J. Agric. Mech., 11 (1):31-42 (2026)

Journal of
Agricultural Mechanization

Online ISSN: 2717-4107
Journal homepage: https://jam.tabrizu.ac.ir/?lang=en

Original Article

Terrain Classification Using Vision Transformer for Autonomous Off-road Vehicle Navigation Systems

Behzad Golanbari[1], Aref Mardani[1*]

1-Department of Mechanics Engineering of Biosystems, Faculty of Agriculture, Urmia University, Urmia, Iran.

ARTICLE INFO

***Keywords*:**
Autonomous Vehicles,
Intelligent Vehicles,
Environmental Perception,
Machine Vision,
Deep Learning

Received:
December 17, 2025

Revised:
May 18, 2026

Accepted:
May 18, 2026

*** Correspondingauthor:**
Arefardania.mardani@urmia.ac.ir

ABSTRACT

The use of autonomous vehicles in off-road environments is growing. The primary challenge of these systems is to navigate diverse and unpredictable terrain, which significantly impacts vehicle performance. This research develops a terrain classification system for off-road autonomous vehicles using the Vision Transformer (ViT) architecture. Considering the challenges of unstructured environments such as texture diversity, lighting variations, and surface roughness, the proposed model is designed based on ViT-base-patch16-224 and trained on a dataset consisting of 1757 images of 6 terrain classes (soil, rocky, grassy, muddy, gravelly, and hard). The results show that the model achieves an overall accuracy of 97% on the test data and an average F1-score of 0.90. The analysis of the confusion matrix and ROC curves indicates the model's high ability in class discrimination. However, challenges were observed in recognizing the soil class (25% error) due to the visual similarity with muddy terrain. Comparing class-by-class metrics, the model performs better in grass (F1=0.96) and gravel (F1=0.94) than in mud (F1=0.82). This study demonstrates that the ViT attention-based architecture, despite data limitations, possesses a high capability in extracting high-level features from complex images and can be utilized as part of the perceptual system of autonomous vehicles in unstructured environments. The findings suggest that adjusting class-by-class thresholds and increasing training data can improve the model's performance, particularly for complex classes.

Introduction

The application of off-road autonomous vehicles is rapidly expanding across diverse sectors such as agriculture, mining, military operations, forestry, environmental monitoring, and planetary exploration. Unlike urban autonomous vehicles that operate within structured environments defined by clear road markings and infrastructure, off-road vehicles must navigate unpredictable and unstructured terrains. The operational efficiency and safety of these vehicles are heavily dependent on their ability to perceive and interpret the ground they traverse. Consequently, developing robust terrain classification systems is essential for identifying traversable paths and potential hazards. While technical literature often uses terms like terrain awareness, scene understanding, and traversability estimation, the core objective remains the same: utilizing sensory data—such as images, LiDAR point clouds, and radar signals—to interpret the surrounding environment. Historically, Convolutional Neural Networks (CNNs) have dominated machine vision tasks; however, the introduction of the Vision Transformer (ViT) has marked a paradigm shift in image processing. By leveraging self-attention mechanisms, ViTs can capture long-range dependencies and complex spatial relationships within images, a capability that is critical for analyzing intricate off-road environments. A significant challenge in deploying these models, however, is their substantial demand for training data. This research aims to evaluate the efficacy of the ViT architecture in a constrained data environment, specifically without employing data augmentation techniques, to classify various off-road terrains.

How to cite:
Golanbari, B., and Mardani, A. (2026). *Terrain Classification Using Vision Transformer for Autonomous Off-road Vehicle Navigation Systems.* Journal of Agricultural Mechanization, 11(1):31-42. https://doi.org/10.22034/jam.2026.68067.1355

Materials and Methods

In this study, a terrain classification system was developed based on the ViT-base-patch16-224 architecture. The dataset utilized comprised 1,757 images curated from the Roboflow platform, categorized into six primary classes: Soil, Rocky terrain, Grass, Muddy terrain, Gravel terrain, and Rough terrain. The data preparation process involved removing irrelevant images, manual verification of labels, and balancing the number of samples across classes to prevent bias. To rigorously assess the model's generalizability, a completely independent and unseen test dataset consisting of 360 images (60 images per class) was established, ensuring that none of these images were involved in the training or validation phases.

Image preprocessing was executed using the ViTImageProcessor module, which included resizing images to 224×224 pixels, normalization, and conversion into PyTorch tensors. The base ViT model, featuring 12 transformer layers and 12 attention heads, was initialized with weights pre-trained on the ImageNet-21k dataset, and the final layer was modified to accommodate the six target classes. Training was conducted in the PyTorch environment using an NVIDIA 1660-Ti GPU. The training parameters were set to 30 epochs, a batch size of 32, a learning rate of 0.0001, and the AdamW optimizer. Cross-entropy loss was employed as the evaluation metric.

Results and Discussion

Experimental results demonstrated that the proposed model achieved a remarkable overall accuracy of 97% on the independent test dataset, despite the limited data volume. The weighted average F1-score was recorded at 0.90, indicating satisfactory model performance. Analysis of the training curves revealed that the model converged rapidly, with the loss value dropping below 0.01 after just a few epochs, suggesting an absence of overfitting. However, minor fluctuations in validation accuracy were observed, attributed to the model's sensitivity to the specific composition of data batches given the limited dataset size.

A detailed class-by-class performance analysis revealed that the model excelled in identifying Grass (F1=0.96) and Gravel (F1=0.94) terrains. This success is likely due to the distinct visual features characterizing these classes. Conversely, the most significant challenge was observed in distinguishing between the "Soil" and "Muddy" classes. The confusion matrix explicitly highlighted that 25% of Soil samples were misclassified as Muddy. Qualitative analysis indicated that visual similarities in color and texture under specific lighting conditions were the primary cause of this systematic error. It appears the model relied heavily on low-level features and failed to extract higher-level attributes such as moisture content or subtle 3D structural differences. The Receiver Operating Characteristic (ROC) curve, with Area Under the Curve (AUC) values close to 1 (ranging from 0.99 to 1.00), confirmed that the model possesses high discriminative power, suggesting that the observed errors could potentially be mitigated through optimized classification thresholding.

Conclusion

This study confirms that the Vision Transformer (ViT) architecture holds significant potential for feature extraction from complex terrain imagery and can classify various surface types with high accuracy. However, visual similarities between specific classes (e.g., soil and mud) present a challenge that requires targeted solutions. To address these limitations, it is recommended to employ targeted data augmentation techniques (such as moisture simulation and artificial lighting), utilize hybrid architectures (combining CNNs and ViTs) for finer feature extraction, and implement class-specific threshold tuning. Furthermore, integrating multi-modal data, such as thermal or depth imagery, could enhance the distinction between visually similar surfaces. This research provides a framework for debugging deep learning model performance in unstructured environments and paves the way for developing more reliable perception systems for autonomous off-road vehicles.

نشریه
مکانیزاسیون کشاورزی
شاپا الکترونیکی: 2717-4107
درگاه نشریه: https://jam.tabrizu.ac.ir

مقاله پژوهشی

طبقه‌بندی انواع زمین با استفاده از ترنسفورمر بینایی برای سامانه‌های پیمایش خودکار وسایل نقلیه خارج از جاده

بهزاد گلعنبری[۱]، عارف مردانی*[۱]

تاریخ دریافت: ۱۴۰۴/۱۲/۲۵ تاریخ بازنگری: ۱۴۰۵/۰۴/۱۸ تاریخ پذیرش: ۱۴۰۵/۰۴/۲۲

۱- گروه مهندسی مکانیک بیوسیستم، دانشکده کشاورزی، دانشگاه ارومیه، ارومیه، ایران

* نویسنده مسئول: E-mail:a.mardani@urmia.ac.ir

چکیده

این پژوهش به توسعه یک سیستم طبقه‌بندی زمین برای وسایل نقلیه خودران خارج از جاده با استفاده از معماری ترنسفورمر بینایی (ViT) می‌پردازد. مدل پیشنهادی بر پایه ViT-base-patch16-224 طراحی و روی مجموعه‌داده‌ای متشکل از ۱۷۵۷ تصویر از ۶ کلاس زمین آموزش داده شد. نتایج نشان می‌دهد مدل با وجود محدودیت در حجم داده و بدون استفاده از تکنیک‌های افزایش داده، به دقت کلی ۹۷٪ و میانگین F1-Score معادل ۰/۹۰ دست یافته است. با این حال، تحلیل ماتریس آشفتگی یک چالش اصلی را آشکار کرد: شباهت بصری بین زمین خاکی و گلی منجر به خطای طبقه‌بندی ۲۵٪ در کلاس خاکی شد. این یافته نشان‌دهنده‌ی محدودیت مدل در تمایز کلاس‌های با ویژگی‌های ظاهری مشابه است. اگرچه مدل در شناسایی زمین چمنی (F1=۰/۹۶) و سنگریزه‌ای (F1=۰/۹۴) عملکرد بهتری داشت، اما نتایج به وضوح نشان می‌دهد که تنظیم آستانه‌های کلاس‌به‌کلاس و به‌ویژه افزایش داده‌های آموزشی برای کلاس‌های مشکل‌دار، راهکار کلیدی برای بهبود بیشتر عملکرد است. این مطالعه قابلیت بالای ViT را در استخراج ویژگی از تصاویر پیچیده تأیید می‌کند، اما بر ضرورت استفاده از راهکارهای هدفمند برای غلبه بر محدودیت‌های داده در محیط‌های بدون ساختار تأکید می‌ورزد.

کلمات کلیدی: وسایل نقلیه خودران، وسایل نقلیه زمینی هوشمند، درک محیط، بینایی ماشین، یادگیری عمیق.

۱- مقدمه

در ادبیات، تکنیک‌های طبقه‌بندی زمین مبتنی بر بینایی عمدتاً از طبقه‌بندی‌کننده‌های دودویی برای تمایز بین زمین‌های جاده‌ای و غیر جاده‌ای استفاده کرده‌اند (Kim et al., 2007; Selvathai et al., 2017). چنین رویکردهایی برای سناریوهای ناوبری خارج از جاده که با انواع مختلف زمین مشخص می‌شوند، ناکافی یا بیش‌ازحد ساده‌انگارانه هستند. در طبقه‌بندی زمین، روش‌های یادگیری عمیق معمولاً به‌عنوان وظایف تشخیص منطقه قابل رانندگی و فضای آزاد فرموله می‌شوند که تقسیم‌بندی معنایی رویکرد رایج‌تری است (Jin et al., 2021; Qiu & Jiang, 2025; Sgibnev et al., 2020; Viswanath et al., 2021; Wu & Liu, 2024). یکی از اولین رویکردها برای پرداختن به چالش کمبود داده‌ها در طبقه‌بندی زمین‌های خارج از جاده توسط هولدر و همکاران، (۲۰۱۶) پیشنهاد شد، که در آن از یادگیری انتقالی برای بهره‌برداری از دانش مدل‌های از پیش آموزش‌دیده استفاده شد. یکی از معایب اصلی مدل‌های مبتنی بر شبکه‌های عصبی کانولوشنی (Convolutional Neural Networks; CNN) در طبقه‌بندی زمین، پیچیدگی محاسباتی بالای آنهاست که استقرار بلادرنگ در سامانه‌های تعبیه‌شده با محدودیت منابع را چالش‌برانگیز می‌کند. برای پرداختن به این موضوع، تحقیقات بعدی در طبقه‌بندی زمین‌های خارج از جاده بر توسعه معماری‌های سبک‌وزن و درعین‌حال به حداقل رساندن افت دقت متمرکز شده است. پنگ و همکاران، (۲۰۲۴) برای پیشبرد بیشتر کارایی محاسباتی، شبکه‌ای به نام شبکه قطعه‌بندی با کمک مرز سبک (LB-ASN) معرفی کردند که ضمن حفظ وزن بسیار سبک مدل، به‌دقت بالایی دست می‌یابد. پس از موفقیت شبکه‌های تبدیل‌کننده (Transformers) در پردازش زبان طبیعی، شبکه‌های بینایی ترنسفورمر (Vision Transformers; ViT) به عنوان جایگزینی قدرتمند برای CNNها در وظایف بصری ظاهر شدند. این شبکه‌ها بر پایه مکانیسم خود-توجهی (self-attention) طراحی شده‌اند که به مدل اجازه می‌دهد تا روابط طولانی‌برد و ویژگی‌های پیچیده مکانی بین بخش‌های مختلف تصویر را استخراج کند. ViTها توانایی تشخیص الگوهای ظریف و اجزای کلیدی تصاویر پیچیده را دارند و برای تحلیل زمین‌های خارج از جاده، استخراج ویژگی‌های سطح بالا و طبقه‌بندی دقیق مسیرهای قابل پیمایش به کار گرفته می‌شوند (Dosovitskiy et al., 2020; Vaswani et al., 2017). ViTها از مکانیسم‌های خود-توجهی بهره می‌برند و به مدل اجازه می‌دهند تا بر روی قسمت‌های مختلف یک تصویر در سطوح مختلف جزئیات تمرکز کند. این امر توانایی ثبت وابستگی‌های دوربرد و روابط مکانی را بهبود می‌بخشد، که برای تقسیم‌بندی معنایی دقیق در محیط‌های پیچیده مانند زمین‌های خارج از جاده بسیار مهم هستند. تحقیقات اخیر فایکوس و همکاران (۲۰۲۴) پتانسیل ViTs را برای درک خارج از جاده نشان دادند. با استفاده از مدل EfficientViT، یک معماری سبک و سخت‌افزار-کارآمد برای تحلیل زمین‌های قابل پیمایش در مجموعه داده CaT (CAVS Traversability Dataset) توسعه یافت؛ CaT یک مجموعه داده برچسب‌خورده برای طبقه‌بندی مسیرهای قابل پیمایش خودروها در محیط‌های ناهموار است که تصاویر آن با توجه به قابلیت عبور انواع مختلف خودروها (مثل sedan، pickup، off-road) برچسب‌گذاری شده‌اند. استنتاج این مدل روی سخت‌افزارهای لبه تا چهار برابر سریع‌تر از مدل‌های CNN مانند SwiftNet گزارش شده است (Cai et al., 2023). بلوک‌های ترانسفورماتور همچنین در معماری‌های شبکه ترکیبی، که دارای ستون فقرات ResNet یا پرسپترون‌های چند لایه (MLP) هستند، ترکیب شده‌اند (Wu & Liu, 2024).

اکثر مطالعات به دقت کلی بالا گزارش شده می‌پردازند، اما تحلیل کیفی شکست‌های مدل و شناسایی نقاط کور خاص آن (مانند خطای سیستماتیک بین دو کلاس) اغلب نادیده گرفته می‌شود. این تحلیل برای توسعه سیستم‌های قابل اعتماد در دنیای واقعی، که در آن هزینه یک خطای طبقه‌بندی می‌تواند بسیار بالا باشد، حیاتی است.

هدف اصلی این تحقیق، توسعه یک سیستم طبقه‌بندی دقیق و کارآمد برای شناسایی انواع زمین‌های خارج از جاده با استفاده از مدل ViT و تکنیک‌های یادگیری عمیق است. با توجه به چالش‌های موجود در محیط‌های خارج از جاده، ازجمله تغییرات روشنایی، ناهمواری‌های سطح زمین و به‌ویژه شباهت بصری بین کلاس‌های خاص (مانند خاکی و گلی) که عملکرد وسیله نقلیه را به‌شدت تحت تأثیر قرار می‌دهد، این پژوهش به ارزیابی دقیق این می‌پردازد که یک مدل استاندارد ViT که با داده‌های کاملاً برچسب‌خورده آموزش دیده، تا چه حد می‌تواند بین انواع زمین‌های خارج از جاده که از نظر بصری شبیه هستند، تمایز قائل شود. به طور خاص بر شناسایی و تحلیل کمّی دقیق‌ترین موارد خطا تمرکز کرده و بینش‌هایی ارائه می‌دهد که مستقیماً به راهکارهای عملی مانند افزایش داده هدفمند و تنظیم آستانه منجر می‌شود. بنابراین، سهم این مطالعه نه در معرفی یک معماری جدید، بلکه در اعتبارسنجی عمیق و ارائه یک چارچوب تحلیلی برای عیب‌یابی عملکرد ViT در یک سناریوی کاربردی و پرچالش است.

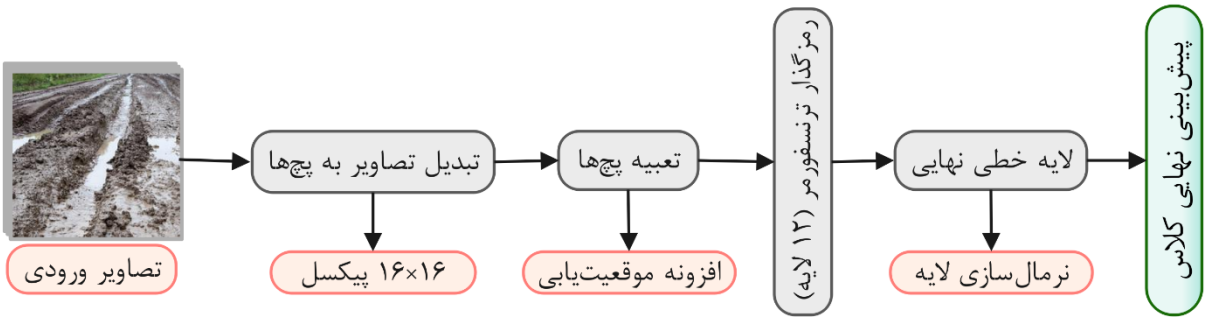
۲- مواد و روش ها

۲-۱- تقسیمات مقاله

در این پژوهش، به‌منظور طبقه‌بندی انواع مختلف زمین برای کاربرد در سامانه‌های خودران خارج از جاده، از مدل ViT استفاده شده است. از یک مجموعه داده سفارشی‌سازی‌شده شامل شش کلاس زمین شامل خاکی (Soil)، صخره‌ای (Rocky terrain)، چمنی (Grass)، گلی (Muddy terrain)، سنگریزه‌ای (Gravel terrain) و زمین سخت (Rough terrain) استفاده شده است. مجموعه داده اصلی مورد استفاده برای آموزش و ارزیابی اولیه، از ادغام و پالایش چندین

مجموعه داده عمومی موجود در پلتفرم Roboflow گردآوری شده است. فرآیند سفارشی‌سازی شامل مراحل زیر بود: (۱) استخراج تصاویر مرتبط با کلاس‌های شش‌گانه هدف از مجموعه‌داده‌های مختلف، (۲) حذف تصاویر نامرتبط یا تکراری، (۳) بازبینی و تأیید دستی برچسب‌های هر تصویر برای اطمینان از صحت، و (۴) متعادل‌سازی مصنوعی تعداد نمونه‌ها در هر کلاس تا از سوگیری (Bias) در آموزش مدل جلوگیری شود. حاصل این فرآیند، مجموعه‌داده‌ای متشکل از ۱۷۵۷ تصویر بود. برای کلاس اول (زمین‌هایی با پوشش سبز و چمنی) ۲۹۱ تصویر، برای کلاس دوم (زمین‌هایی با پوشش سنگریزه‌ای) ۲۷۱ تصویر، برای کلاس سوم (زمین‌های گلی) ۲۹۴ تصویر، برای کلاس چهارم (زمین‌های سنگلاخی و صخره‌ای) ۲۸۹ تصویر، برای کلاس پنجم (زمین‌های سفت) ۳۰۵ تصویر و برای کلاس ششم (زمین‌های خاکی) ۳۰۷ تصویر در نظر گرفته شد که یک توزیع متعادل را در برمی‌گیرد. داده‌ها به‌صورت تقسیم دستی ثابت (Manual Fixed Split) به سه بخش آموزش، اعتبارسنجی و آزمون تفکیک شده‌اند. ساختار مجموعه داده به‌گونه‌ای طراحی شده که هر کلاس در یک زیرپوشه جداگانه قرار دارد و کلاس‌ها از طریق نام پوشه به‌صورت خودکار شناسایی می‌شوند. علاوه بر این، برای ارزیابی نهایی و سنجش قابلیت تعمیم‌پذیری (Generalization) مدل در محیطی کاملاً جدید، یک مجموعه داده کاملاً مستقل و جداگانه متشکل از ۳۶۰ تصویر (۶۰ تصویر از هر کلاس) تهیه شد. این تصاویر در هیچ یک از مراحل آموزش یا اعتبارسنجی استفاده نشدند. بنابراین، این داده‌ها دیده نشده (Unseen) هستند و عملکرد گزارش‌شده در تحقیق بر اساس این مجموعه داده مستقل است که اعتبار نتایج را افزایش می‌دهد.

برای پیش‌پردازش تصاویر، از ماژول ViTImageProcessor متعلق به کتابخانه Transformers شرکت Hugging Face استفاده شده است. این ماژول فرآیندهایی نظیر تبدیل تصاویر به فضای رنگی RGB، تغییر اندازه به ابعاد استاندارد ورودی مدل (۲۲۴×۲۲۴ پیکسل)، نرمال‌سازی مقادیر پیکسل‌ها و تبدیل به قالب تانسوری PyTorchرا به‌صورت خودکار انجام می‌دهد. برای مدیریت داده‌ها، یک کلاس داده سفارشی با نام TerrainDataset طراحی شده است که مسئول بارگذاری تصاویر و برچسب‌های مربوطه در طول فرآیند آموزش و اعتبارسنجی است.



شکل ۱- ساختار کلی مدل ViT مورد استفاده در این پژوهش

Fig 1. Overall architecture of the ViT model employed in this study

مدل مورد استفاده، نسخه پایه ViT با ساختار vit-base-patch16-224 است. ساختار کلی مدل ViT مورد استفاده در این پژوهش در شکل ۱ نشان داده شده است. این ساختار شامل ۱۲ لایه ترنسفورمر، ۱۲ هد توجه و یک لایه MLP با اندازه ۳۰۷۲ است که وزن‌های اولیه آن از پیش‌آموزش‌دیده روی مجموعه داده ImageNet-21k بارگذاری شده‌اند. به‌منظور انطباق با مسئله طبقه‌بندی زمین، لایه نهایی مدل با یک لایه خطی جدید با تعداد خروجی برابر با شش (تعداد کلاس‌ها) جایگزین شده است. مدل انتخاب شده با توجه به توان سخت‌افزاری در دسترس بود.

فرآیند آموزش مدل در محیط PyTorch و با استفاده از کارت گرافیکی NVIDEA 1660-Ti با حافظه ۶ گیگابایت انجام شده است. پارامترهای آموزشی شامل اندازه دسته ۳۲، تعداد ۳۰ دوره آموزشی، نرخ یادگیری ۰/۰۰۰۱ و استفاده از بهینه‌ساز AdamW با ضریب کاهش وزن ۰/۰۱ بوده‌اند. تابع خطای مورد استفاده، آنتروپی متقاطع چند کلاسه است. در هر دوره، مدل طی دو مرحله آموزش و اعتبارسنجی اجرا می‌شود و بهترین وزن‌های مدل بر اساس بالاترین دقت در مجموعه اعتبارسنجی ذخیره می‌شوند.

برای ارزیابی عملکرد مدل، از معیارهای استاندارد مانند دقت (Accuracy)، دقت مثبت (Precision)، بازخوانی (Recall) و میانگین هماهنگ آن‌ها (F1-Score) به‌صورت کلاس به کلاس و میانگین وزنی استفاده شده است.

برای شفاف‌سازی نحوه محاسبه شاخص‌های ارزیابی، در این پژوهش از چهار معیار اصلی استفاده شد. دقت کلی یا Accuracy نسبت تعداد پیش‌بینی‌های صحیح به کل نمونه‌های آزمون را نشان می‌دهد و از رابطه ۱ محاسبه می‌شود.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱)$$

در این رابطه، TP تعداد نمونه‌هایی است که به‌درستی به‌عنوان کلاس مثبت تشخیص داده شده‌اند، TN تعداد نمونه‌هایی است که

به‌درستی به‌عنوان کلاس منفی تشخیص داده شده‌اند، FP تعداد نمونه‌هایی است که به‌اشتباه مثبت تشخیص داده شده‌اند و FN تعداد نمونه‌هایی است که به‌اشتباه منفی تشخیص داده شده‌اند. شاخص Precision نشان می‌دهد از میان نمونه‌هایی که مدل به یک کلاس نسبت داده است، چه تعداد واقعاً متعلق به همان کلاس بوده‌اند و از رابطه ۲ محاسبه می‌شود.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۲)$$

شاخص Recall بیان می‌کند که مدل چه نسبتی از نمونه‌های واقعی یک کلاس را به‌درستی شناسایی کرده است و از رابطه ۳ قابل محاسبه است.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (۳)$$

در نهایت، F1-Score میانگین هماهنگ Precision و Recall است و زمانی اهمیت بیشتری دارد که لازم باشد تعادل بین خطای مثبت کاذب و منفی کاذب بررسی شود. این شاخص از رابطه ۴ به دست می‌آید.

$$\text{F1} - \text{Score} = 2 \times \frac{Precision.Recall}{Precision + Recall} \quad (۴)$$

با توجه به ماهیت چندکلاسه مسئله، این معیارها برای هر کلاس به‌صورت جداگانه محاسبه شدند و سپس میانگین وزنی آنها با در نظر گرفتن تعداد نمونه‌های هر کلاس گزارش شد.

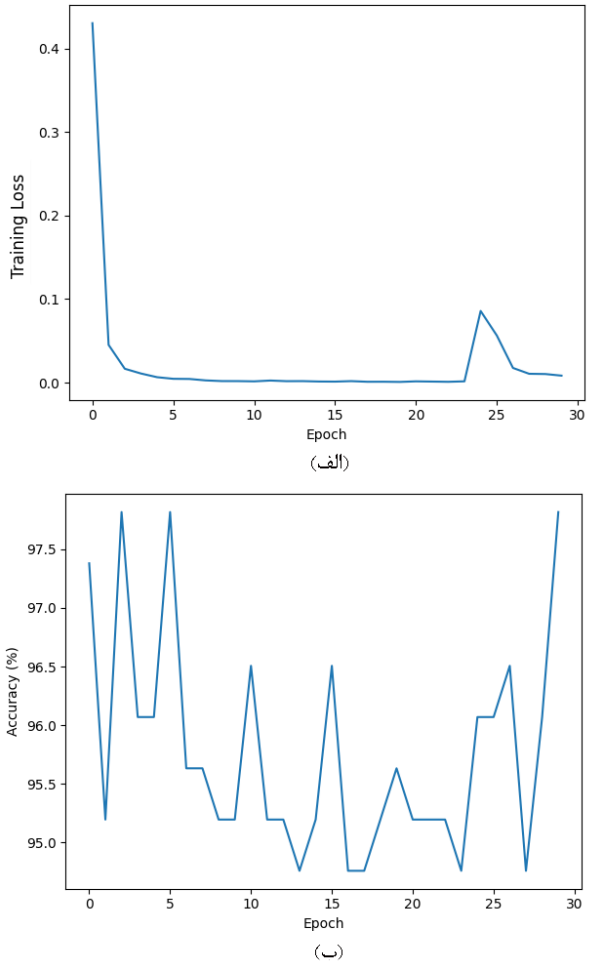
علاوه‌براین، ماتریس آشفتگی برای بررسی توانایی مدل در تمایز بین کلاس‌های مشابه ترسیم شده است. همچنین برای تحلیل عمیق‌تر عملکرد، نمودارهای روند آموزش شامل تغییرات خطا و دقت در طول دوره‌های آموزشی و نیز نتایج پیش‌بینی بر روی مجموعه آزمون ارائه شده‌اند.

درنهایت، گزارش طبقه‌بندی نهایی به همراه فایل وزن‌های مدل، نمودارهای آموزشی و خروجی‌های تصویری از پیش‌بینی‌های انجام‌شده برای نمونه‌های آزمون ذخیره و تحلیل شده‌اند. استفاده از مدل ViT در قالب انتقال یادگیری منجر به بهبود چشم‌گیر در سرعت آموزش و دقت نهایی در طبقه‌بندی تصاویر شده است. این نتایج نشان می‌دهند که معماری مبتنی بر توجه (Attention) در ViT قابلیت بالایی در استخراج ویژگی‌های سطح بالا از تصاویر پیچیده زمین دارد و می‌تواند نقش مؤثری در سامانه‌های درک محیطی ماشین‌های خودران ایفا کند.

۳- نتایج و بحث

در این بخش، عملکرد مدل ViT در طبقه‌بندی انواع مختلف زمین مورد ارزیابی و تحلیل قرار می‌گیرد. با توجه به اهمیت دقت و قابلیت تعمیم مدل در کاربردهای عملی مانند سامانه‌های خودران خارج از جاده، بررسی کمی و کیفی نتایج به‌دست‌آمده ضروری است. بدین منظور، معیارهای استاندارد ارزیابی شامل دقت کلی، دقت طبقه‌ای، بازخوانی، میانگین F1-Score و همچنین ماتریس آشفتگی مورد استفاده قرار گرفته‌اند تا نقاط قوت و ضعف مدل در تشخیص کلاس‌های مختلف زمین شناسایی شود.

شکل ۲-الف روند تغییرات خطا (loss) در مجموعه آموزش و شکل ۲-ب دقت در مجموعه اعتبارسنجی را طی ۳۰ دوره آموزشی (epoch) نمایش می‌دهد. همان‌طور که در نمودار خطا مشاهده می‌شود، مدل با سرعت قابل قبولی به همگرایی رسیده و از ابتدای آموزش، کاهش چشم‌گیری در مقدار خطا رخ داده است؛ به‌طوری‌که مقدار loss پس از تنها چند دوره به کمتر از ۰/۰۱ رسیده و در ادامه تقریباً پایدار باقی مانده است که نشانه‌ای از یادگیری مناسب و عدم بیش برازش است.



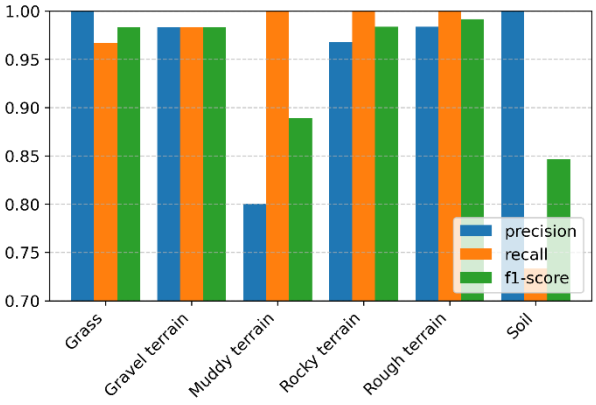
شکل ۲- روند تغییرات خطا (الف) و دقت یادگیری مدل (ب) در مرحله آموزش

Fig 2. Variation of the training error (a) and learning accuracy (b) during the training phase

نوسانات مشاهده‌شده در دقت اعتبارسنجی (شکل ۲-ب) را می‌توان عمدتاً به محدودیت حجم داده‌های آموزشی و عدم استفاده از افزایش داده (Data Augmentation) نسبت داد. با وجود توزیع متعادل، حجم کلی داده (۱۷۵۷ تصویر) برای آموزش یک مدل پارامتر-حجیم مانند ViT-base ممکن است در برخی دوره‌ها منجر به برازش ناقص (underfitting) جزئی شود، به خصوص زمانی که دسته‌ای (batch) از نمونه‌های دشوار یا غیرمعمول برای مدل ارائه می‌شود. این نظریه با مشاهده اینکه نوسانات پس از دوره‌های اولیه کاهش می‌یابد (زمانی که مدل بر ویژگی‌های اصلی تسلط می‌یابد) و نیز با عملکرد پایدار مدل بر روی مجموعه آزمون مستقل (۹۷٪ دقت) تأیید می‌شود. عملکرد مناسب روی داده آزمون نشان می‌دهد که این نوسانات نشان‌دهنده بیش برازش نیست، بلکه نشان‌دهنده حساسیت مدل به ترکیب دقیق داده در هر دسته در حین آموزش است که با افزایش حجم داده یا استفاده از افزایش داده می‌توان آن را کاهش داد.

نمودار ارائه‌شده در شکل ۳ معیارهای دقت (Precision)، بازخوانی (Recall) و امتیاز F1 را برای هر یک از ۶ کلاس زمین روی تصاویر دیده نشده نمایش می‌دهد. این نمودار نشان می‌دهد که مدل چگونه در تمایز بین انواع سطوح با ویژگی‌های مشابه عمل کرده است.

مدل به میانگین وزنی امتیاز F1 حدود ۰/۹۰ دست یافته که نشان‌دهنده عملکرد مناسب در طبقه‌بندی چند کلاسه است. بالاترین امتیاز F1 مربوط به کلاس زمین چمنی با مقدار ۰/۹۶ است که احتمالاً به دلیل ویژگی‌های بصری متمایز مانند بافت یکنواخت و رنگ سبز مشخص است. از سوی دیگر، کلاس زمین گلی با امتیاز F1 پایین‌تر حدود ۰/۸۲ بیشترین چالش را برای مدل ایجاد کرده است.



شکل ۳- نمودار دقت، بازخوانی و امتیاز F1 مدل روی تصاویر دیده نشده

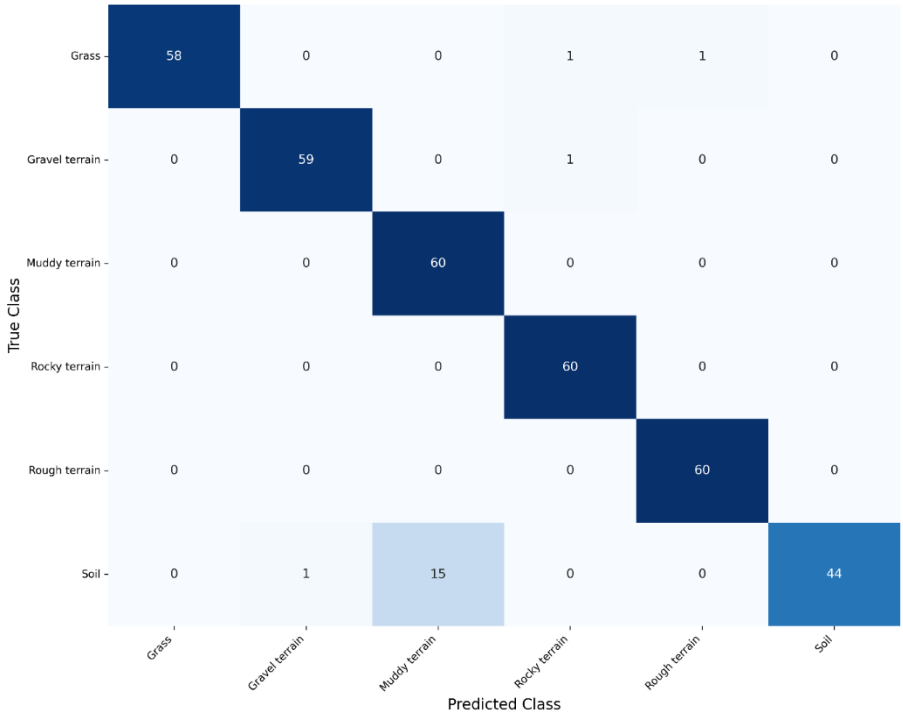
Fig 3. Performance metrics—accuracy, recall, and F1-score—of the model evaluated on unseen images

زمین گلی پایین‌ترین بازخوانی معادل ۰/۷۸ را ثبت کرده است که نشان می‌دهد مدل در شناسایی کامل این کلاس ضعف دارد. این احتمالاً ناشی از شباهت بصری با کلاس خاکی است. زمین سفت با دقت نسبتاً پایین (۰/۸۴)، نشان می‌دهد مدل گاهی نمونه‌های این کلاس را با سنگلاخی یا سنگریزه‌ای اشتباه می‌گیرد. کلاس چمنی با بالاترین دقت (۰/۹۷) نشان می‌دهد که مدل در پیش‌بینی این کلاس با چالش جدی روبرو نیست. کلاس سنگریزه‌ای دارای تعادل عالی بین دقت (۰/۹۳) و بازخوانی (۰/۹۴) نشان‌دهنده توانایی مدل در هم شناسایی و هم تمایز این کلاس است.

نمودار ماتریس آشفتگی ارائه‌شده در شکل ۴ نشان می‌دهد مدل در اکثر کلاس‌ها عملکردی قابل‌قبول داشته است. پنج کلاس از شش کلاس (چمنی، سنگریزه‌ای، گلی، سنگلاخی و سفت) با دقتی بیش از ۹۶٪ شناسایی شده‌اند که نشان‌دهنده توانایی بالای مدل در تمایز این سطوح است. به‌ویژه، کلاس‌های گلی، سفت و سنگلاخی با دقت ۱۰۰٪ شناسایی شده‌اند که نشان می‌دهد ویژگی‌های بصری متمایز این کلاس‌ها به‌خوبی توسط مدل تشخیص داده می‌شوند.

کلاس خاکی تنها کلاسی است که با چالش‌های جدی مواجه بوده است. ۱۵ نمونه از کلاس خاکی به‌اشتباه به‌عنوان گلی طبقه‌بندی شده‌اند (۲۵٪ خطا) ۱ نمونه نیز به‌اشتباه سنگریزه‌ای تشخیص داده شده است. این خطاهای سیستماتیک نشان می‌دهد که مدل در تمایز بین خاک خشک و زمین‌های گلی با مشکل مواجه است، منبع احتمالی این خطاها می‌تواند شباهت رنگ و بافت در شرایط خاص نوری باشد. ۱ نمونه از کلاس چمنی به‌اشتباه به‌عنوان سنگی و ناهموار طبقه‌بندی شده است و ۱

نمونه از کلاس سنگریزه‌ای به‌عنوان سنگی تشخیص داده شده است. این خطاهای پراکنده احتمالاً ناشی از موارد مرزی (borderline cases) در داده‌های آزمون بوده‌اند که حتی برای چشم انسان نیز ممکن است چالش‌برانگیز باشند.



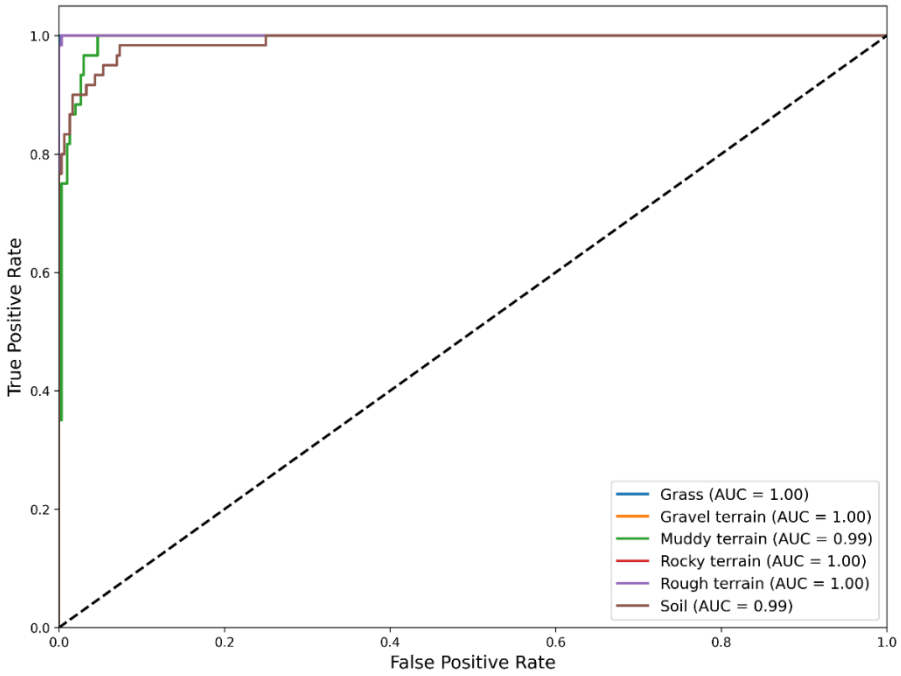
شکل ۴- نمودار آشفتگی مدل برای تشخیص کلاس‌های مختلف در مرحله آزمون

Fig 4. Confusion matrix illustrating the model's classification performance across different classes in the testing phase

بررسی کیفی نمونه‌های اشتباه طبقه‌بندی شده نشان می‌دهد که شباهت رنگ و بافت در شرایط نوری خاص، عامل اصلی خطای مدل بین کلاس‌های خاکی و گلی است. به عنوان مثال، خاک خشک و روشن اغلب دارای رنگ قهوه‌ای مایل به زرد و بافتی یکدست است که در شرایط خاص، با زمین گلی که آب خود را از دست داده و ترک‌خورده است، شباهت بصری زیادی پیدا می‌کند. از طرف دیگر، زمین گلی تیره و مرطوب نیز ممکن است با خاک اشباع شده اشتباه گرفته شود. به نظر می‌رسد مدل فعلی بیش از حد به ویژگی‌های سطح پایین مانند رنگ و بافت کلی تکیه دارد. در حالی که برای تمایز قاطعانه بین چنین کلاس‌های مشابهی، مدل نیازمند یادگیری نشانه‌های ظریف‌تر و مفهومی‌تر است؛ نشانه‌هایی مانند حالت رطوبت و درخشندگی سطح، الگوهای خاص ترک‌خوردگی، یا ساختار سه‌بعدی ریزناهمواری‌ها که حتی برای چشم انسان نیز در نگاه اول ممکن است ناپیدا باشند. این ویژگی‌های سطح بالا معمولاً از ترکیب پیچیده‌ای از ویژگی‌های سطح پایین در لایه‌های عمیق‌تر شبکه به دست می‌آیند. به نظر می‌رسد حجم داده‌های آموزشی حاضر برای یادگیری مطمئن این ویژگی‌های بسیار پیچیده کافی نبوده است. شایان ذکر است که مدل مورد استفاده در این پژوهش صرفاً تصاویر RGB را تحلیل می‌کند و قادر به تشخیص مستقیم ویژگی‌های غیرتصویری مانند رطوبت دقیق خاک یا درصد ترکیبات آن نیست. برای تحلیل چنین ویژگی‌هایی، نیاز به داده‌های تکمیلی مانند تصاویر چندطیفی، حرارتی یا اطلاعات عمق (Depth) و آزمایش‌های جداگانه است. این فرضیه با مشاهده بهبود عملکرد مدل در کلاس‌های متمایزتر (مانند چمن و سنگ) که به ویژگی‌های سطح پایین‌تر وابسته هستند، تایید می‌شود.

منحنی ROC ارائه‌شده نشان‌دهنده در شکل ۵ عملکرد مناسب مدل در تمایز بین انواع سطوح زمین است. چهار کلاس از شش کلاس (چمنی، سنگریزه‌ای، سنگی، سفت و خاکی) به مقدار AUC=۱/۰۰ دست یافته‌اند که نشان‌دهنده قدرت تفکیک کامل مدل برای این کلاس‌هاست. کلاس‌های گلی و خاکی با AUC=۰/۹۹ عملکردی اندکی پایین‌تر دارد، اما همچنان در محدوده بسیار مناسبی قرار می‌گیرد. این نتایج تأیید می‌کند که مدل قادر است به‌خوبی بین انواع زمین‌ها تمایز قائل شود.

کلاس چمنی، سنگریزه‌ای، گلی و سنگلاخی عملکرد بی‌نقص (AUC=۱/۰۰) نشان می‌دهد که مدل به‌طور کامل می‌تواند این سطوح را از دیگران تشخیص دهد. کلاس گلی با AUC=۰/۹۹، اگرچه اندکی پایین‌تر است، اما همچنان نشان‌دهنده عملکرد بسیار قوی است. این ممکن است ناشی از موارد مرزی با کلاس خاکی باشد که در ماتریس درهم‌ریختگی نیز مشاهده شد. کلاس خاکی باوجود چالش‌های مشاهده‌شده در ماتریس درهم‌ریختگی، AUC=۰/۹۹ نشان می‌دهد که مدل به‌طورکلی توانایی خوبی در تشخیص این کلاس دارد. مقادیر AUCنزدیک به ۱ نشان می‌دهند که مدل در تمام آستانه‌های طبقه‌بندی، توانایی بالایی در تفکیک کلاس‌ها دارد. این نتایج حاکی از آن است که مدل ویژگی‌های متمایزکننده هر کلاس را به‌خوبی یاد گرفته است، همچنین توزیع امتیازات طبقه‌بندی برای کلاس‌های مختلف همپوشانی ناچیزی دارد. از طرفی مدل در برابر تغییر آستانه‌های طبقه‌بندی مقاوم است.



شکل ۵- نمودار منحنیROC برای تصاویر دیده نشده بدون توجه به آستانه بندی مدل

Fig 5. ROC curves for unseen images, independent of the model's thresholding settings

منحنی ROC با AUC= ۰/۹۹ -۱/۰۰ توانایی مدل در تمایز بین کلاس‌ها را بدون در نظر گرفتن آستانه طبقه‌بندی می‌سنجد. مقدار بالا نشان می‌دهد مدل می‌تواند نمونه‌های هر کلاس را از دیگر کلاس‌ها جدا کند، حتی اگر در عمل برخی اشتباهات رخ دهد. ماتریس آشفتگی عملکرد مدل را در یک آستانه خاص (۰/۵) نشان می‌دهد. خطاهای مشاهده‌شده (مثل ۱۵ نمونه خاکی که به‌عنوان گلی طبقه‌بندی شدند) مربوط به این آستانه ثابت هستند. AUC بالا نشان می‌دهد که اگر آستانه بهینه شود، می‌توان خطاها را کاهش داد. در ماتریس آشفتگی، ممکن است آستانه بهینه برای کلاس خاکی متفاوت از ۰/۵ باشد. بااین‌حال آستانه ۰/۵ به‌عنوان یک آستانه استاندارد در مطالعات طبقه‌بندی مرسوم است و با کاهش آن دقت مدل پایین خواهد آمد و ممکن است تصاویری که مربوط به یک کلاس نباشد هم در آن کلاس قرار گیرد.

انتخاب معماری مدل، نوع ویژگی‌ها، پارادایم یادگیری و معیارهای ارزیابی می‌تواند تأثیر بسزایی در عملکرد نهایی چنین سیستم‌هایی داشته باشد. برای درک بهتر جایگاه پژوهش حاضر در مقایسه با دیگر مطالعات مطرح در این حوزه، جدول ۱ به مقایسه سیستماتیک چهار مطالعه مشابه (شامل پژوهش حاضر و سه کار مرتبط دیگر) از جنبه‌های مختلف می‌پردازد.

جدول ۱- مقایسه تطبیقی مطالعات مختلف در حوزه طبقه‌بندی زمین برای ناوبری خودران بر اساس معیارهای کلیدی
Table 1. Comparative summary of recent studies on terrain classification for autonomous navigation based on key performance criteria

معیار مقایسه	مطالعه حاضر	Zhang et al. (2016)	Selvathai et al. (2017)	Kim et al. (2007)
هدف اصلی	طبقه‌بندی ٦ کلاس زمین خارج از جاده	طبقه‌بندی دو کلاس زمین	طبقه‌بندی دوکلاس جاده و غیرجاده	مقایسه قطعات و ابرپیکسل‌ها
معماری استفاده شده	ترنسفورم بینایی	جنگل تصادفی	شبکه عصبی پرسپترون چندلایه	طبقه‌بند مبتنی بر قانون کدبوک
تعداد کلاس‌ها	٦ کلاس (خاکی، سنگلاخی، چمنی، گلی، سنگریزه‌ای، سفت)	۲ کلاس (قابل پیمایش، غیرقابل پیمایش)	۲ کلاس (جاده، غیرجاده)	۲ کلاس (قابل پیمایش، غیرقابل پیمایش)
حجم مجموعه داده	۱۷۵۷ تصویر	ترکیب داده لیدار و تصویر (حجم دقیق ذکر نشده)	۱۰۰ تصویر (٦۲۵۰ نمونه در هر کلاس)	۳ دنباله ویدیویی (با ۱۰ فریم آموزشی برای هر کلاس)
ویژگی‌های استخراج شده	ویژگی‌های سطح بالا به صورت خودکار توسط ViT	رنگ، بافت (GLCM)، هندسه	رنگ، کنتراست، بافت (GLCM)، فرکانس فضایی (SFM)	رنگ (RGB, HSV)، هیستوگرام Hue/Saturation
نوع یادگیری	انتقال یادگیری، نظارت شده	خود-نظارتی	نظارت شده	خود-نظارتی (برچسب‌گذاری با استریو)
واحد پردازش تصویر	پچ‌های ١٦×١٦ورودی استاندارد ViT	بلوک‌های تصویری	بلوک‌های ۵۰×۵۰ پیکسل	پچ‌های ثابت در مقابل ابرپیکسل‌ها
دقت کلی (Accuracy)	۹۷٪	۹۱٪/۷۵	۹۳٪	وابسته به اندازه پچ/ابرپیکسل بر اساس منحنی(ROC)

۴- نتیجه‌گیری کلی

در این پژوهش، مدل Vision Transformer (ViT) به‌عنوان یک رویکرد مبتنی بر مکانیسم توجه در طبقه‌بندی زمین‌های خارج از جاده مورد استفاده قرار گرفت. اگرچه مدل به دقت کلی بالای ۹۷٪ روی یک مجموعه داده تست مستقل دست یافت، اما تحلیل‌های ما دو محدودیت اصلی را آشکار کرد:

الف) چالش در تمایز کلاس‌های بصری مشابه: مدل در تمایز بین زمین‌های خاکی و گلی با خطای ۲۵٪ مواجه شد که عمدتاً ناشی از شباهت رنگ و بافت در شرایط نوری مختلف و عدم یادگیری ویژگی‌های سطح بالاتر (مانند نشانه‌های رطوبت یا ساختار سه‌بعدی) بود.

ب) حساسیت به حجم داده: نوسانات در دقت اعتبارسنجی و خطای مشاهده‌شده در کلاس‌های مشکل‌دار، تا حد زیادی به محدودیت حجم داده‌های آموزشی و عمدتاً عدم استفاده از تکنیک‌های افزایش داده نسبت داده می‌شود.

برای رفع این محدودیت‌ها و هموار کردن مسیر تحقیقات آینده، راهکارهای عملی زیر پیشنهاد می‌شود:

- افزایش داده هدفمند: به‌جای افزایش داده عمومی، استفاده از تکنیک‌های هدفمند مانند شبیه‌سازی رطوبت روی خاک خشک یا افزودن نورپردازی مصنوعی در تصاویر برای تشدید تفاوت‌های کلیدی بین کلاس‌های مشابه.

- تغییر معماری برای اســتخراج ویژگی‌های ســطح بالا: ادغام ViT با لایه‌های کانولوشــنی (مانند یک مدل هیبریدی) یا اســتفاده از ماژول‌های توجه پیشرفته‌تر که بتوانند بر روی نواحی بسیار ظریف تصــویر (مانند بافت ســطوح در مقیاس بســیار کوچک) تمرکز کنند.
- تنظیم آســتانه‌های کلاس‌به‌کلاس: به کارگیری اســتراتژی هزینــه-حســاس (Cost-Sensitive Learning) در حین آموزش یا استنتاج، که در آن penalty اشــتباه طبقه‌بندی کردن خاک به عنوان گلی (یــا بــالعکس) افزایش می‌یــابد تــا مــدل در تشخیص این کلاس‌ها محتاط‌تر عمل کند.
- تلفیق داده‌هــای چنــدمُــدی: در صــورت امکــان، جمع‌آوری و استفاده از داده‌های عمق (Depth) یا حرارتی (Thermal) alongside تصــاویر RGB، زیرا این داده‌ها می‌توانند اطلاعات مکملی (مانند بافت ســه‌بعدی یا محتوای رطوبت) ارائه دهند که تمایز بین سطوح مشابه را ممکن می‌سازد.

یافته‌های این مطالعه نه تنها توانایی‌های ViT، بلکه چالش‌های عملی پیاده‌سازی آن در محیط‌های واقعی را نیز برجسته می‌کند. با توجه به محدودیت‌های این تحقیق از طریق راهکارهای پیشنهادی، می‌توان به سیستم‌های ادراکی قوی‌تر و قابل اعتمادتری برای وسایل نقلیه خودران خارج از جاده دست یافت.

منابع

Cai, H., Li, J., Hu, M., Gan, C., & Han, S. (2023). *Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction.* Proceedings of the IEEE/CVF International Conference on Computer Vision, 17302–17313.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* ArXiv Preprint ArXiv:2010.11929.

Faykus III, M. H., Pickeral, A., Marquez, E., Smith, M. C., & Calhoun, J. C. (2024). *Efficient Vision Transformers for Autonomous Off-Road Perception Systems.* Journal of Computer and Communications, 12(9), 188–207.

Holder, C. J., Breckon, T. P., & Wei, X. (2016). *From on-road to off: Transfer learning within a deep convolutional neural network for segmentation and classification of off-road scenes.* Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14, 149–162. https://doi.org/10.1007/978-3-319-46604-0_11

Jin, Y., Han, D., & Ko, H. (2021). *Memory-based semantic segmentation for off-road unstructured natural environments.* 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 24–31. https://doi.org/10.1109/IROS51168.2021.9636620

Kim, D., Oh, S. M., & Rehg, J. M. (2007). *Traversability classification for UGV navigation: A comparison of patch and superpixel representations*. 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, 3166–3173. https://doi.org/10.1109/IROS.2007.4399610

Peng, J., Hua, G., Tian, S., Wu, Y., & Zou, W. (2024). L*ightweight boundary-assisted network for freespace segmentation in unstructured road scenes*. Displays, 83, 102688. https://doi.org/10.1016/j.displa.2024.102688

Qiu, J., & Jiang, C. (2025). *A bilateral semantic guidance network for detection of off-road freespace with impairments based on joint semantic segmentation and edge detection*. Computers and Electrical Engineering, 123, 110045. https://doi.org/10.1016/j.compeleceng.2024.110045

Selvathai, T., Varadhan, J., & Ramesh, S. (2017). *Road and off road terrain classification for autonomous ground vehicle.* 2017 International Conference on Information Communication and Embedded Systems (ICICES), 1–3. https://doi.org/10.1109/ICICES.2017.8070724

Sgibnev, I., Sorokin, A., Vishnyakov, B., & Vizilter, Y. (2020). *Deep semantic segmentation for the off-road autonomous driving.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 617–622. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-617-2020

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need.* Advances in Neural Information Processing Systems, *30*.

Viswanath, K., Singh, K., Jiang, P., Sujit, P. B., & Saripalli, S. (2021). *Offseg: A semantic segmentation framework for off-road driving.* 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), 354–359. https://doi.org/10.1109/CASE49439.2021.9551643

Wu, L., & Liu, T. (2024). *Drivable area detection in off-road for Autonomous Driving*. 2024 IEEE 7th International Conference on Electronic Information and Communication Technology (ICEICT), 515–520. https://doi.org/10.1109/ICEICT61637.2024.10670777