰٫۹	۰٫۹
جاده	**۰٫۷۹**	**۰٫۷۹**	۰٫۷۸	۰٫۸۱	**۰٫۸۲**	۰٫۷۳	**۰٫۸**	**۰٫۸**	۰٫۷۶
درخت	**۰٫۷۸**	۰٫۷۳	۰٫۷۱	۰٫۸۳	**۰٫۸۶**	۰٫۶۹	**۰٫۸۱**	۰٫۷۹	۰٫۷
پوشش گیاهی کم	**۰٫۷۹**	۰٫۷۸	۰٫۶۱	**۰٫۶۶**	۰٫۵۵	۰٫۵۹	**۰٫۷۲**	۰٫۶۴	۰٫۶
خودرو	**۰٫۷۹**	۰٫۷۳	۰٫۷۸	۰٫۶۲	**۰٫۶۸**	۰٫۶	۰٫۶۹	**۰٫۷۱**	۰٫۶۸

با بررسی جدول (۲) می‌توان کیفیت بالای معماری پیشنهادی را در قطعه‌بندی معنایی اکثر دسته‌های مناطق شهری مشاهده کرد. روش پیشنهادی بر اساس معیار صحت در تمام دسته‌ها دقتی بالاتر یا مساوی روش‌های مقایسه‌ای دارد. هر چند روش *Seg-Net* بر اساس معیار بازیابی در سه دسته (جاده، درخت و خودرو)، دقت بالاتری در قطعه‌بندی از روش پیشنهادی دارد. نهایتاً بر اساس معیار $F_1 Score$ روش پیشنهادی در همه دسته‌ها به جز خودرو بالاترین دقت را دارد. معماری پیشنهادی در دسته ساختمان با امتیاز F_1 بالای ۹۰ درصد عملکرد قابل قبولی را در مقایسه با سایر دسته‌ها و همچنین دیگر معماری‌ها ارائه کرده است.یکی دیگر از معیارهای کلیدی و تاثیرگذار در ارزیابی نتایج قطعه‌بندی معنایی، محاسبه معیار *IOU* (بررسی درصد همپوشانی بین برچسب‌های پیش‌بینی‌شده با واقعیت زمینی) برای هردسته است. جدول (۳) مقادیر *IOU* شش دسته را در نتایج بدست آمده از روش پیشنهادی در مقایسه با دو شبکه عصبی *Seg-Net* و *U-Net* نمایش می‌دهد.

مقادیر *IOU* گزارش شده در جدول (۳) بیانگر پتانسیل بالای روش یادگیری انتقال پیشنهادی در مقایسه با دو روش یادگیری عمیق دیگر است. از بین شش دسته موجود در مجموعه داده مورد مطالعه، روش پیشنهادی در چهار دسته ساختمان، درخت، پوشش گیاهی کم و خودرو عملکرد به نسبت بهتری را در همپوشانی نقشه‌های قطعه‌بندی پیش‌بینی شده در مقایسه با واقعیت زمینی داشته و در دو دسته جاده و خودرو اختلاف حدوداً یک درصدی با روش *Seg-Net* دارد. به منظور ارزیابی کلی قطعه‌بندی معنایی، سه معیار دقت کلی، میانگین *IOU* و میانگین *BFScore* برای روش پیشنهادی محاسبه شده و با نتایج معماری‌های *Seg-Net* و *U-Net* مقایسه شد (شکل (۷)).

جدول ۳: معیار *IOU* برای هردسته (برحسب درصد)

دسته	روش		
	DeepLabV3 Plus	Seg-Net	U-Net
پس‌زمینه	**۵۰٫۷۳**	۴۸٫۴۲	۴۱٫۶
ساختمان	**۸۴٫۳**	۸۲٫۶۵	۸۱٫۱۶
جاده	۶۶٫۸	**۶۷٫۲۵**	۶۰٫۹۲
درخت	**۶۷٫۴**	۶۵٫۱	۵۴٫۱
پوشش گیاهی کم	**۵۶**	۴۷٫۶	۴۲٫۷
خودرو	۵۳٫۱۲	**۵۴٫۵۲**	۵۱٫۵۵



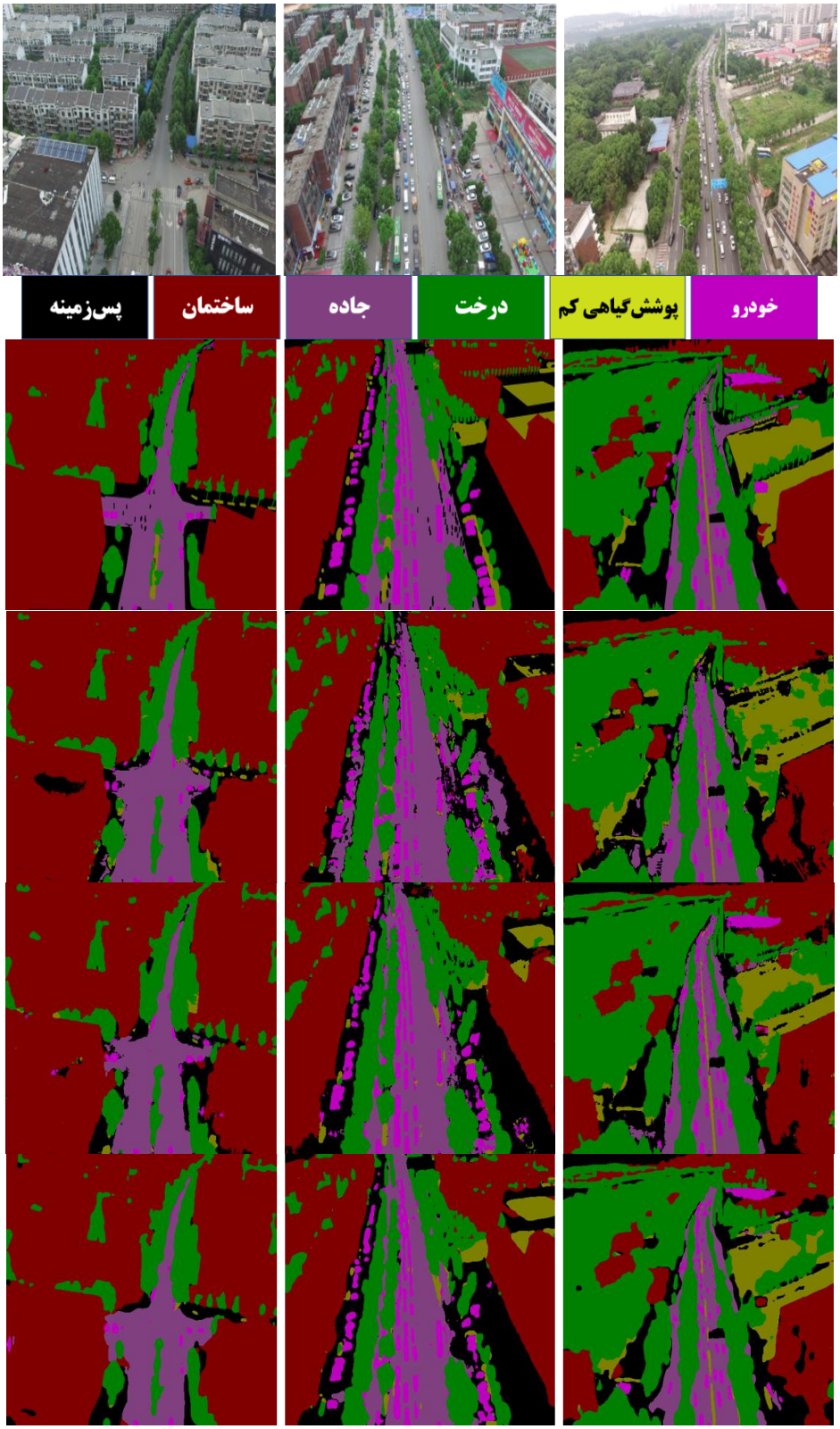
	دقت کلی	میانگین IOU	میانگین BFScore
U-Net	۷۴٫۳۵	۵۵٫۳۳	۶۲٫۰۲
Seg-Net	۷۹٫۱۵	۶۰٫۹۲	۶۷٫۵۴
DeepLabv3Plus	۸۱٫۹۳	۶۳٫۰۴	۶۸٫۸

شکل ۷: مقایسه معیارهای دقت کلی، میانگین *IOU* و میانگین *BFScore* برای سه مدل *U-Net*، *Seg-Net* و *DeepLabV۳Plus*

نتایج آماری بدست آمده در نمودار شکل (۷) توانایی بالای روش پیشنهادی را در قطعه‌بندی مناطق پیچیده شهری اثبات می‌کند. دقت بالای ۸۰٪ در تصاویر مایل پهپاد در حالتی‌که تصاویر به دلیل محدودیت منابع محاسباتی به میزان قابل توجهی کوچک شده‌اند ، دقت مطلوبی محسوب می‌شود. همچنین مقایسه روش پیشنهادی با سایر روش‌های یادگیری عمیق نشانگر توانایی یادگیری انتقال در بهبود عملکرد شبکه‌های یادگیری عمیق است. روش پیشنهادی در هر سه معیار (دقت کلی، میانگین *IOU* و میانگین *BFScore*) با توجه به رویکرد یادگیری انتقال (انتقال وزن‌های شبکه از قبل آموزش‌دیده *ResNet-۵۰* در فرآیند آموزش شبکه *DeepLabV۳Plus*) عملکرد بهتری را نسبت به دیگر معماری‌های عمیق پایه برای قطعه‌بندی معنایی ارائه داده است.

در شکل (۸) سه تصویر از مجموعه تصاویر آزمایشی مورد مطالعه به همراه واقعیت زمینی متناظر و نقشه-های طبقه‌بندی مدل پیشنهادی در مقایسه با دیگر مدل‌های یادگیری عمیق با هدف تحلیل بصری نتایج آورده شده است.

در یک بررسی کلی نتایج بصری (ردیف‌های سوم و چهارم شکل (۸))، به وضوح پیداست‌که رویکردهای یادگیری عمیق *UNet* و *SegNet*، در تطبیق مرزهای پیش‌بینی‌شده با واقعیت زمینی قطعه‌بندی ضعیفی را ارئه کردند. در مقابل، معماری مبتنی بر رویکرد یادگیری انتقال *DeepLabV۳Plus* (ردیف پنجم شکل(۸)) نسبت به دو معماری دیگر عملکرد قابل قبولی در قطعه‌بندی معنایی تصاویر مایل مبتنی‌بر پهپاد شهری ارائه داده است. شبکه عصبی *UNet* در طبقه‌بندی اکثر دسته‌ها عملکرد ضعیفی داشته است. روش پیشنهادی در مقایسه با دو روش دیگر، توانایی بالایی در قطعه‌بندی معنایی اکثر دسته‌ها، به ویژه چهار دسته ساختمان، درخت، پوشش گیاهی کم و خودرو ارائه کرده است.



شکل ۸: (الف) تصویر ورودی (ب) واقعیت زمینی (ج) نتایج تقسیم بندی مدل *U-Net*، (د) نتایج تقسیم بندی مدل *Seg-Net* و (و) نتایج تقسیم بندی مدل *DeepLabV۳Plus*

۴- نتیجه‌گیری و پیشنهادها

با پیشرفت تکنولوژی سکوها و سنجنده‌ها، تصاویر با قدرت تفکیک بالا در زمان‌های مختلف در دسترس می‌باشد؛ از سوی دیگر با افزایش قدرت تفکیک مکانی و طیفی و تغییر زاویه دید سنجنده، استخراج اطلاعات با چالش‌هایی مواجه شده است. در این مطالعه قطعه-بندی معنایی تصاویر مایل پهپاد با قدرت تفکیک مکانی بالا مورد بررسی قرار گرفت. با توجه به پیچیدگی‌های بالای عوارض در مناطق شهری، الگوریتم‌های یادگیری ماشین سنتی با محدودیت‌های جدی مواجه هستند. در دهه اخیر به‌واسطه پیشرفت الگوریتم‌های یادگیری عمیق، توانسته‌اند که با موفقیت در کاربردهای مختلف مورد استفاده قرار گیرند . در این مطالعه، توانایی یادگیری انتقال در قطعه‌بندی معنایی تصاویر مایل پهپاد بررسی شده است. یکی از محدودیت‌های شبکه-های یادگیری عمیق مرحله آموزش آن می‌باشد که نیاز به داده‌های آموزشی زیاد و حجم محاسباتی بالا می-باشد. یادگیری انتقال راهکاری برای کاهش این محدودیت‌ها می‌باشد که در آن از شبکه‌های از پیش آموزش‌دیده‌استفاده شده و پارامترهای شبکه با داده-های آموزشی مورد نظر بهبود داده می‌شود.

رویکرد پیشنهادی برای قطعه‌بندی معنایی تصاویر پهپاد-مبنا نواحی شهری متشکل از شبکه عصبی پیچشی از قبل آموزش دیده *ResNet*-۵۰ و شبکه عصبی رمزگذار-رمزگشا پیچشی *DeepLabV۳Plus* است. فرآیند آموزش مدل پیشنهادی مبتنی‌بر روش یادگیری انتقال برمبنای شبکه عصبی پیچشی *ResNet*-۵۰ با وزن‌های از قبل آموزش‌دیده است. پتانسیل روش پیشنهادی با استفاده از مجموعه داده پهپاد مبنا *UAVid* در مقایسه با شبکه‌های قدرتمند *Seg-Net* و *U-Net* ارزیابی شد. نتایج بدست آمده حاکی از برتری روش پیشنهادی با دقت بالای ۸۱٪، *IOU* متوسط ۶۳٪ و *BFScore* متوسط ۶۸٪ است. همچنین در اکثر دسته‌های موجود در تصویر، روش پیشنهادی با دقت بالاتری تفکیک دسته‌ها را انجام داد. دسته ساختمان با دقت کاربر و تولیدکننده بیش از ۸۸٪ و ۹۴٪ با دقت بالایی در تصاویر مایل طبقه‌بندی شد. همچنین دسته خودروبا ۷۹٫۲۱٪ دقت از پنج دسته دیگر تفکیک شده است. مقایسه روش پیشنهادی با دو روش یادگیری عمیق شناخته شده *U-Net* و *SegNet* بیانگر توانایی بالای روش *DeepLabV۳Plus* در قطعه‌بندی معنایی نواحی شهری پهپاد مبنا است.

با توجه به نتایج قابل قبول روش ارائه شده، پیشنهاد می‌شود در کارهای آتی تصویر در ابعاد بزرگ‌تر با استفاده از سخت‌افزارهای قوی‌تر پردازش شود. همچنین به منظور اثبات توانایی الگوریتم‌های یادگیری عمیق می‌توان روش‌های سنتی را پیاده سازی کرده و نتایج بدست آمده را با آن‌ها مقایسه کرد. در این مطالعه چند شبکه یادگیری عمیق محدود پیاده‌سازی شده است که در مطالعات بعدی می‌توان توانایی شبکه-های دیگر را نیز بررسی کرد.

مراجع

[1] *S. Girisha, M.M. Manohara Pai, U. Verma and R. M. Pai, "Semantic Segmentation of UAV Aerial Videos using Convolutional Neural Networks", 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 21-27, Sardinia, Italy, 2019.*

[2] *X. Yuan, J. Shi and L. Gu. "A review of deep learning methods for semantic segmentation of remote sensing imagery", Expert Systems with Applications, Vol. 169, p. 114417, 2021.*

[3] *S. Girisha, U. Verma, M.M. Pai and R.M. Pai, "Uvid-net: Enhanced semantic segmentation of uav aerial videos by embedding temporal information", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 14, pp. 4115-4127, 2021.*

[4] *H. Yao, R. Qin and X. Chen, "Unmanned aerial vehicle for remote sensing applications—A review", Remote Sensing, Vol. 11(12), p. 1443, 2019.*

[5] *N.M. Noor, A. Abdullah and M. Hashim, "Remote sensing UAV/drones and its applications for urban areas: A review", In IOP conference series: Earth and environmental science, Vol. 169, No. 1, p. 012003, UK, 2018.*

[6] *S. Minaee, Y.Y. Boykov, F. Porikli, A.J. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image segmentation using deep learning: A survey", in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, no. 7, pp. 3523-3542, 2022.*

[7] *M.N. Reza, I.S. Na, S.W. Baek and K.H. Lee, "Rice yield estimation based on K-means clustering with graph-cut segmentation using low-altitude UAV images", Biosystems engineering, Vol. 177, pp.109-121, 2019.*

[8] *R. Azhar, D. Tuwohingide, D. Kamudi and N. Suciati, "Batik image classification using SIFT feature extraction, bag of features and support vector machine", Procedia Computer Science, Vol. 72, pp. 24-30, 2015.*

[9] *D.A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization", Canadian Journal of remote sensing, Vol. 28(1), pp. 45-62, 2002.*

[10] *N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, pp. 886-893, San Diego, CA, USA, 2005.*

[11] *H. Hasani, F. Samadzadegan and P. Reinartz, "A metaheuristic feature-level fusion strategy in classification of urban area using hyperspectral imagery and LiDAR data", European Journal of Remote Sensing, 50(1), pp. 222-236, 2017.*

[12] *P. Mohammadpour, D.X. Viegas and C. Viegas, "Vegetation Mapping with Random Forest Using Sentinel 2 and GLCM Texture Feature—A Case Study for Lousã Region, Portugal", Remote Sensing, 14(18), p. 4585, 2022.*

[13] *P. Sturgess, K. Alahari, L. Ladicky and P.H. Torr, "Combining appearance and structure from motion features for road scene understanding", In BMVC-British Machine Vision Conference, London, UK, 2009.*

[14] *A.S. Laliberte and A. Rango, "Texture and scale in object-based analysis of subdecimeter resolution unmanned aerial vehicle (UAV) imagery", IEEE Transactions on Geoscience and Remote Sensing, 47(3), pp.761-770, 2009.*

[15] *C. Zhang, L. Wang and R. Yang, "Semantic segmentation of urban scenes using dense depth maps", In Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, pp. 708-721, Berlin, Heidelberg, Germany, 2010.*

[16] *T. Moranduzzo and F. Melgani, "Detecting Cars in UAV Images With a Catalog-Based Approach", in IEEE Transactions on Geoscience and Remote Sensing, Vol. 52, no. 10, pp. 6356-6367, 2014.*

[17] *A. Vezhnevets, V. Ferrari and J.M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model", In 2011 international conference on computer vision (ICCV 2011), pp. 643-650, Barcelona, Spain, 2011.*

[18] *M. Qi, Y. Shi, Y. Qi, C. Ma, R. Yuan, D. Wu and Z.J. Shen, "A practical end-to-end inventory management model with deep learning", Management Science, 69(2), pp.759-773, 2023.*

[19] *L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin and B.A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review", ISPRS journal of photogrammetry and remote sensing, Vol. 152, pp. 166-177, 2019.*

[20] *J. Yosinski, J. Clune, Y. Bengio and H. Lipson, "How transferable are features*

in deep neural networks?", Advances in neural information processing systems, Vol. 27, p. 1-14, 2014.

[21] M. Oquab, L. Bottou, I. Laptev and J. Sivic, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks", 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717-1724, Columbus, OH, USA, 2014.

[22] J. Senthilnath, N. Varia, A. Dokania, G. Anand, and J. A. Benediktsson, "Deep TEC: Deep transfer learning with ensemble classifier for road extraction from UAV imagery", Remote Sensing, vol. 12, no. 2, pp. 245, 2020.

[23] F.G. Zanjani and M. van Gerven, "Improving semantic video segmentation by dynamic scene integration", Paper presented at The Netherlands Conference on Computer Vision (NCCV 2016), Lunteren, Netherlands, 2016.

[24] X. Wei, K. Fu, X. Gao, M. Yan, X. Sun, K. Chen, and H. Sun, "Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model", Remote Sensing Letters, vol. 9, no. 3, pp. 199-208, 2018.

[25] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling", IEEE Access, vol. 7, pp. 128774-128786, 2019.

[26] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 165, pp. 108-119, 2020.

[27] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation", Applied Soft Computing, vol. 70, pp. 41-65, 2018.

[28] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang and C. Liu, "A survey on deep transfer learning", In International conference on artificial neural networks, Springer, Cham, pp. 270-279, Greece, 2018.

[29] B. Cui, X. Chen, and Y. Lu, "Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection", Ieee Access, vol. 8, pp. 116744-116755, 2020.

[30] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 9, pp. 3252-3261, 2018.

[31] X. Zhang, Z. Xiao, D. Li, M. Fan and L. Zhao, "Semantic Segmentation of Remote Sensing Images Using Multiscale Decoding Network", in IEEE Geoscience and Remote Sensing Letters, Vol. 16, no. 9, pp. 1492-1496, 2019.

[32] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning", Remote Sensing, Vol. 11(1), p. 83, 2019.

[33] Y. Liu, Y. Kong, B. Zhang, X. Peng, and H. Leung, "A Novel Deep Transfer Learning Method for Airborne Remote Sensing Semantic Segmentation Based on Fully Convolutional Network.", In 2020 4th International Conference on Imaging, Signal Processing and Communications (ICISPC), pp. 13-19, Kumamoto, Japan, 2020..

[34] L. Zhang, M. Wang, Y. Fu, and Y. Ding, "A Forest Fire Recognition Method Using UAV Images Based on Transfer Learning," Forests, vol. 13, no. 7, pp. 975, 2022.

K. N. TOOSI UNIVERSITY OF TECHNOLOGY
FACULTY OF GEODESY AND GEOMATICS ENGINEERING

Journal of Geospatial Information Technology
Vol.10, No.4, Winter 2023

Research Paper

Transfer Learning Framework for Semantic Segmentation of High-Resolution UAV-based images in Urban Area

Abbas Majidizadeh [1], Hadiseh Hasani [2]*, Marzieh Jafari [2]

1- MSc Student, Department of Geodesy and Surveying Engineering, Tafresh University, Tafresh
2- Assistant Professor, Department of Geodesy and Surveying Engineering, Tafresh University, Tafresh

Abstract

Semantic segmentation technique for Unmanned Aerial Vehicle (UAV) data processing has been one of the leading researches in photogrammetry, remote sensing, and computer vision in recent years. This technique has attracted increasing attention from industry and academia (a wide range of academic and real-world applications). Many applications, including aerial mapping of urban scenes, positioning objects in aerial images, automatic extraction of buildings from remote sensing or high-resolution aerial images, etc., require accurate and efficient segmentation algorithms. However, proper and accurate semantic segmentation using a deep learning approach (overall training of a deep neural network with random weighting) requires a large amount of training and labeled images. As we are facing the challenge of a lack of labeled data in the field of urban aerial images, we used the Transfer Learning Approach for the semantic segmentation of the UAV-based images of urban areas in this paper. The proposed method implements a transfer learning framework based on DeepLabV3Plus convolutional encoder-decoder architecture with ResNet-50 pre-trained model in ImageNet collection for semantic segmentation of the urban scenes. The dataset studied in this research is the UAVid2020, an urban UAV-based semantic segmentation dataset from the International Society for Photogrammetry and Remote Sensing (ISPRS). We used traditional deep learning models (U-Net and Seg-Net convolutional encoder-decoder neural networks) to evaluate the semantic segmentation performance of the proposed method. Finally, the results of the semantic segmentation of UAV-based images show the effectiveness of the proposed transfer learning framework compared to the deep learning models, in terms of the overall accuracy metric. The DeepLabV3Plus-ResNet50 architecture achieved the best result with 81.93% compared to U-Net and Seg-Net neural networks with 74.35% and 79.15% respectively.

Key words: *Semantic Segmentation, Unmanned Aerial Vehicle (UAV), Transfer Learning, Convolutional Encoder-Decoder Deep Neural Network, DeepLabV3Plus.*

Correspondence Address *: Department of Geodesy and Surveying Engineering, Tafresh University, Tafresh, Iran.*
Tel*: +988622367430.*
Email*: h.hasani@tafreshu.ac.ir*