Available in:
Journal.isrc.ac.ir

Journal of
Space Science, Technology
& Applications (Persian)

Vol. 4, No. 2, pp.: 49-62
2025

DOI:
10.22034/jssta.2024.462010.1173

Article Info

Received: 2024-06-12
Accepted: 2024-12-29

Keywords

Remote Sensing, Image Registration, Region of Interest, Transformer Deep Neural Network

How to Cite this article

S.Mohammad Hoseini Panah, etal.," Multi-Temporal Remote Sensing Image Registration with Deep Neural Networks and Region of Interes", Journal of Space Science, Technology and Applications, vol.4(2), p.: 49-62, 2025.

Original Article

Multi-Temporal Remote Sensing Image Registration with Deep Neural Networks and Region of Interest

Seyed Mohammad Hoseini Panah[1], Mohsen Soryani[2], Masoud Khoshsima[3]

1.Master of Artificial Intelligence, Iran University of Science and Technology, Faculty of Computer Engineering, Tehran, Iran, seyedmasoudhp@gmail.com

2.Associate Professor, Iran University of Science and Technology, Faculty of Computer Engineering, Tehran, Iran, soryani@iust.ac.ir

3.Assistant Professor, Iran Space Research Institute, Tehran, Iran, m.khoshsima@isrc.ac.ir

Abstract

The purpose of image registration is to align two or more images taken from the same scene at different times and/or from different perspectives and/or using different devices. In recent years, with the continuous improvement of human ability to observe the earth, the accuracy and quality of remote sensing images have increased. Therefore, the need for new image registration models that can perform high calculations of these images and also have good accuracy is observed. In this thesis, we have used a new method to solve these problems. The proposed solution includes the use of regions of interest in order to reduce the search area and increase the accuracy. For this purpose, first, the areas that are the same between two images are identified, and then, the image is registered according to the similar areas. To find the region of interest, a deep transformer neural network model is used. The proposed deep neural network of the transformer includes several layers of inner-attention and cross-attention, which has the task of learning the importance of different positions within an image and between two images. The proposed model is a self-supervised method that generate training data using the segment swapping. The training data was collected from Google Earth images and annotated by us. After training the model and obtaining the similar regions, we use the common SIFT model to obtain the image registration. For testing, we have used Sentinel-2 aerial images. To quantitatively evaluate the result, we use the root mean square error. Quantitative and qualitative results show a significant performance gap in cost and accuracy, compared to conventional methods of capturing aerial images.

دسترس‌پذیر در نشانی:
Journal.isrc.ac.ir

دو فصلنامه
علوم، فناوری و کاربردهای فضایی

سال چهارم، شماره ۲، صفحه ۶۲-۴۹
پاییز و زمستان ۱۴۰۳

DOI:
10.22034/jssta.2024.462010.1173

تاریخچه داوری

دریافت: ۱۴۰۳/۰۳/۲۳
پذیرش: ۱۴۰۳/۱۰/۰۹

واژه‌های کلیدی

سنجش از دور، تطبیق تصویر، ناحیه مورد علاقه، شبکه عصبی عمیق ترانسفورمر.

نحوه استناد به مقاله

سید محمد حسینی پناه و همکاران، " ثبت تصویر سنجش از دور چند زمانی با استفاده از شبکه‌های عصبی عمیق و نواحی مورد علاقه"، دوفصلنامه علوم، فناوری و کاربردهای فضایی، جلد چهارم، شماره دوم، صفحات ۴۹-۶۲، ۱۴۰۳.

مقاله پژوهشی

ثبت تصویر سنجش از دور چند زمانی با استفاده از شبکه‌های عصبی عمیق و نواحی مورد علاقه

سید محمد حسینی پناه[۱]، محسن سریانی[۲]، مسعود خوش سیما[۳]

۱- کارشناسی ارشد هوش مصنوعی دانشگاه علم وصنعت ایران – seyedmasoudhp@gmail.com

۲- دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران – soryani@iust.ac.ir

۳- استادیار پژوهشگاه فضایی ایران – m.khoshsima@isrc.ac.ir

چکیده

هدف از ثبت تصویر، تراز کردن دو یا چند تصویر است که از یک صحنه، در زمان های متفاوت و یا از دیدگاه های گوناگون و با استفاده از دستگاه های مختلف گرفته شده است. در سال های اخیر با بهبود مستمر توانایی انسان در رصد زمین، دقت و کیفیت تصاویر سنجش از دور افزایش یافته است. بنابراین نیاز به مدل های جدید ثبت تصویر که بتواند محاسبات بالای این تصاویر را انجام دهد و همچنین از دقت خوبی برخوردار باشد، مشاهده می‌شود. در این مقاله،از یک روش جدید برای حل این مشکلات استفاده شده است. راه حل پیشنهادی شامل استفاده از نواحی مورد علاقه به منظور کاهش ناحیه جست و جو و افزایش دقت است. برای این منظور ابتدا ناحیه هایی که بین دو تصویر یکسان هستند، شناسایی می‌شوند و سپس، ثبت تصویر با توجه به ناحیه های مشابه صورت می‌گیرد. برای پیدا کردن ناحیه مورد علاقه، از یک مدل شبکه عصبی عمیق ترانسفورمری استفاده شده است. شبکه پیشنهادی شامل چندین لایه توجه درونی و توجه متقاطع است که وظیفه یادگیری اهمیت موقعیت های مختلف در درون یک تصویر و بین دو تصویر را دارد. مدل پیشنهادی یک مدل خودنظارتی است که از روش تعویض بخش برای تولید داده های آموزشی استفاده می‌کند. داده های آموزشی از تصاویر Google Earth جمع آوری شده است و توسط ما نشانه گذاری شده است. پس از آموزش مدل و بدست آوردن ناحیه هایی مشابه، از ویژگی‌های SIFT برای ثبت تصویر استفاده می‌کنیم. برای آزمایش، از تصاویر هوایی Sentinel_2 استفاده شده است. برای ارزیابی کمی نتیجه، از ریشه میانگین مربعات خطا استفاده می‌کنیم. نتایج کمی و کیفی نشان دهنده بهبود عملکرد قابل توجه‌ای در هزینه و دقت، در مقایسه با روش استفاده از ویژگی‌های SIFT و نیز یک روش مبتنی برشبکه عصبی عمیق برای ثبت تصاویر هوایی است؛ بطوری که میانگین خطای ثبت تصاویر بر حسب پیکسل ۳/۵ برابر نسبت به مدل SIFT و ۱۷ برابر نسبت به مدل شبکه عصبی عمیق کاهش داشته است

۱- مقدمه

ثبت تصویر[۱]، به فرآیند تراز کردن تصاویر یک صحنه گرفته شده توسط حسگر(های) یکسان یا متفاوت، در زمان‌های مختلف و یا از دیدگاه‌های مختلف تعریف می‌شود. در ثبت تصویر یکی از تصاویر را تصویر مرجع[۲] و تصویر دیگر را تصویر دریافتی[۳] می‌نامند. هدف اصلی ثبت تصاویر، تراز کردن دقیق تصویر دریافت شده با تصویر مرجع است. اگر ثبت بین تصاویر گرفته شده در زمان های مختلف انجام شود، ثبت تصویر چند زمانی[۴] نامیده می‌شود.

سنجش از دور[۵] به فرآیند به دست آوردن اطلاعات در مورد یک شی توسط یک سیستم تصویربرداری بدون تماس فیزیکی با جسم گفته می‌شود. ماهواره های رصد زمین و سیستم های تصویربرداری هوابرد به طور مرتب با استفاده از حسگرهای منفرد یا چندگانه، تصاویر سطح زمین را به دست می آورند و این تصاویر به عنوان تصاویر سنجش از دور شناخته می‌شوند. در طول دو دهه گذشته، حجم و کیفیت تصاویر سنجش از دور به شدت افزایش یافته است. در دسترس بودن گسترده تصاویر منجر به افزایش قابل توجهی در زمینه های کاربردی شده است. این تصاویر به طور گسترده در بسیاری از برنامه ها مانند تشخیص تغییر، موزاییک سازی تصویر و ترکیب تصاویر استفاده می‌شود. در این برنامه ها، از ثبت تصویر به عنوان یک مرحله اساسی استفاده می‌شود که تراز دقیق تصویر به تصویر را انجام می دهد. ثبت تصویر سنجش از دور با ثبت تصویر طبیعی یا پزشکی از طرق مختلف متفاوت است. به طور کلی، تصاویر سنجش از دور اندازه قابل توجهی دارند و طیف گسترده ای ازحسگرهای سنجش از دور وجود دارد. علاوه بر این، شرایط گرفتن تصویر به طور قابل توجهی در این مورد تغییر می‌کند. روش‌های ثبت سنجش از دور شامل تمام این عوامل می‌شود. در این مقاله هدف، تراز کردن تصاویر نوری به نوری است که در زمان های مختلف گرفته شده است.

شناسایی الگوهای مکرر در قلب مشکل بینایی کامپیوتر نهفته است و جزء کلیدی در هوش مصنوعی است. با این حال، در عمل، بهترین روش‌ها برای انجام چنین وظیفه‌ای اساسی، اغلب نقص‌هایی به همراه دارد. در حالی که اکنون روش‌های متعددی برای کشف تطابق الگوهای دقیق (که به طور گسترده برای یافتن نقض حق نسخه‌برداری[۶] استفاده می‌شود)، و همچنین تطابق تقریبی اشیاء برجسته[۷] وجود دارد، شناسایی الگوهای مشابه بصری در یک زمینه بصری بزرگتر به طرز شگفت آوری دشوار است. یافتن الگوهای تکرار شده می تواند عملکرد را در بومی سازی بصری برای تشخیص مکان تقویت کند[1]. هم‌بخش‌بندی تصویر[۸] و شناسایی نقاط متناظر[۹] قابل اعتماد می‌تواند کشف شی را در مجموعه‌های تصویری امکان‌پذیر کند[2].

پیدا کردن ناحیه‌های مشابه چندین مزیت دارد. این کار باعث می‌شود تا هنگام تطبیق تصویر، ناحیه جست‌وجو برای یافتن نقاط متناظر کاهش یابد و به طبع آن هزینه محاسباتی نیز کاهش می‌یابد. همچنین نرخ تناظر نقاط پرت[۱۰] کاهش یافته و از ناحیه‌های گرفتگی[۱۱] مانند ابر برای تناظریابی استفاده نمی شود.

ترانسفورمرها[3]، به عنوان یک بلوک ساختمانی جدید مبتنی بر توجه برای ترجمه ماشینی معرفی شده اند. مکانیسم های توجه[4]، لایه های شبکه عصبی هستند که اطلاعات را از کل توالی ورودی جمع می کنند. پس از آنکه ترانسفورمرها لایه‌های خودتوجهی را معرفی کنند هر عنصر یک دنباله را اسکن می‌کند و با جمع‌آوری اطلاعات از کل دنباله، آن را به روز رسانی می‌کند. یکی از مزیت‌های اصلی مدل‌های مبتنی بر توجه، محاسبات سراسری[۱۲] و حافظه نگهدارنده عالی آن‌ها است. در این پژوهش از شبکه‌های ترانسفورمر برای پیدا کردن ناحیه های مشابه بین تصاویر ورودی استفاده شده است و بعد از آن با توجه به ناحیه های استخراج شده، تطابق با بهره گیری از ویژگی های SIFT [5] صورت گرفته و سپس نتایج آن ارزیابی شده است.

به دلیل در دسترس نبودن یک مجموعه داده معیار[۱۳] ، پژوهش حاضر بر روی یک مجموعه داده تصویر ماهواره‌ای چند زمانی که

---

[۱] Image registration
[۲] Reference image
[۳] Sensed image
[۴] Multi-temporal image registration
[۵] Remote sensing
[۶] Copyright infringements
[۷] Approximate matches of salient objects
[۸] Co-segmentation
[۹] Correspondence identification
[۱۰] Outlier
[۱۱] Occlusion
[۱۲] Global computation
[۱۳] Benchmark

توسط پژوهشگران تهیه شده است، مورد ارزیابی قرار گرفته است.

۲- پژوهش‌های مرتبط

۲-۱- ثبت تصویر

مزیت اصلی روش‌های مبتنی بر ویژگی این است که این روش‌ها می‌توانند تفاوت‌های هندسی و همچنین رادیومتری قابل توجهی را بین تصاویر سنجش از دور کنترل کنند[6]. با این حال، این روش‌ها در مواردی که ویژگی‌های مناسب استخراج و تطبیق داده می‌شوند، قابل اجرا هستند. استخراج ویژگی، توصیف ویژگی، تطبیق ویژگی، و حذف موارد پرت نقشی حیاتی در ثبت تصویر مبتنی بر ویژگی دارند[7,8]. استخراج ویژگی های قابل تکرار خوب یک کار چالش برانگیز در ثبت تصاویر سنجش از دور است[9]. [31] در یک روش ثبت تصویر جدید، شامل دو نوع آشکارساز ویژگی و یک استراتژی محدودیت مرزی منطقه برای تطبیق، پیشنهاد شده است. دو نوع ویژگی شناسایی شده توسط تبدیل ویژگی تغییرناپذیر مقیاس و عملگرهای هریس از مزایای نگهداری اطلاعات ساختاری مختلف در تصویر و افزایش تعداد نقاط کلیدی برای تطبیق بعدی برخوردارند. پس از آن، یک استراتژی محدودیت مرزی منطقه بر اساس نقشه طرح تصویر[۱۴] در مرحله تطبیق استفاده می‌شود. بسیاری از نقاط کلیدی در تصویر به یکدیگر نزدیک هستند، که ممکن است تطبیق یک به چند و بسیاری از نقاط پرت در فرآیند تطبیق رخ دهد. بنابراین، به منظور کاهش ابهام نتایج تطبیق، از نقشه طرح استفاده می‌شود تا نقاط گوشه و نقاط بافت در ناحیه ساختاری و ناحیه غیرساختاری را به ترتیب با توجه به موقعیت این نقاط محدود کنیم.

در پاره ای از موارد روش‌های مبتنی بر ویژگی و همچنین مبتنی بر شدت در ثبت تصاویری که حاوی تغییر شکل شدید و تأثیر نویز هستند، با شکست مواجه می شوند. روش‌های متداول مبتنی بر ویژگی، ویژگی‌های دست‌ساز، مانند اطلاعات لبه، گوشه‌ها، بافت و گرادیان را که فاقد اطلاعات معنایی سطح بالا هستند استخراج می‌کنند. در این روش ها، هیچ بازخوردی+ بین استخراج ویژگی و تطبیق استفاده نمی‌شود. بنابراین، در دسترس نبودن ویژگی‌های قوی کافی، به طور مستقیم بر عملکرد تطبیق ویژگی تأثیر می‌گذارد. به منظور حل این مسائل، یادگیری عمیق به طور رایج برای ثبت تصاویر سنجش از دور استفاده شده است. در[10] یک روش ثبت، مبتنی بر شبکه عصبی عمیق را برای یادگیری ویژگی های تصویر پیشنهاد شده است. روش های مبتنی بر یادگیری عمیق به طور خودکار ویژگی ها را یاد می گیرند و لایه های پنهان و لایه خروجی آن برای استخراج ویژگی و هدف تطبیق ویژگی استفاده می‌شود. به طور کلی، در این روش از نتایج تطبیق ویژگی برای هدایت فرآیند استخراج ویژگی استفاده می‌شود. با این حال، چالش اصلی این روش ها پیچیدگی محاسباتی است[11,12,13].

۲-۲- هم‌بخش‌بندی و پیدا کردن ناحیه مشترک

SIFT-Flow [14] روش اولیه‌ای بود که صحنه های بصری متمایز را با ترکیب ویژگی های بصری، مانند SIFT در رویکردهای جریان نوری تراز می کرد. در SIFT-Flow با الهام از روش‌های جریان نوری، که قادر به ایجاد تناظرهای متراکم پیکسل به پیکسل بین دو تصویر هستند، چارچوب محاسباتی جریان نوری را با تطبیق توصیف‌گرهای SIFT جایگزین پیکسل‌های خام می‌کند. در SIFT-flow، یک توصیفگر SIFT در هر پیکسل استخراج می شود تا ساختارهای تصویر محلی را مشخص کند و اطلاعات متنی را کدگذاری کند. یک الگوریتم تخمین جریان گسسته، برای حفظ ناپیوستگی تطبیق توصیفگرهای SIFT بین دو تصویر استفاده می‌شود. استفاده از ویژگی های SIFT امکان تطبیق قوی در ظواهر مختلف صحنه/اشیاء را فراهم می کند و مدل فضایی با حفظ ناپیوستگی امکان تطبیق اشیاء واقع در قسمت های مختلف صحنه را فراهم می کند.

معماری‌های مبتنی بر مکانیسم‌های توجه و ترانسفورمرها برای پیش‌بینی تطابقات تصویری، از اهمیت خاصی برخوردار هستند. Super-Glue [15]، یک شبکه عصبی گراف مبتنی بر توجه است که برای تطبیق نقاط بین دو تصویر استفاده می‌شود. در این کار، تطبیق ویژگی به عنوان یافتن تخصیص جزئی[۱۵] بین دو مجموعه از ویژگی های محلی بین دو تصویر در نظر گرفته می شود. در این شبکه عصبی گراف با الهام از شبکه ترانسفورمر، از توجه درون تصویری[۱۶] و بین تصویری[۱۷] استفاده می کند تا از روابط فضایی نقاط کلیدی و ظاهر بصری آنها استفاده کند.

[۱۴] Image sketch map

[۱۵] Partial assignment

[۱۶] Self-attention

[۱۷] Cross-attention

COTR [16] یک معماری ترانسفورمر دنباله به دنباله است که یک تصویر و مختصات دوبعدی از نقاط پرس و جو را به عنوان ورودی برای پیش بینی تناظرها می گیرد. در این مدل با توجه به دو تصویر و یک نقطه پرس و جو در یکی از آنها، تناظر را در دیگری پیدا می کند. با انجام این کار، فرد می‌تواند فقط نقاط مورد نظر را جستجو کند و متناظرهای پراکنده را بازیابی کند، یا تمام نقاط یک تصویر را پرس و جو کند و نگاشت‌های متراکم را به دست آورد.

در LoFTR [17] یک رویکرد درشت به ریز برای تناظریابی با استفاده از کدگذار ترانسفورمر انجام شده است. ابتدا تناظرهای متراکم پیکسلی را در سطح درشت[۱۸] ایجاد می‌کند و پس از آن تناظرهای دقیق[۱۹] را در سطحی پیکسل اصلاح می‌کند. برخلاف روش‌های متراکمی، که از هزینه زیادی برای جستجوی تناظرها استفاده می‌کنند، در این روش از لایه‌های خود توجهی و بین توجهی در ترانسفورمر برای به دست آوردن توصیفگرهای ویژگی هر دو تصویر، استفاده می‌شود.

همه این روش‌ها بر روی یک مجموعه داده بزرگ با موقعیت‌ها و عمق درستی مرجع[۲۰] آموزش داده شده است. طیف گسترده ای از رویکردها با هدف کشف اشیا[۲۱] و مکان آنها از تصاویر بدون برچسب وجود دارد. بسیاری ازاین روش‌ها، از طرح‌های جعبه مرزی[۲۲] استفاده می‌کنند و کشف شی را به عنوان یک مسئله بهینه‌سازی فرموله می‌کنند[18,19]. این موارد معمولاً برای داده‌های تصاویر هوایی سازگار نیستند.

سایر رویکردها بر پیش بینی ماسک اشیاء برجسته[۲۳] به طور مستقیم تمرکز دارند. برخی از آنها برای آموزش به ماسک های پیش زمینه[۲۴] نیاز دارند، در حالی که برخی دیگر برای بخش بندی اشیاء تکراری رایج در یک مجموعه تصویر طراحی شده اند. این رویکردها مفروضات محکمی در مورد فراوانی اشیاء ایجاد می کنند، در حالیکه، در بسیاری از سناریوهای عملی، اشیاء تکراری نادر هستند و کشف آنها مانند به دنبال سوزن در انبار کاه بودن است[20].

[۱۸] Coarse
[۱۹] Fine
[۲۰] Ground truth
[۲۱] Object discovery
[۲۲] Bounding box
[۲۳] Salient objects
[۲۴] Foreground

۳- روش پیشنهادی

روش پیشنهادی پژوهش حاضر شامل سه بخش اصلی است. بخش اول شامل تهیه مجموعه داده وتولید مجموعه داده ساختگی است. در بخش دوم به ارائه مدل هم‌بخش‌بندی[۲۵] پرداخته شده است. در بخش سوم با توجه به مدل هم‌بخش‌بندی آموزش دیده، ابتدا ناحیه‌های مشابه احصا و سپس با توجه به ناحیه‌های مشابه بدست آمده، با استفاده از ویژگیهای SIFT ثبت تصویر انجام می‌شود. در ادامه به صورت کامل به بخش‌های اشاره شده پرداخته می شود.

۳-۱- مجموعه داده آموزش

مجموعه داده‌ای که توسط پژوهشگران گردآوری شده است، برای هم‌بخش بندی تصویر سنجش از دور با وضوح بالا ایجاد شده است، که هدف از آن پیدا کردن ناحیه‌های مشابه در تصاویر ماهواره‌ای به دست آمده در زمان های مختلف است. مجموعه داده ما شامل تصاویر ماهواره‌ای به دست آمده از منابع متعدد، از آرشیوهای سنجش از دور در دسترس عموم Google Earth است.

برای تولید جفت‌های تصاویر آموزشی، ابتدا نیاز به تصاویر هوایی است که، شی‌های اصلی موجود در تصویر را بخش بندی کند. شی‌های اصلی بخش‌هایی از تصویر هستند که نسبت به تغییر زمان ثابت باشند. زیرا برای تطبیق ناحیه‌های مشابه نیاز به بخش‌هایی است که نسبت به زمان بدون تغییر باقی بمانند. بطور مثال اشیائی مانند خودرو، ابر و بطور کلی اشیائی که حرکت می‌کنند شی‌های خوبی برای ناحیه‌های مورد علاقه نیستند. همچنین ناحیه‌های مشابه باید یکتا باشند، به این معنی که ناحیه مورد علاقه باید به قدری بزرگ باشد که احتمال وجود چندین ناحیه مشابه در تصویر تقریبا صفر است. به عنوان مثال بخش‌بندی یک ساختمان، امکان وجود چندین ساختمان مشابه را افزایش می دهد.

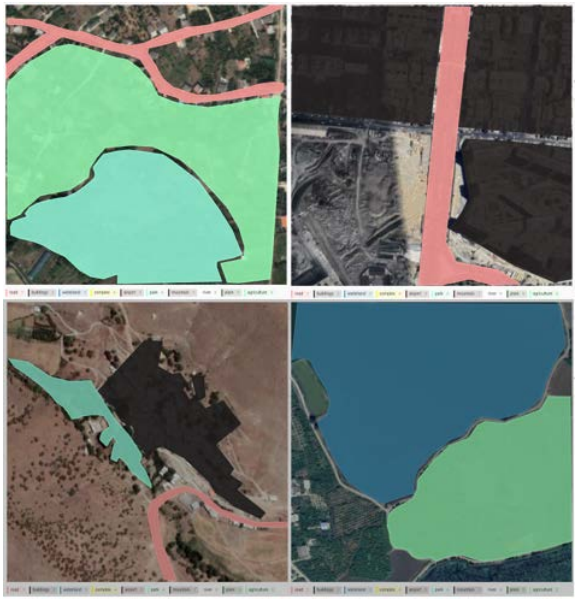
مجموعه داده شامل تصاویر ماهواره‌ای سه کاناله طیف نوری است. هر تصویر معمولاً دارای وضوح مکانی بالایی است که بین ۱ تا ۵ متر در هر پیکسل متغیر است و امکان تجزیه و تحلیل بصری دقیق را فراهم می کند. مجموعه داده شامل تصاویر با وضوح بالا است، که در آن هر تصویر شامل بخش‌بندی معنایی[۲۶]

[۲۵] Co-segmentation
[۲۶] Semantic segmentation

است که بصورت دستی و توسط پژوهشگران پژوهش حاضر انجام شده است. تصاویر در زمان‌های مختلف و یا با حسگرهای نوری متفاوت گرفته شده‌اند. مجموعه داده جمع‌آوری‌شده انواع پوشش زمین، از جمله مناطق شهری، مزارع کشاورزی، جنگل ، آب و مناظر طبیعی در فصول مختلف را شامل می شود. درستی مرجع[۲۷] مجموعه داده با بخش بندی معنایی توسط ما و بصورت دستی برچسب گذاری می شود که نشان دهنده اشیاء داخل تصویر است. این حاشیه نویسی ها[۲۸] با استفاده از نرم افزار label-studio و توسط انسان انجام می‌شوند.

مجموعه داده به مجموعه های آموزشی و اعتبار سنجی تقسیم می شود. مجموعه آموزشی شامل تعداد زیادی جفت تصویر ساختگی برای آموزش و بهینه سازی الگوریتم های ثبت تصویر است که توسط مجموع داده جمع‌آوری شده بدست می‌آید. مجموعه اعتبارسنجی برای تنظیم فراپارامترها و انتخاب مدل استفاده می‌شود.

مجموعه داده جمع آوری شده شامل ۴۰۰ تصویر RGB با اندازه ۵۱۲×۵۱۲ و تفکیک مکانی[۲۹] مختلف است. مجموعه داده دارای ده کلاس است، که شامل: جاده، منطقه ساختمانی، آب، مجتمع، فرودگاه، زمین سبز، مناطق کوهستانی، رودخانه، جلگه و زمین کشاورزی هستند. این کلاس‌ها با توجه به مشخصاتی که در قبل توضیح داده شد، انتخاب شده‌اند. مجموعه‌ای از این تصاویر را در شکل۱ مشاهده می شود.



شکل۱: چهار نمونه از تصاویر نشانه گذاری شده در مجموعه داده تولیدشده.

۳-۱-۱- تولید داده‌های آموزشی با تعویض بخش

برای تولید داده‌های آموزش، از رویکرد تبادل قطعه[۳۰] که در[21] اشاره شده، استفاده شده است. برای این کار، ابتدا از یک تصویر منبع، یک بخش که متعلق به یکی از کلاس‌ها است، نمونه برداری می‌شود. سپس تصویر هدف با اعمال تبدیل‌های هندسی روی ناحیه جداشده و ترکیب آن در یک تصویر پس‌زمینه تصادفی با استفاده از ترکیب پواسون بدست می‌آید[22]. تبدیل‌های هندسی شامل چرخش، انتقال، تغییر اندازه و اسپلاین صفحه نازک[۳۱] (TPS) است.

ترکیب پواسون این امکان را می دهد تا ظاهر تصویر را به صورت یکپارچه، در یک منطقه انتخاب شده تغییر دهد. این تغییرات را می‌توان به گونه ای ترتیب داد که بر روی بافت، روشنایی و رنگ اشیاء موجود در منطقه تأثیر بگذارد و تصویر را یکنواخت کند.

سپس با استفاده از مدل AdaIN [32] [23] آموزش داده شده بر روی مجموعه داده بروگل [24] ، یک انتقال سبک بر روی هر دو تصویر منبع و هدف انجام می شود. انتقال سبک روشی است که شامل تبدیل سبک یک تصویر به تصویر دیگر و در عین حال حفظ محتوای تصویر دوم است. AdaIN یک روش شبکه عصبی را برای انتقال سبک دلخواه در زمان واقعی پیشنهاد می‌کند که میانگین و واریانس ویژگی های محتوا و سبک را تراز می کند. این کار باعث می‌شود مدل نسبت به تغییرات محتوا ثابت شود و تغییراتی مانند فصل و نوع تصویربرداری بر روی عملکرد تاثیر نگذارد.

نمونه ای از جفت داده آموزشی را می توان در شکل۲ مشاهده کرد. در جفت‌های تصویر تولید شده، به ماسک‌های ناحیه‌های مشابه جفت‌های تصویر و همچنین تناظر بین نقاط که به عنوان ناظر برای آموزش کمکی شبکه استفاده می‌شود، دسترسی داریم.

---

[۲۷] Ground truth

[۲۸] Annotations

[۲۹] Spatial resolution

[۳۰] Segment swapping

[۳۱] Thin-plate-spline

[32] Adaptive Instance Normalization

شکل۲: تولید داده توسط تعویض بخش. پس از انتقال پیش‌زمینه از تصویر منبع به تصویر پس زمینه، از ترکیب پواسون و انتقال سبک برای تولید داده آموزشی بهتر استفاده می‌کنیم.

۳-۲- معماری هم‌بخش‌بندی

معماری شبکه به این صورت است که ابتدا یک تصویر منبع I^s و یک تصویر هدف I^t را به عنوان ورودی می‌گیرد، سپس توسط شبکه ستون فقرات[۳۳] استخراج ویژگی، نقشه‌های ویژگی F^s و F^t با ابعاد فضایی $W \times H$ استخراج می‌شوند. سپس این نقشه‌های ویژگی توسط ترانسفورمر متقاطع تصویر پردازش می‌شوند تا هر دو ماسک ناحیه مشابه در تصاویر منبع $M^s \in [0,1]^{W \times H}$ و هدف $M^t \in [0,1]^{W \times H}$ و تناظر بین نقاط از منبع به هدف $C^{s \to t}$ و هدف به منبع $C^{t \to s}$ بدست آید. $C^{s \to t}$ و $C^{t \to s}$ با ماتریس هایی با اندازه $W \times H \times 2$ نشان داده می شوند که در آن آرایه اول، مختصات x و ارایه دوم، مختصات y نقاط متناظر بر روی تصویر اصلی است.

معماری شبکه با استفاده از کتابخانه Pytorch پیاده سازی شده است. PyTorch یک چارچوب[۳۴] یادگیری عمیق منبع باز[۳۵] محبوب است که ابزارها و قابلیت هایی را برای ساخت و آموزش شبکه های عصبی فراهم می کند.

۳-۲-۱ شبکه ستون فقرات

برای ویژگی های ستون فقرات از ویژگی های لایه conv4 یک ResNet-50 [25] آموزش دیده در ImageNet [26] با MOCO-v2 [27] استفاده می کنیم. در زمینه یادگیری عمیق، اصطلاح ستون فقرات معمولاً به یک معماری شبکه عصبی از پیش آموزش دیده اشاره می‌کند که به عنوان پایه‌ای برای یک کار خاص عمل می‌کند. در اینجا، ویژگی های ستون فقرات، نمایش های میانی هستند که از یک لایه خاص از شبکه عصبی به دست می‌آیند.

MOCO-v2 به عنوان ستون فقرات بسیاری از مدل‌های یادگیری خودنظارتی انتخاب می‌شود، زیرا این مدل به طور مؤثری در یادگیری ویژگی‌های باکیفیت و عمومی، بدون برچسب‌های داده، عمل می‌کند.

MoCo-v2به طور خاص برای یادگیری خودنظارتی طراحی و بهینه‌سازی شده است. این مدل از یک مکانیسم کنتراست استفاده می‌کند که امکان می‌دهد نمایش‌های مقاوم به نویز و متمایز برای نمونه‌های مختلف تولید کند. این امر باعث می‌شود مدل بتواند ویژگی‌های کلی را از داده‌ها استخراج کند و نیاز به داده‌های برچسب‌دار را کاهش دهد.

ستون فقرات در طول آموزش منجمد می‌شود، زیرا آموزش پارامترهای زیاد ستون فقرات منجر به بیش برازش[۳۶] بر روی مجموعه آموزشی مصنوعی می‌شود.

۳-۲-۲- ترانسفورمر متقاطع تصویر

معماری شبکه یک معماری مبتنی بر کدگذار ترانسفورمر کلاسیک [27] است که از چند بلوک توجه چند سر[۳۷] و بلوک کاملاً متصل پیشخور[۳۸] (FFN) تشکیل شده است. دلیل استفاده از مدل ترانسفورمر بجای مدل‌های معمول کانولوشنی این است که در این مدل نگاه شبکه بصورت سراسری است و شبکه می‌تواند ارتباط ناحیه‌های مختلف تصویر را آموخته و این در حالی است که در روش‌های معمول، یادگیری بصورت محلی است. بلوک های FFN شامل دو لایه با تابع فعال سازی[۳۹] ReLu[۴۰] هستند.

---

[۳۳] Backbone
[۳۴] Framework
[۳۵] Open source

[۳۶] Overfitting
[۳۷] Multi-headed attention
[۳۸] Feed-forward networks blocks
[۳۹] Activation function
[۴۰] Rectified Linear unit

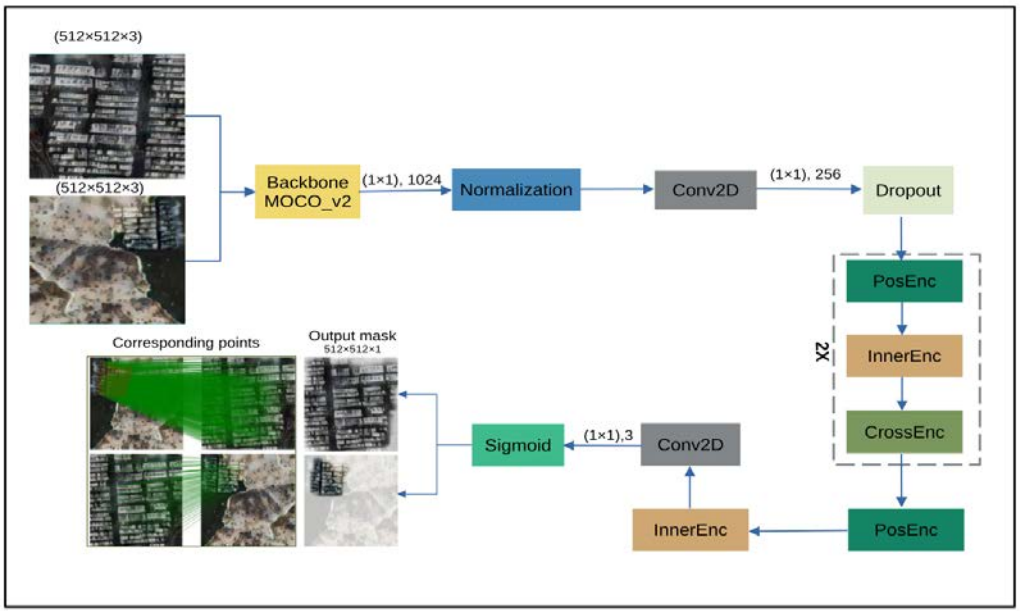
مشابه [15]، در این معماری از دو نوع لایه توجه استفاده شده است: یک لایه استاندارد توجه به خود[41] (SA) که فقط به ویژگی‌های داخل یک تصویر می‌پردازد، دیگری یک لایه توجه متقاطع[42] (CA) است که در آن توجه بین ویژگی‌های بین تصاویر محاسبه می‌شود.

در این معماری از کدگذار مکانی دوبعدی، مشابه [28] برای کدگذاری نقشه ویژگی، قبل از هر بلوک SA استفاده شده است. معماری ترانسفورمر در شکل۳ نشان داده شده است، که شامل پنج بلوک توجه و FFN است. هر لایه توجه دارای ۲ سر و ابعاد ۲۵۶ است. لایه آخر با یک تابع فعال‌سازی sigmoid دنبال می‌شود و دارای سه خروجی است که برای هر تصویر شامل ماسک و تناظر بین نقاط دو تصویر است.

مدل ترانسفورمر طراحی شده دارای ۲ نوع لایه است:

لایه کدگذار توجه درونی(InnerEnc) یک لایه کدگذار ترانسفورمر با توجه درونی را نشان می دهد. پارامترهای ورودی آن شامل:

- d-model به ابعاد نمایش های داخلی مدل اشاره دارد و نشان‌دهنده تعداد ویژگی ها یا کانال ها در نمایش داخلی مدل در هر موقعیت در دنباله ورودی است. در این کار d-model روی ۲۵۶ تنظیم می‌شود و به این معنی است که هر موقعیت در دنباله ورودی با یک بردار ۲۵۶تایی نشان داده می‌شود.
- nhead به تعداد سرها در مکانیسم خودتوجهی چند سر اشاره دارد. مکانیسم خودتوجهی به مدل اجازه این امکان را می‌دهد تا اهمیت موقعیت‌های مختلف را هنگام پیش بینی یک موقعیت معین بهطور متفاوت ارزیابی کند. توجه چند سر این ایده را با استفاده از چندین سر توجه به صورت موازی گسترش می دهد. افزایش تعداد سرها به مدل اجازه می دهد تا انواع مختلفی از وابستگی ها و روابط را در داده ها ثبت کند. در اینجا تعداد سرها روی ۲ تنظیم شده است و خروجی نهایی یک الحاق یا ترکیب خطی از خروجی های این ۲ سر خواهد بود.
- dim-feedforward به ابعاد لایه شبکه عصبی پیشخور، درون بلوک ترانسفورمر اشاره دارد. شبکه عصبی پیشخور به طور مستقل برای هر موقعیت در دنباله اعمال می شود و از دو تبدیل خطی با یک تابع فعال سازی غیر خطی در بین آن تشکیل شده است. پارامتر dim-feedforward ابعاد لایه میانی در این شبکه پیشخور را مشخص می‌کند. در این مدل مقدار آن را روی ۲۵۶ تنظیم کرده‌ایم.
- در زمینه شبکه‌های عصبی، از جمله معماری ترانسفورمر، dropout یک روش منظم‌سازی است که در طول آموزش برای جلوگیری از برازش بیش از حد



شکل۳: عملکرد ترانسفورمر متقاطع تصویر [۲۱] روی جفت های تولید شده آموزشی به همراه ماسک ها و نقاط متناظر.

[41] Self-Attention

[42] Cross attention

استفاده می‌شود. در اینجا مقدار dropout را مساوی ۰/۱ قرار می‌دهیم.

- انتخاب تابع فعال سازی به ویژگی های خاص مسئله در دست، بستگی دارد ولی ReLU یک انتخاب پیش فرض رایج در بسیاری از معماری های یادگیری عمیق، از جمله ترانسفورمرها است.
- pos-weight وزنی است که در هنگام تشکیل بردارهای ویژگی نهایی به عبارت کدگذاری موقعیتی اعمال می شود. کدگذاری موقعیتی به ویژگی های ورودی اضافه می شود تا اطلاعاتی در مورد موقعیت هر عنصر در دنباله به مدل ارائه دهد. با تنظیم این وزن می توان تأثیر ویژگی های اصلی و اطلاعات موقعیتی را در بردارهای ویژگی نهایی کنترل کرد. این کار برای این است که مدل در طول آموزش روی جنبه‌های خاصی از داده‌های ورودی تمرکز بیشتری داشته باشد. مقدار این فراپارامتر را درطول آموزش برابر با ۰.۱ قرار می‌دهیم.
- feat-weight وزنی است که هنگام تشکیل بردارهای ویژگی نهایی به ویژگی های ورودی اصلی اعمال می شود. با تنظیم این وزن ها، می‌توان سهم ویژگی های اصلی و اطلاعات موقعیتی را در بردارهای ویژگی نهایی کنترل کرد. این نوع وزن دهی اگر بخواهیم مدل در طول آموزش بر جنبه های خاصی از داده های ورودی تأکید کند می‌تواند مفید باشد. مقدار این فراپارامتر را درطول آموزش برابر با ۱ قرار داده شده است.

لایه کدگذار توجه متقاطع(CrossEnc) یک لایه کدگذار ترانسفورمر با توجه متقاطع را نشان می دهد. پارامترهای ورودی آن شامل:

- d-model روی ۲۵۶ تنظیم می‌شود و به این معنی است که هر موقعیت در دنباله ورودی با یک بردار ۲۵۶ بعدی نشان داده می‌شود.
- nhead، در این پژوهش تعداد سرها روی ۲ تنظیم شده است و خروجی نهایی یک الحاق یا ترکیب خطی از خروجی های این ۲ سر خواهدبود.
- dim-feedforward، مقدار آن را روی ۲۵۶ تنظیم شده است.
- مقدار dropout را مساوی ۰.۱ تنظیم شده است.
- در این پژوهش نیز از ReLU به عنوان تابع فعال سازی استفاده شده است.

تصاویر ورودی را به شبکه ستون فقرات اعمال شده و سپس تنسور خروجی را، از یک لایه عادی‌سازی عبور داده می شود. هدف از این مرحله عادی سازی، اطمینان از این است که ویژگی ها دارای مقیاس سازگار هستند که می تواند برای آموزش شبکه های عصبی مفید باشد. سپس تنسور خروجی را به مدل ترانسفورمر منتقل می شود. مدل ترانسفورمر به ترتیب شامل لایه‌های، توجه درونی، توجه متقاطع، توجه درونی، توجه متقاطع و توجه درونی است. پس از آن خروجی را از یک لایه

$$L^s = \underbrace{CE(M_{gt}^s, M^s)}_{L_m} + \underbrace{CE(M_{gt}^s, M^t(C^{s\to t}))}_{L_{tm}} + \underbrace{\eta \frac{1}{\sum_{i,j} M_{gt}^s(i,j)} \sum_{i,j} M_{gt}^s(i,j) || C^{s\to t}(i,j) - C_{gt}^{s\to t}(i,j) ||}_{L_{corr}} \quad (۱)$$

کانوولوشنی و تابع فعالسازی سیگموئید[۴۳] عبور داده می شود.

۳-۲-۳- تابع زیان

$$CE(M_{gt}, M) = -\frac{1}{W \times H} \sum_{i,j} M_{gt}(i,j) log(M(i,j)) + (1 - M_{gt}(i,j)) log(1 - M(i,j)) \quad (۲)$$

در داده‌های مصنوعی آموزش[۴۴]، به ماسک‌های درستی مرجع M_{gt}^t و M_{gt}^s و نقاط متناظر درستی مرجع بین تصاویر از منبع به هدف $C_{gt}^{t\to s}$ و از هدف به منبع $C_{gt}^{s\to t}$ ، دسترسی داریم. تابع زیان، مجموع دو عبارت یکسان برای منبع و هدف است، برای سادگی، فقط تابع زیان منبع L^s را می‌نویسیم. که شامل یک مقدار زیان آنتروپی متقاطع[۴۵] (CE) روی ماسک پیش‌بینی‌شده L_m و ماسک انتقال‌یافته L_{tm} و همچنین یک مقدار زیان رگرسیونی L_{corr} در نقاط متناظر بین دو تصویر است:

که در آن i و j مختصات ویژگی‌ها هستند، η یک فراپارامتر اسکالر است و مقدار $CE(M_{gt}, M)$ از رابطه (۲) بدست می‌آید.

باید توجه داشت که این تلفات هم برای جفت‌های مثبت (جفت‌های منبع و هدف ایجاد شده توسط تعویض بخش) و هم

---

[۴۳] Sigmoid activation function

[۴۴] Synthetic training data

[۴۵] Cross-entropy

برای جفت‌های منفی (نمونه‌گیری از دو جفت مختلف، بدون ناحیه مشابه) محاسبه می‌شود که $M_{qt}^{e} = M_{qt}^{t} = 0$ و بر اساس قرارداد $L_{corr} = 0$ است.

برای همه آزمایش‌ها، زیان تعریف‌شده در رابطه (۱)، با توجه به [21]، معادل $\eta = 8$ قرار می‌دهیم و از بهینه‌ساز Adam [29] با شرایط حرکت $\beta1 = 0.5$ و $\beta2 = 0.999$ استفاده می‌کنیم. گرادیان ها با استفاده از پس انتشار محاسبه می شوند و مدل آموزش می‌بینید.

برای آموزش مدل از ۱۰۰۰۰ جفت تصویر تولید شده که دارای ماسک و تناظر بین بخش‌ها است، استفاده شده است. در هر تکرار[۴۶]، از ۵ جفت تصویر دارای ناحیه مشترک و ۱۵ جفت تصویر بدون ناحیه مشترک نمونه برداری شده است. در ابتدا با نرخ یادگیری[۴۷] 2e-6، آموزش شروع می شود و با افزایش نرخ یادگیری تا ۹۰۰۰ تکرار نرخ یادگیری را به 2e-4 رسانده می شود. پس از آن نرخ یادگیری را ثابت نگه داشته و یادگیری تا ۱۷۰۰۰ تکرار ادامه داده می شود.

۳-۳- ثبت تصویر

در این قسمت پس ازآموزش مدل هم‌بخش‌بندی، از آن برای بدست آوردن ناحیه‌های مشابه استفاده شده است. سپس از ناحیه‌های مشابه برای تطبیق تصاویر استفاده می شود. برای تطبیق ناحیه‌های مشابه از روش SIFT-RANSAC[48] [5] استفاده شده است.

الگوریتم SIFT-RANSAC به دلیل مقیاس پذیری، پایداری در برابر چرخش و کشیدگی، مقاوم بودن در برابر تغییرات روشنایی و کنتراست و نویز در تصاویر سنجش از دور بسیار کاربرد داشته است و پس از پیدا کردن ناحیه‌های مشابه روش مناسبی برای یافتن نقاط متناظر و ثبت تصویر است.

ثبت تصویر به وسیله SIFT-RANSAC روشی است که در بینایی کامپیوتری و پردازش تصویر برای تخمین تبدیل هندسی بین دو تصویر استفاده می شود. این الگوریتم استخراج ویژگی SIFT را با الگوریتم RANSAC ترکیب می‌کند و بدین صورت پارامترهای تبدیل، حتی در حضور نقاط پرت یا نویز به طور قوی قابل تخمین است.

الگوریتم SIFT یک روش پرکاربرد برای استخراج ویژگی های متمایز از یک تصویر است. نقاط کلیدی را در یک تصویر شناسایی و توصیف می‌کند که نسبت به مقیاس، چرخش و تبدیل‌های وابسته ثابت هستند. این نقات کلیدی توسط یک بردار توصیفگر نشان داده می شوند که ظاهر محلی آنها را به تصویر می کشد و می تواند در بین تصاویر مختلف مطابقت داده شود.

۴- نتایج

مجموعه داده ارزیابی، برای ثبت تصویر سنجش از دور با وضوح بالا ایجاد شده است، که هدف آن تراز و ثبت دقیق جفت تصاویر ماهواره‌ای به دست آمده در زمان های مختلف است. به دلیل در دسترس نبودن تصاویر با وضوح فضایی بالا برای ارزیابی، بهترین تصاویر موجود برای ارزیابی کمی مدل، تصاویر Sentinel-2 است که دارای وضوح فضایی ۱۰ متر در هر پیکسل است.

مجموعه داده شامل جفت‌هایی از تصاویر است، که در زمان‌های مختلف گرفته شده‌اند، اما منطقه جغرافیایی یکسانی را پوشش می‌دهند. مجموعه داده انواع مختلف پوشش زمین، از جمله مناطق شهری، کشاورزی، جنگل ها و آبها را پوشش می دهد.

جفت‌های تصویر در مجموعه داده، شامل باندهای طیفی قرمز، سبز و آبی است. مجموعه داده با مطابقت های درستی مرجع برچسب گذاری می شود که نشان دهنده تراز مکانی دقیق بین جفت های تصویر است. این حاشیه نویسی ها با استفاده از روش های ارجاع جغرافیایی دقیق[۴۹] توسط نرم‌افزار ENVI و ناظر انسانی انجام شده است. در هر جفت تصویر حس شده و ثبت شده، ۱۰ جفت نقطه متناظر بصورت دستی شناسایی و ثبت می شود و خطا با توجه به فاصله بین هر جفت نقطه مشخصه اندازه گیری می شود. جفت های تصویر در مجموعه داده ارزیابی، دارای تغییرات روشنایی، جابه‌جایی، چرخش ویا تغییر دیدگاه هستند. در جدول۱ مشخصات کامل تصاویر استفاده شده برای ارزیابی، آمده است. در ادامه ابتدا، به ارزیابی و تحلیل مدل بخش‌بندی مشترک پرداخته، سپس، ارزیابی و مقایسه مدل پیشنهادی برای ثبت تصویر بررسی می‌شود.

[۴۶] Iteration
[۴۷] Learning rate
[48] Random Sample Consensus
[۴۹] Accurate georeferencing techniques

جدول۱: مشخصات تصویر استفاده شده برای ارزیابی تصویر

ماهواره	نوع	اندازه تصویر (pixel)	تفکیک پذیری تصویربرداری (meter/pixel)	مکان تصویر	چالش
Sentinel-2	S2A	۷۳۵۳ × ۱۰۹۸۰	۱۰	بندر ماهشهر	تغییر روشنایی، جابه‌جایی
Sentinel-2	S2B	۶۷۰۷ × ۱۰۹۸۰	۱۰	بندر ماهشهر	
Sentinel-2	S2A	۸۰۰۰ × ۸۰۰۰	۱۰	فارس	چرخش، وجود ابر
Sentinel-2	S2B	۷۳۳۹ × ۸۱۴۵	۱۰	فارس	
Sentinel-2	S2A	۱۰۹۸۰ × ۱۰۹۸۰	۱۰	دریاچه نمک	اندازه بزرگ تصویر، جابه‌جایی
Sentinel-2	S2A	۱۰۹۸۰ × ۱۰۹۸۰	۱۰	دریاچه نمک	
Sentinel-2	S2A	۲۰۰۰ × ۲۰۰۰	۱۰	تهران	چرخش، جابه‌جایی، تغییرات زمینی
Sentinel-2	S2B	۲۰۰۰ × ۲۰۰۰	۱۰	تهران	

۴-۱- ارزیابی هم‌بخش‌بندی

پیدا کردن هم‌بخش‌بندی گامی اصلی در ثبت تصویر و بهبود آن در این روش است. برای نشان دادن اثربخشی روش پیشنهادی ابتدا به طور مستقیم مورد بررسی قرار می گیرد. پس از آن، عملکرد ثبت تصویر در مجموعه داده‌های مختلف ارزیابی می شود.

۴-۱-۱- معیار

برای ارزیابی بخش بندی مشترک از رابطه (۳) که در [3] آمده ، استفاده شده است.

$$accuracy = 0.5 \times accuracy_{positive} + 0.5 \times accuracy_{negative} \quad (۳)$$

که $accuracy_{positive} = \frac{T_P}{P}$ ، نشان‌دهنده دقت بخش‌های مشترک شناسایی‌شده است. T_P به تعداد پیکسل‌های مشترک، که به درستی در تصاویر شناسایی شده اند، اشاره دارد و P نشان دهنده تعداد کل پیکسل‌های مشترک در درستی مرجع است.

$accuracy_{negative} = \frac{T_N}{N}$ نشان دهنده دقت بخش‌های غیرمشترک (پس‌زمینه یا مناطق نامربوط) را نشان می‌دهد که به درستی شناسایی شده‌اند. T_N به پیکسل‌های غیرمشترک در تصویر که به درستی شناسایی شده اند، اشاره دارد. N نشان دهنده کل پیکسل‌های غیر مشترک در درستی مرجع است.

این معادلات امکان ارزیابی الگوریتم هم‌بخش‌بندی را با اندازه‌گیری توانایی آن در شناسایی بخش‌های مرتبط و تمایز آن‌ها از پس‌زمینه یا مناطق نامربوط فراهم می‌کند.

۴-۱-۲- نتایج

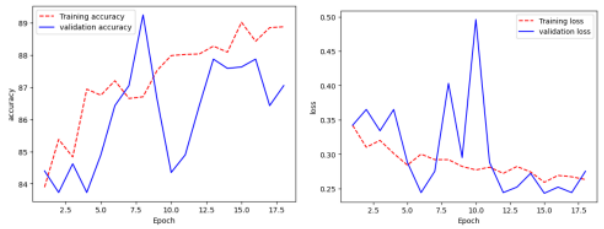
در شکل۴، نتایج بدست آمده از آموزش مدل، اشاره شده است.

در شکل۴-آ، منحنی یادگیری مدل برحسب تابع زیان (۱) در هر دور برای داده آموزش و ارزیابی نشان داده شده است.

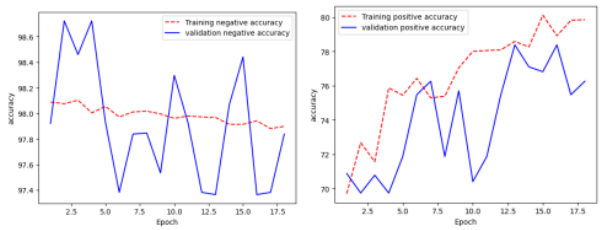
در شکل۴-ب، منحنی یادگیری مدل برحسب مقدار دقت (۳) در هر دور برای داده آموزش و ارزیابی نشان داده شده است.

در شکل۴-پ، منحنی یادگیری مدل برحسب مقدار دقت ناحیه‌های مشترک در هر دور برای داده آموزش و ارزیابی نشان داده شده است.

در شکل۴-ت، منحنی یادگیری مدل برحسب مقدار دقت ناحیه‌های غیرمشترک در هر دور برای داده آموزش و ارزیابی نشان داده شده است. کل تنظیم[۵۰] ترانسفورمر از پیش آموزش داده شده، تقریباً ۴ ساعت بر روی یک GPU واحد nvidia-RTX3060 طول می‌کشد.



(آ) منحنی یادگیری مدل برحسب تابع زیان (ب) منحنی یادگیری مدل برحسب مقدار دقت



(پ) منحنی یادگیری مدل برحسب مقدار دقت ناحیه مشترک (ت) منحنی یادگیری مدل برحسب مقدار دقت ناحیه غیرمشترک

شکل۴: نتایج بدست آمده از مدل آموزشی

با توجه به نمودارهای بدست آمده مدل به خوبی آموزش دیده است. علاوه بر این در شکل۴-ت مدل توانسته است نمونه‌های غیرمشترک را تشخیص دهد. همچنین در شکل۴-پ همانطور که نمایش دارد که مدل توانسته در نمونه‌های مشترک از دقت ۶۲٪ قبل از آموزش، به دقت ۷۹/۸٪ در ایپاک هفده برسد. این

[۵۰] Fine-tune

نشان می‌دهد که مدل به خوبی توانسته مفهوم تصاویر هوایی را درک کند و رابطه بین آنها را پیدا کند.
بخاطر تعداد نمونه‌های داده تست، مدل با نمونه‌های بسیار محدودی ارزیابی می‌شود که ممکن است نماینده‌ی خوبی برای تمام الگوهای موجود در داده‌های واقعی نباشند. به همین دلیل، نوسانات در داده ارزیابی افزایش یافته و واریانس آن زیاد شده‌است. در این حالت، حتی تغییرات کوچک در پارامترهای مدل، نتایج تست را به طور قابل توجهی تغییر می‌دهد.

۴-۲- ارزیابی ثبت تصویر

ارزیابی ثبت گامی مهم در تجزیه و تحلیل تصویر سنجش از دور است. این شامل ارزیابی دقت هندسی و تراز مجموعه داده‌های تصویری است که در زمان‌های مختلف و یا با وضوح‌های متفاوت به دست آمده است. ارزیابی ثبت در سنجش از دور نقش حیاتی در حصول اطمینان از صحت، قابلیت اطمینان و سودمندی داده های به دست آمده ایفا می کند. کیفیت تجزیه و تحلیل را افزایش می دهد، یکپارچگی داده ها را تسهیل می کند و امکان تصمیم گیری موثر در زمینه های مختلف از جمله نظارت بر محیط زیست، کشاورزی، مدیریت بلایای طبیعی، برنامه ریزی شهری و مدیریت منابع طبیعی را فراهم می آورد. در ادامه، معیار ارزیابی و نتایج بدست آمده تشریح می‌شود.

۴-۲-۱ معیار

خطای افکنش یک معیار ارزیابی رایج در ثبت تصویر است که بصورت ریشه میانگین مربعات خطا[۵۱] (RMSE) برای تمام نقاط کلیدی مطابق زیر تعریف می‌شود.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}((lx_i - lx_{i\prime})^2 + (ly_i - ly_{i\prime})^2)}{M}} \quad (۴)$$

که در آن lx_i و ly_i مکان i امین نقاط تطبیق هستند، $lx_{i\prime}$ و $ly_{i\prime}$ موقعیت های پیش بینی شده i امین نقاط تطبیق هستند و M تعداد نقاط تطبیق به دست آمده است. RMSE با اندازه پیکسل نرمال می شود[30] .

۴-۲-۲- نتایج

ارزیابی داده ارزیابی بر روی یک GPU واحد nvidia-GTX1060 انجام شده‌است. برای مقایسه روش پیشنهادی از دو مدل استفاده شده است. یک مدل قدیمی SIFT [5] و یک مدل مبتنی بر شبکه عصبی عمیق[13]، که در بخش کارهای مرتبط به آن اشاره شده است. نتایج در جدول۲ آمده است.
نتایج بدست آمده نشان می‌دهد که با اینکه مدل ثبت تصویر بر روی تصاویر با وضوح بالا آموزش دیده است، در تصاویر آزمایشی با وضوح کمتر به خوبی عمل کرده است. این نتایج نشان می‌دهد که مدل، احتمالاً ویژگی های قوی و متمایز کننده ای را یاد گرفته است که نسبت به تغییرات وضوح ثابت است. همچنین نتایج نشان می‌دهد که در همه موارد افزایش دقت وجود داشته است. دلیل آن هم، این است که فضای جستجو برای پیدا کردن نقاط متناظر و ثبت کاهش پیدا کرده و همچنین از تناظر بین نقاط پرت تا حد امکان جلوگیری شده است. علاوه بر افزایش دقتی که در همه مجموعه داده ارزیابی بدست آمده است، در زمان پردازش هم کاهش قابل توجهی مشاهده شد.

۵- نتیجه‌گیری، مسائل باز و کارهای قابل انجام

از مهمترین مزایای یافتن ناحیه‌های مشابه این است که هنگام تطبیق تصویر، ناحیه جستوجو برای پیدا کردن نقاط متناظر کاهش می‌یابد که به طبع آن هزینه محاسباتی و نرخ تناظر نقاط پرت کاهش می‌یابد و از ناحیه‌های انسداد، مانند ابر برای تناظریابی استفاده نمی‌شود.
در این مقاله از مدلی استفاده شد که ماسک های ناحیه تکرار شده را پیش بینی کند. علاوه بر پیش بینی ماسک‌های متناظر در دو تصویر، پیدا کردن نقاط متناظر بین دو ناحیه مشترک، یک کار کمکی مهم برای آموزش مدل است. برای آموزش مدل از یک معماری مبتنی بر مدل‌های ترانسفورمر از پیش آموخته، استفاده شده است. در این روش از ماسک و نقاط متناظر بین تصاویر ساختگی برای بازیابی و آموزش بخش‌بندی مشترک جفت تصویر بهره گرفته شده است. در ادامه با توجه به ناحیه‌های مشابه بدست آمده، ثبت تصویر با روش SIFT انجام شد.
در ادامه به بررسی نتایج ثبت تصاویر سنجش از دور پرداخته می شود. آزمایش‌ها با استفاده از مجموعه داده‌های حاوی تصاویر گرفته شده توسط حسگر نوری غیرفعال در شرایط محیطی مختلف انجام شد. مجموعه داده شامل تصاویر ماهواره ای از منابع مختلف، مانند آرشیوهای سنجش از دور در دسترس عموم، Sentinel_2 است. مجموعه داده شامل جفت‌هایی از تصاویر است که هر کدام منطقه جغرافیایی یکسانی را نشان می‌دهند اما در زمان های مختلف گرفته شده‌اند. این تصاویر انواع پوشش

[۵۱] Root Mean Square Error

جدول۲: نتایج ارزیابی و مقایسه ثبت تصاویر مجموعه داده با مدل پیشنهادی

تصویر	زمان هم‌بخش‌بندی(ثانیه)	خطای ثبت تصویر (پیکسل)			زمان ثبت تصویر (ثانیه)		
		SIFT+Co-Segmentation	SIFT	مدل [13]	SIFT+Co-Segmentation	SIFT	مدل [13]
دریاچه نمک	۲٫۵	۰٫۵۵۶	۲٫۷۶۲	۱۰٫۴۶	۸٫۵	۱۵٫۲	۳٫۵
فارس	۲٫۲	۰٫۳۶۸	۱٫۳۷۵	۹٫۷۱	۶٫۲	۱۰۲٫۲	۳٫۷
بندر ماهشهر	۲٫۸	۰٫۵۷۷۵	۲٫۷۱۴	۱۲٫۹۵	۸٫۹	۱۷٫۵	۴٫۰
تهران	۲٫۸	۰٫۱۰۶	۰٫۵۶۹	۳٫۲۶	۶٫۳	۲۰۸	۳٫۲

زمین از جمله مناطق شهری، مزارع کشاورزی، جنگل ها، تالاب ها و مناظر طبیعی را در بر می‌گیرد. معیار ارزیابی استفاده شده ریشه میانگین مربعات خطا(RMSE) است. هدف، مقایسه عملکرد الگوریتم ثبت رایج SIFT و یک کار مبتنی بر شبکه عصبی عمیق با مدل پیشنهادی و ارزیابی اثربخشی آن در تراز کردن تصاویر سنجش از دور است.

نتایج بدست آمده نشان می‌دهد که با اینکه مدل ثبت تصویر بر روی تصاویر با وضوح بالا آموزش دیده است، در تصاویر آزمایشی با وضوح کمتر به خوبی عمل کرده است. این نتایج نشان می‌دهد که مدل، احتمالاً ویژگی های قوی و متمایز کننده ای را آموخته که نسبت به تغییرات وضوح ثابت است. همچنین نتایج نشان می‌دهد که در همه موارد افزایش دقت وجود داشته است. دلیل آن هم این است که فضای جست‌جو برای پیدا کردن نقاط متناظر و ثبت کاهش پیدا کرده و همچنین از تناظر بین نقاط پرت تا حد امکان جلوگیری شده است. علاوه بر افزایش دقتی که در همه مجموعه داده تست بدست آمده است، در زمان پردازش هم کاهش قبل توجهی مشاهده شد.

توجه به این نکته الزامی است که در این پژوهش ما به دنبال یک مدل کاربردی برای استفاده در محصولات سنجش از دور بودیم. برای این منظور در تمامی مراحل از داده‌های واقعی چه در زمان آموزش و چه در زمان ارزیابی مدل استفاده کرده ایم.

روش‌های مختلف ثبت مبتنی بر ویژگی برای ترکیب خاصی از تصاویر سنجش از دور پیشنهاد شده‌اند. به عنوان مثال، الگوریتم UR-SIFT برای ثبت تصویر نوری توسعه یافته‌است، در حالی که الگوریتم SAR-SIFT به طور ویژه برای تصاویر SAR پیاده‌سازی شده‌است. با این حال، ثبت تمام ترکیبات ممکن از تصاویر چند حسگر هنوز یک موضوع چالش برانگیز است. روش‌های تبدیلی که برای استخراج ویژگی و تطبیق در ثبت بر اساس ویژگی استفاده می‌شوند، در بسیاری از موارد برای ثبت تصاویر چند حسگر شکست می‌خورند. استخراج و تطبیق ویژگی مبتنی بر یادگیری عمیق می تواند راه حلی ممکن برای رسیدگی به این مسائل باشد.

الگوریتم های تکاملی به طور گسترده ای در ثبت تصاویر سنجش از دور استفاده شده اند و به عملکرد عالی دست یافته اند. یک الگوریتم تکاملی یک الگوریتم بهینه‌سازی سراسری اکتشافی[۵۲] است که ایده آن شبیه‌سازی تکامل بقای بهترین‌ها در اکوسیستم است. از آنجایی که وظیفه ثبت تصویر سنجش از دور می تواند با ایجاد یک مدل معقول به عنوان یک مسئله بهینه سازی در نظر گرفته شود و الگوریتم های تکاملی عملکرد بسیار خوبی در مسائل بهینه سازی دارند، می‌توان از آنها برای پیدا کردن ناحیه مشترک استفاده کرد.

با توجه به توانایی یادگیری قدرتمند یادگیری عمیق، به طور گسترده در وظایف بینایی کامپیوتری مانند ثبت تصویر، تشخیص اشیا، تشخیص تغییر و ترکیب تصویر استفاده شده است. به خصوص با معرفی یک سری از شبکه های استخراج ویژگی برجسته، مانند AlexNet، VGGNet و GoogleNet، ثبت تصویر سنجش از دور مبتنی بر یادگیری عمیق عملکرد رضایت بخشی را به دست آورده است. به منظور بهبود عملکرد مدل، می‌توان شبکه استخراج ویژگی را با یک شبکه که بر روی تصاویر ماهواره‌ای آموزش دیده است جایگزین کرد تا مدل ویژگی‌هایی با معنای بیشتری را دریافت کند و عملکرد آن بهبود یابد.

تعارض منافع

"هیچ‌گونه تعارض منافع توسط نویسندگان بیان نشده است."

مراجع

[1] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally global descriptors for place recognition," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.14141–14152, June 2021.

[2] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.43, no.10, pp.3632–3647, 2021.

[۵۲] Heuristic global optimization

*Computer Vision andPattern Recognition (CVPR) Workshop*s, pp.5082–5092, June 2022.
[22] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Gra*ph., vol.22, pp.313–318, 2003.
[23] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," pp.1510–1519, 2017.
[24] X. Shen, A. Efros, and M. Aubry, "Discovering visual patterns in art collections with spatially-consistent feature learning," pp.9270–9279, 2019.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp.770–778, 2016.
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: a large-scale hierarchical image database," pp.248–255, 2009.
[27] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020.
[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-End Object Detection with Transformers, pp.213–229. 2020.
[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.
[30] H. Goncalves, J. A. Goncalves, and L. Corte-Real, "Measures for an objective evaluation of the geometric correction process quality," *IEEE Geoscience and Remote Sensing Letters*, vol.6, no.2, pp.292–296, 2009.
[31] W. Ma, Y. Wu, Y. Zheng, Z. Wen, and L. Liu, "Remote sensing image registration based on multi-feature and region division," *IEEE Geoscience and Remote Sensing Letters*, vol.14, no.10, pp.1680–1684, 2017.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," *in Advances in Neural Information Processing Systems (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.)*, vol.30, Curran Associates, Inc., 2017.
[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ArXiv*, vol.1409, 09 2014.
[5] D. Lowe, "Object recognition from local scale-invariant features," vol.2, pp.1150 – 1157, 1999.
[6] Y. Ye and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *ISPRS Journal of Photogrammetry and Remote Sensing*,vol.90, p.83–95, 2014.
[7] Z. Hossein-nejad and M. Nasri, "Rkem: Redundant keypoint elimination method in image registration," *IET Image Processing*, vol.11, 2017.
[8] W. Ma, Y. Wu, Y. Zheng, Z. Wen, and L. Liu, "Remote sensing image registration based on multi-feature and region division," *IEEE Geoscience and Remote Sensing Letters*, vol.14, no.10, pp.1680–1684, 2017.
[9] W. Ma, Y. Wu, S. Liu, Q. Su, and Y. Zhong, "Remote sensing image registration based on phase congruency feature detection and spatial constraint matching," *IEEE Access*, vol.6, pp.77554–77567, 2018.
[10] D. Quan, S. Wang, M. Ning, T. Xiong, and L. Jiao, "Using deep neural networks for synthetic aperture radar image registration," pp.2799–2802, 2016.
[11] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, and L. Jiao, "A deep learning framework for remote sensing image registration," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol.145,2018.
[12] F. Ye, Y. Su, H. Xiao, X. Zhao, and W. Min, "Remote sensing image registration using convolutional neural network features," *IEEE Geoscience and Remote Sensing Letters*, vol.PP, pp.1–5, 2018.
[13] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol.PP, pp.1–1, 2018.
[14] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications.," *IEEE Trans. Pattern Anal. Mach. Intell*., vol.33, pp.978–994, 2011.
[15] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. binovich, "Superglue: Learning feature matching with graph neural networks," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[16] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. Yi, "Cotr: Correspondence transformer for matching across images," pp.6187–6197, 2021.
[17] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," pp.8918–8927, 2021.
[18] V. H. Vo, F. Bach, M. Cho, K. Han, Y. Lecun, P. Perez, and J. Ponce, "Unsupervised image matching and object discovery as optimization," 2019.
[19] V. H. Vo, F. Bach, M. Cho, K. Han, Y. Lecun, P. Perez, and J. Ponce, "Unsupervised image matching and object discovery as optimization," 2019.
[20] X. Shen, A. Efros, and M. Aubry, "Discovering visual patterns in art collections with spatially-consistent feature learning," pp.9270–9279, 2019.
[21] X. Shen, A. A. Efros, A. Joulin, and M. Aubry, "Learning co-segmentation by segment swappingfor retrieval and discovery," *in Proceedings of the IEEE/CVF Conference on*