machine learning & knowledge extraction

MDPI

*Article*

# XTrail-ID: An Explainable AI Human Footprint Trail Identification on Soil Substrate Using Unsupervised Machine Learning from UAV Imagery

**Wazha Mmereki [1,*], Rodrigo S. Jamisola, Jr. [1], Zoe C. Jewell [2,3], Tinao Petso [1], Oduetse Matsebe [1] and Sky K. Alibhai [2,3]**

[1] Department of Mechanical, Energy and Industrial Engineering, Botswana International University of Science and Technology (BIUST), Palapye Private Bag 16, Botswana; jamisolar@biust.ac.bw (R.S.S.J.); pt19100041@biust.ac.bw (T.P.); matsebeo@biust.ac.bw (O.M.)
[2] Nicholas School of the Environment, Duke University, Durham, NC 27708, USA; zoejewell228@gmail.com (Z.C.J.); sky.alibhai@gmail.com (S.K.A.)
[3] Wildtrack Inc., Box 90328, Durham, NC 27708, USA
* Correspondence: mw19100038@biust.ac.bw

**Abstract**

This paper investigates human–AI collaboration through explainable AI where we interpret the results of barefoot print clustering using unsupervised machine learning. This can be used to identify the number of individuals from barefoot prints on the ground as a tool in forensics or anti-poaching. A self-supervised vision transformer, DINOv2, is used to automatically extract feature embeddings from localized barefoot-print regions to identify trails belonging to an individual on soil substrate. Furthermore, we introduce an Embedding Spatial Attribution Module (ESAM) to generate spatial attribution heatmaps, enabling visualization of discriminative regions that contribute to individual-specific trail identification and improving model explainability. The proposed method is named XTrail-ID, an explainable human footprint trail identification framework with two variants, OBB-XTrail-ID (oriented bounding box-based), and SEG-XTrail-ID (segmentation-based). We quantify embedding similarity using three complementary metrics: cosine similarity, Pearson correlation coefficient, and Spearman rank correlation. Twenty adults (ten males, ten females) participated, with a total of 1000 trail images extracted from UAV imagery. SEG-XTrail-ID using cosine similarity yielded the highest performance, with (3.21) discriminability and (94.2%) accuracy, while OBB-XTrail-ID using cosine similarity achieved (2.54) discriminability and (91.5%) accuracy. In addition, the latter exhibited reduced consistency in footprint grouping when more than three individuals were present within a single frame.

**Keywords:** human–AI collaboration; explainable AI; barefoot trail; soil substrate; UAV imagery; self-supervised learning; DINOv2 vision transformer; heatmaps; similarity metrics; forensic applications

Check for updates

Academic Editor: Randy Goebel

Received: 5 May 2026
Revised: 12 June 2026
Accepted: 15 June 2026
Published: 18 June 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Human footprint trails remain one of the most informative forms of trace evidence encountered in outdoor environments, providing valuable clues for forensic investigations, military tracking, border surveillance, and search-and-rescue operations, particularly when other biometric cues are unavailable [1,2].

Recent advances in computer vision for barefoot print analysis have focused primarily on single-footprint identification and matching for forensic analysis [3]. However, in practical forensic scenarios, footprint evidence may occur as trails composed of multiple impressions rather than isolated footprints, particularly when recovered in outdoor environments [4].

The goal of this work is to group a trail of barefoot prints belonging to one individual from UAV captured imagery using unsupervised machine learning. The explainable artificial intelligence (XAI) part is archived by spatial attribution heatmaps based on normalized feature vectors of footprint morphology.

Nevertheless, most existing approaches continue to rely on statistical analysis and supervised learning, using labeled data to learn and compare morphological features for associating impressions with a suspect, limiting their applicability to such complex scenarios [5]. Recent methods have explored self-supervised learning frameworks that reduce reliance on annotated data while enabling more interpretable feature representations [6,7]. Despite these advances, the automatic self-supervised grouping of barefoot trail impressions belonging to the same individual on soil substrates remains largely unexplored. And lastly, to the best of our knowledge, there is no publicly available dataset for barefoot trail identification on soil substrates.

Our proposed method utilizes normalized feature embeddings together with similarity-based feature association to group unlabeled impressions into individual-specific trails based on shared visual morphology. This avoids the black-box nature of supervised machine learning methods that raise concerns regarding explainability in forensic applications, where understanding the basis of predictions is essential [8,9]. This work collected data from 20 individuals and a total of 1000 images of trails, with 6 to 8 footprints per trail. In addition, this study promotes interpretability through the Embedding Spatial Attribution Module (ESAM), which leverages similarity analysis of feature embeddings and activation heatmaps to show that the model focuses on discriminative footprint regions. The proposed framework may support forensic reconstruction tasks, including movement pattern interpretation and estimation of the number of individuals present within a scene.

## 2. Related Work

The research in this paper draws on methodologies from diverse domains, namely human barefoot print analysis and self-supervised deep learning. Consequently, the related work is structured into two subsections: one focusing on human barefoot print analysis methods, and the other discussing self-supervised methods for visual feature learning, which, although not previously applied to footprint identification, provide a theoretical foundation for the framework proposed in this study.

### *2.1. Human Barefoot Print Analysis Methods*

This section reviews an array of studies that utilize conventional methodologies, sensor-based approaches, machine learning algorithms, and deep learning architectures for the examination of barefoot prints across diverse domains. Early studies relied primarily on manual landmark and morphometric measurements for barefoot print identification and comparison, including techniques such as the Gunn method [10], the Overlay method [11], the Optical Center method [12], the Robbins technique [13], and the Reel method [14].

Subsequently, research shifted towards digital image processing and statistical similarity metrics; nevertheless, these investigations continued to depend on contour-based features, linear and angular measurements, or predetermined geometric descriptors. For example, Osisanwo et al. [15] examined human footprints for forensic applications by

photographically capturing crime-scene impressions and subjecting them to conventional image-processing algorithms to evaluate their uniqueness. Švábová et al. [16] examined stride length derived from dynamic footprints on controlled walkways to infer biological attributes, including sex and stature. Their methodology used conventional statistical techniques, without machine learning, and omitted any documentation of predictive accuracy.

Domjanic et al. [17] employed geometric morphometric analysis to investigate the morphology of young women's footprints, using landmarks and semi-landmarks to derive Procrustes coordinates and principal components, thus quantifying variation in the shape of the footprint. Liu et al. [18] developed a methodology to quantify the similarity of barefoot prints in forensic applications. The investigation assesses various measurement techniques for comparing two-dimensional footprint images, encompassing contour-based and linear similarity metrics, to determine whether two footprints originate from the same individual. Kennedy et al. [19] focused on single barefoot prints captured on inked paper and analyzed using statistical linear and angular measurements to assess uniqueness and similarity. Sun et al. [20] conducted a reliability study of the Chinese method for linear barefoot print measurement, showing that it yields high test–retest and consistency for both inked and dust footprints, and demonstrating its practical value as a reliable quantitative approach in forensic footprint analysis.

However, these traditional footprint recognition methods are constrained by their reliance on hand-made features, which often lack robustness to variations in walking style and are highly sensitive to changes in gait dynamics and environmental conditions, such as soil substrate variability [21]. In addition, they are typically developed and validated in controlled acquisition settings such as inked prints, paper impressions, or laboratory walkways—and are primarily designed for single-print comparisons rather than trail-based analysis in unconstrained environments. Moreover, the dependence on manual landmark placement and measurement introduces examiner subjectivity and susceptibility to human error, potentially leading to inconsistent and non-reproducible outcomes. In addition to traditional morphometric methods, several studies have explored the use of sensor-based systems to capture dynamic characteristics of barefoot prints and gait, providing quantitative insights that are difficult to obtain through manual observation alone.

For example, in ref. [22], Zhou et al. proposed a person identification approach using pressure footsteps captured through a carpet-integrated fabric sensor system, extracting features such as center-of-gravity shifts and maximum pressure per step and achieving about 76.9% accuracy in classifying individuals based on single steps. Zhang et al. [23] developed a rule-based algorithm to extract and recognize footprints from dynamic plantar pressure images, using segmented pressure distributions and geometric features for identification. Xiang et al. [24] applied supervised machine learning (PCA/RFE with SVM and Naïve Bayes) to classify habitually barefoot and shod populations using controlled plantar pressure and foot shape data, achieving a precision of 93% to 95%. Bao et al. [25] used a CNN based on supervised attention to classify barefoot pressure images acquired from sensor hardware under controlled conditions. Unlike these sensor-based approaches, Our study utilizes self supervised vision-based deep learning on soil footprint images, enabling automated feature learning and trail-level identity matching in natural outdoor environments.

Recent advances in machine learning and deep learning have shifted footprint research towards automated feature extraction and classification frameworks that offer greater stability and robustness. Abuqadumah et al. [26] applied deep transfer learning for human identification using footprint images by fine-tuning various pre-trained CNN architectures such as AlexNet, VGG, GoogLeNet and Inception-V3 on a small labeled dataset, achieving 90%+ recognition accuracies under controlled conditions. Hild et al. [27] developed a person identification method based on the 3D shape of the barefoot sole, using geo-

metric features from 3D scans of the sole surface to achieve robust biometric recognition. Mmereki et al. [28] investigated deep learning approaches based on bounding-boxes, landmarks and segmentation for the automatic classification of human barefoot prints on soil substrates, achieving a maximum precision of 93.5%. Shen et al. [29] developed a footprint detection method by integrating an enhanced StarNet backbone into YOLOv8, achieving the latest barefoot print detection accuracy with mAP = 99.5%, while also reducing model complexity and computational overhead. Jin et al. [30] proposed a personal recognition algorithm that extracts multi-scale features from barefoot print images using a ResNet-based network combined with Horizontal Pyramid Matching and metric learning, achieving a top recognition accuracy of 96.2% on a large barefoot dataset.

Although these approaches achieve high accuracy, their dependence on labeled data, single-footprint analysis, and controlled laboratory or sensor-based acquisition limits their applicability to real-world forensic scenarios, as summarized in Table 1. Direct comparisons of accuracy can therefore be misleading, as previous studies rely on data and conditions that do not reflect the variability of natural soil-based footprints.

**Table 1.** Summary of prior methods, learning approaches, and analytical techniques in barefoot-print research, with comparison to the present study.

| Study | Method | Category & Learning | Barefoot Print Representation | Feature Used | Acquisition Medium | Performance Metrics | Analytical Task |
|---|---|---|---|---|---|---|---|
| [10,12,13] | Manual Landmarking | Traditional morphometric | 2D single footprint | Landmarks/ linear measurements | Footprint image | Measurement reliabilty & morphological agreement | Footprint identification |
| [11,14] | Overlay/ Statistical | Traditional morphometric | 2D single footprint | Overlay comparison/ statistical metrics | Footprint image | Anthropometric measurement consistency | Footprint comparison |
| [15] | Pattern matching | Traditional & Rulebased | 2D single footprint | Contour-based | Crime scene footprint image | Morphological pattern matching | Footprint uniqueness analysis/matching |
| [16] | Gait analysis | Traditional & Statistical | Dynamic footprints | Stride length, step distance | Controlled walkway (paper surface) | Pearson correlation coefficient 0.295 | Biological profiling (sex/stature) |
| [17] | Geometric morphometric | Traditional & Statistical | 2D Single footprint | Landmark coordinates | Foot scan | $p$-value 0.003 | Footprint shape variation analysis |
| [18] | Measurement based | Traditional & Statistical | 2D single footprints | Linear and angular measures | Inked paper footprints | Similarity measurement score | Footprint similarity measurement |
| [19] | Statistical analysis | Traditional & Statistical | 2D single footprint | Linear and angular measures | Inked paper footprints | statistical significance tests | Footprint uniqueness and similarity |
| [20] | Linear measurement | Traditional & Statistical | 2D single footprint | Morphometric linear distances | Ink/dust footprint images | intraclass correlation coefficient 0.999 | Reliability of linear footprint measurements |
| [22] | Pressure mat | Sensor based | Dynamic footsteps | Pressure distribution | Fabric pressure sensor under carpet | Accuracy 76.9% | Person identification from footstep pressure |
| [23] | segmentation | Sensor based | 2D single footprint | Plantar pressure | Flexible force sensor | Accuracy 97% | Footprint extraction and recognition |
| [24] | PCA,SVM Naïve Bayes | Sensor based learning | 2D single footprint | Foot metrics + dynamic plantar pressure | Gait lab pressure plate + foot scans | Accuracy 95% | Population classification barefoot vs shod |
| [25] | Deep CNN Selective fusion | Sensor based deeplearning | 2D single footprint | Regions of high /low pressure | Pressure sensing | Accuracy 93.6% | Footprint/pressure image recognition |
| [26] | CNN | machine learning | 2D single footprint | 2D single footprint | Controlled 2D image capture | Accuracy 98.33% | Human identification from footprint images |
| [27] | PCA,/LDA /SVM | Machine learning | 3D single footprint | Shape descriptors | Laser scanner 3D capture | Accuracy 94% | Person identification |
| [28] | DeepFIT | Supervised deep learning | 2D Single footprint | Geometric and deep features | Soil substrate imagery | Accuracy 96% | Person individual identification |

**Table 1.** *Cont.*

| Study | Method | Category & Learning | Barefoot Print Representation | Feature Used | Acquisition Medium | Performance Metrics | Analytical Task |
|---|---|---|---|---|---|---|---|
| [29] | YOLOv8-StarNet | Supervised deep learning | 2D single footprint | CNN-based spatial features | Inked paper footprint | mAP 99.5% | Footprint detection |
| [30] | ResNet | Supervised deep learning | 2D single footprint | Multi-scale deep features (HPM) | Specialized sensor | Accuracy 96.2% | Personal identification via footprint features |
| This Study | XTrail-ID | Self supervised | 2D trail footprints | Normalized feature vectors | UAV imagery soil substrate | Accuracy 94.2% | Explainable individual trail identification |

Note: CNN, Convolutional Neural Network; DeepFIT, Deep Learning Footprint Identification Technology; LDA, Linear Discriminant Analysis; PCA, Principal Component Analysis; SEG, Segmentation; SVM, Support Vector Machine.

This underscores the need for an unsupervised framework capable of learning discriminative features directly from soil-imprinted footprints, enabling trail-based identity matching across multiple footprints without requiring labels or controlled acquisition conditions. In summary, the key contributions of this study are as follows.

1. We implemented an unsupervised framework, XTrail-ID, to group barefoot prints belonging to the same individual based on normalized morphological feature vectors extracted by a vision transformer from localized footprint regions. These feature representations enabled pairwise cosine similarity analysis for footprint association and grouping.
2. This study promotes explainability through the Embedding Spatial Attribution Module (ESAM), which converts frozen vision transformer patch embeddings into spatial attribution heatmaps showing low and high relevance areas, enabling visualization of the footprint regions that contribute most strongly to embedding similarity and identity grouping.
3. This study provides a comparative evaluation of two footprint localization methods, namely oriented bounding boxes (OBB) and segmentation masks (SEG), and investigates how each representation influences self-supervised feature embedding extraction and barefoot trail grouping performance.
4. This study introduces and evaluates a UAV-acquired trail soil dataset for the assessment of unsupervised barefoot trail grouping, providing a realistic benchmark for associating footprints into individual-specific trails under real world field conditions.

### *2.2. Self-Supervised Methods for Visual Feature Learning*

Recent developments in self-supervised learning and foundation models have shown that effective visual representations can be learned without relying on task-specific labels. Instead of supervised signals, these approaches exploit the intrinsic structure of visual data to learn features that generalize across a wide range of downstream tasks [31]. Reconstruction-based approaches, such as masked image modeling [32], mainly focus on pixel-level recovery, which may not explicitly preserve the discriminative identity-related structure required for fine-grained matching.

Contrastive learning frameworks such as SimCLR [33] and MoCo [34] learn instance-level representations but depend on carefully designed augmentations and large batch sizes and may struggle to preserve subtle morphological variations critical for distinguishing similar footprints. In contrast, DINO [35] and its successor DINOv2 [36] released by Meta in April 2023, based on iBOT [37], represent significant advances in self-supervised representation learning. These models employ a self-distillation framework built upon Vision Transformers (ViTs), enabling the learning of robust and transferable visual features.

Moreover, its framework incorporates large-scale curated datasets and improved training strategies, resulting in more stable optimization and significantly improved performance [38]. ViTs process images as sequences of patches and leverage self-attention mechanisms to capture both fine-grained local details and invariant global image structures, while exhibiting robustness to multiple views and crops and benefiting from large-scale pre-training. DINO-based models learn structured feature embeddings in which visually similar instances cluster together, enabling similarity-based inference via metrics like cosine similarity. They also exhibit emergent behaviors such as unsupervised object segmentation, with attention maps often aligning to meaningful object boundaries without explicit supervision [39].

To the best of our knowledge, self-supervised methods have not yet been applied to barefoot print analysis. Frameworks such as DINOv2 can encode structural information from raw soil footprint images, allowing trail based identity matching and clustering across impressions, and helping to address variability and annotation challenges in forensic settings. Although self-supervised models such as DINOv2 are expensive to pretrain, pretrained frozen encoders can be directly used to obtain rich, invariant visual representations for similarity-based and downstream vision tasks [40].

## 3. Dataset

### *3.1. Data Acquisition Platform and Sensor Configuration*

Aerial imagery was acquired using an Inspired Flight IF1200A (Inspired Flight Systems Inc.), equipped with a Gremsy T3 gimbal (Gremsy, Vietnam) carrying a Sony Alpha 7R IV camera (Sony), with a 35 mm lens. Flights were conducted at 8 m above ground level with the gimbal fixed in the nadir position, producing imagery with a ground sampling distance of approximately 1.49 mm/pixel. Flight planning used 80% forward and 70% side overlap, while image acquisition settings were fixed at f/5.6, ISO 400, and 1/1000 s shutter speed. Georeferencing relied on dual GNSS receivers onboard without ground control points.

### *3.2. Participants and Ethical Approval*

The study involved 20 healthy adult participants (10 males and 10 females) who provided their written informed consent. Ethical approval was obtained from the Human Ethics Research Committee of the Botswana International University of Science and Technology (HREC-12, approved 10 August 2023), with additional permission granted by the Ministry of Environment and Tourism (Permit No. ENT 8/36/4 LVII 7).

### *3.3. Data Collection Protocol and Imaging Conditions*

The participants walked side-by-side and along a single trail path. A total of 1000 images of the trail were extracted from 30-s aerial video recordings. The membership of the ground-truth trail was established during data collection and independently verified by an expert tracker specializing in military tracking within the Botswana Defense Force. To our knowledge, no publicly available benchmark dataset currently exists for barefoot trail identification on soil substrates.

### *3.4. Substrate Conditions and Environmental Characteristics*

Two soft-soil substrate conditions with differing compaction characteristics were considered during data acquisition to incorporate substrate variability. The first condition comprised compact soft soil, which provided firmer ground and improved footprint retention, whereas the second comprised loose soft soil characterized by a finer granular texture and reduced impression definition. To incorporate environmental variability, data collection was conducted in the morning and evening under different lighting conditions

and in some instances footprints were partially occluded by grass. Table 2 presents the substrate characteristics, and environmental conditions observed during data collection.

**Table 2.** Environmental and Substrate Conditions During Data Collection

| Soil Type | Soil Moisture | Relative Compaction | Grain Size | Surface Roughness | Vegetation Coverage | Environmental Disturbance | Footprint Age | Illumination | Occlussion Level |
|---|---|---|---|---|---|---|---|---|---|
| Compact soft soil | Dry | Medium | Fine | Smooth | None | Lightning & shadows | Immediately captured | Morning & evening | No occlusion |
| Loose soft soil | Dry | Lower | Fine | Smooth | Sparse grass | Partially occlusion | Immediately captured | Morning & evening | Partial overlap |

## 4. Methodology

### *4.1. Proposed Framework*

Building on the dataset described above, we propose XTrail-ID, a self-supervised pipeline designed to group barefoot prints based on their normalized feature embeddings captured on natural soil substrates. An overview of the framework is illustrated in Figure 1. Barefoot prints are localized using a detection module that outputs either OBB or SEG. The OBB representation explicitly captures the orientation of the footprint, enabling robust localization under rotational variability. Segmentation provides pixel-level delineation mask, focusing on feature extraction within the footprint region. Despite these differences, both localization strategies produce a unified footprint region $b_k$, defined directly by an OBB or derived from a SEG mask-based region. DINOv2 extracts latent embeddings, and similarity metrics compute pairwise similarity while the proposed ESAM transforms similarity-driven embedding activations into spatially attributable heatmaps for explainable Trail-ID grouping.

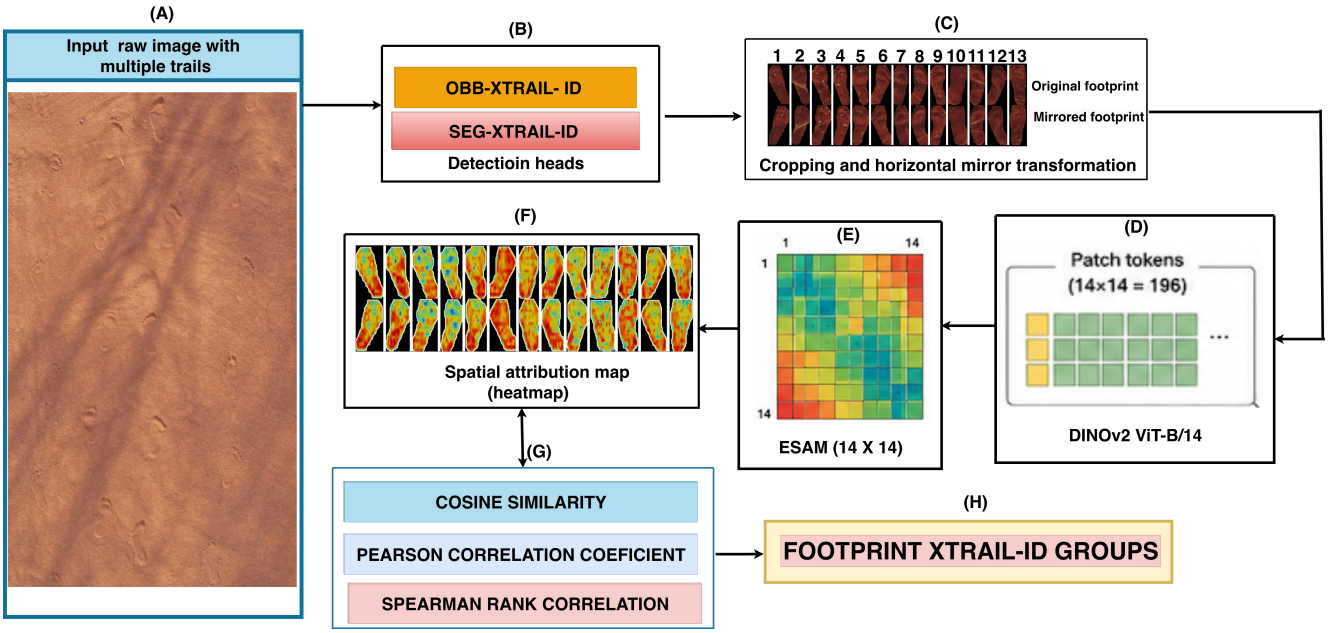



**Figure 1.** Overview of the proposed XTrail-ID framework. Footprint regions are localized using OBB and SEG, followed by feature embedding extraction using the self-supervised DINOv2 ViT-B/14 model (released by Meta in April 2023 and available on GitHub: https://github.com/facebookresearch/dinov2 accessed on 20 February 2026) for robust feature representation, similarity metric computation, and finally ESAM module, which converts embeddings into spatially interpretable heatmaps.

XTrail-ID comprises five main stages: (Section 4.2.1) detection and localization of barefoot prints, (Section 4.2.2) cropping with horizontal mirror normalization, and (Section 4.2.3) extraction of DINOv2 features. In addition, two supporting components are presented: (Section 4.2.4) the use of oriented bounding boxes instead of axis-aligned bounding boxes and (Section 4.2.5) motivation for using DINOv2 and ESAM.

### 4.2. Implementation Details

The XTrail-ID pipeline was implemented in Python 3.10 using PyTorch 2.2 on an NVIDIA RTX 2080 Ti GPU. The feature extraction backbone is a frozen DINOv2 vision transformer model DINOv2 (ViT-B/14 variant) with an input resolution of 518 × 518 pixels and a patch size of 14 × 14 pixels, producing 768-dimensional embeddings. Table 3 presents the operational workflow of the X-Trail-ID pipeline.

**Table 3.** Operational workflow of the XTrail-ID system for independent implementation.

| Stage | Input | Operation | Output | Parameters |
|---|---|---|---|---|
| Detection | UAV soil imagery with trails | Barefoot print localization model | OBB & SEG | Confidence threshold = 0.7 |
| Cropping | OBB & SEG | Crop each detected footprint ROI | Rotated rectangle & binary mask | Zero-valued background padding |
| Orientation normalization | Cropped patch | Horizontal flip (left/right standardisation) | Aligned footprint | Geometric heuristic |
| Resizing | Aligned cropped image | Resize to fixed input size | Standardized image | 518 × 518 |
| Feature encoding | 518 × 518 image | Feature extraction using frozen transformer | 768-dimensional embedding vector | DINOv2 ViT-B/14 |
| Normalization | Embedding vector | unit vector normalization | L2-normalized unit vector (768-D) | L2 norm scaling |
| Similarity computation | L2-normalized 768-D embeddings | Pairwise similarity computation | Similarity score/trail-id | Cosine similarity |
| ESAM | DINOv2 patch tokens + global embedding | Token attribution mapping (patch-to-spatial) | 2D spatial attribution heatmap | Token relevance (patch-level) |

#### 4.2.1. Barefoot Print Detection and Localization

In this work, the location of the barefoot prints is performed using an adapted version of DeepFIT [28], a deep learning footprint identification technology previously validated for identifying footprints on soft and sandy soil substrates. Although the original DeepFIT model was developed for single footprint images, transfer learning was used to extend its applicability to trail image scenes containing multiple footprint instances within a single image. Specifically, the pretrained DeepFIT weights were used to initialize the network and subsequently fine-tuned on a small trail dataset of compact and loose soil substrates using a learning rate of 0.001, batch size of 16, momentum of 0.99, weight decay of 0.0005, and 300 training epochs.

This adaptation enabled the model to retain the generalized morphological representations of the footprint learned from the original data set while accommodating the increased complexity of the new scene, background variability, and footprint arrangements characteristic of the trail imagery. Following standard transfer learning practices for adapting pretrained object detectors to new domains [41], the network uses its multi-head detection architecture, including the extra small detection head (XSDH), to localize multiple barefoot footprint instances within a single image. Importantly, DeepFIT is used solely as a footprint location module, generating accurate footprint regions that are subsequently used for unsupervised feature extraction and footprint matching, which constitute the primary focus of this study.

4.2.2. Footprint Cropping and Horizontal Mirror Transformation

We normalized each footprint by cropping and mirroring, ensuring that feature extraction focuses on the footprint itself rather than the background soil textures. Mirroring aligns the left and right footprints of the same individual, resulting in comparable feature embeddings, consistent with the established behavior of deep networks in producing mirror-symmetric representations [42,43]. To illustrate this formulation, let $\{b_k\}_{k=1}^{M}$ denote the detected footprint regions, each represented as an oriented bounding box (OBB) or derived from a segmentation mask (SEG). The corresponding cropped footprint instances $\{I_k\}_{k=1}^{M}$ are obtained as

$$I_k = \mathrm{Crop}(I_{\mathrm{trail}}, b_k), \quad k = 1, \ldots, M. \tag{1}$$

where $I_{\mathrm{trail}}$ is the original footprint trail image, $b_k$ specifies the region corresponding to the $k$-th footprint, and $I_k$ corresponds to the cropped image of the $k$-th individual footprint. The mirroring operation can be expressed as

$$\mathcal{I}_k = \{I_k, \mathrm{Flip}(I_k)\}, \quad k = 1, \ldots, M, \tag{2}$$

where $\mathrm{Flip}(\cdot)$ denotes horizontal reflection. Let $f(\cdot)$ represent the feature extraction function that maps an input image to a $d$-dimensional embedding space, i.e.,

$$\mathbf{v}_k = f(I_k), \quad f : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^d. \tag{3}$$

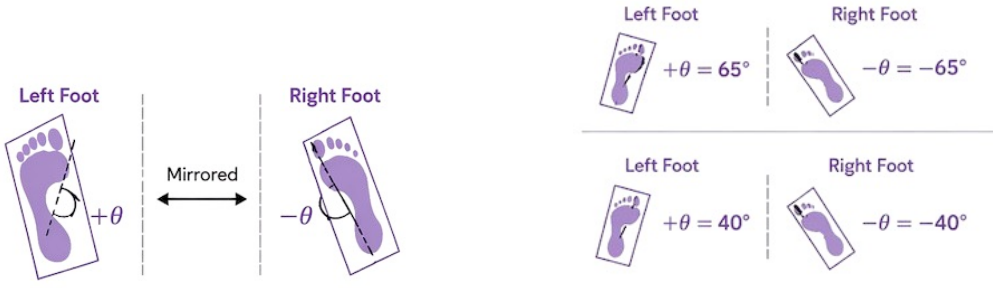
For each pair mirrored in $\mathcal{I}_k$, the learned representation exhibits an approximate invariance to horizontal reflection, which can be expressed as

$$f(I_k) \approx f(\mathrm{Flip}(I_k)). \tag{4}$$

This indicates that both the original and mirrored images produce closely aligned feature embeddings in the representation space. Importantly, horizontal mirroring does not correspond to an inversion of the feature direction, i.e.,

$$f(\mathrm{Flip}(I_k)) \neq -f(I_k), \tag{5}$$

but rather preserves semantic content, resulting in embeddings with similar orientation. This behavior reflects the ability of the feature extractor to capture semantic equivalence while remaining invariant to geometric transformations, and enables a more robust representation of gait patterns and improves the consistency of footprint association. Figure 2a shows the mirroring effect, while Figure 2b shows that the footprints are correctly identified as similar after normalization.




(**a**) Mirrored (**b**) Normalization

**Figure 2.** Images of human barefoot prints showing (**a**) mirrored orientations of left and right footprints, and (**b**) their alignment after applying orientation normalization.

#### 4.2.3. Self-Supervised Feature Extraction Using DINOv2

Given a cropped footprint image $I_k$, a pre-trained DINOv2 vision transformer is used to extract patch-level representations from the final transformer layer, producing a set of spatial feature embeddings $\mathbf{f}_k \in \mathbb{R}^d$:

$$\mathbf{F}_k = \text{DINOv2}(I_k) = \{\mathbf{f}_{k,1}, \ldots, \mathbf{f}_{k,N}\}, \tag{6}$$

where

- $I_k$ denotes the cropped image of the $k$-th individual footprint,
- $\mathbf{F}_k$ represents the set of feature embeddings extracted from $I_k$,
- $N$ is the number of image patches,
- $\mathbf{f}_{k,i} \in \mathbb{R}^d$ denotes the embedding vector of the $i$-th patch.

Each footprint crop produces its own heat-map or patch-level feature set. These patch embeddings are reshaped into a spatial feature map, yielding a dense heat-map representation that highlights discriminative footprint regions. To obtain a compact footprint-level descriptor, the spatial features are aggregated via average pooling:

$$\mathbf{v}_k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{f}_{k,i} \tag{7}$$

Each embedding vector is $\ell_2$-normalized to form a compact latent representation of the visual and morphological characteristics of each localized barefoot print. Normalization enforces scale invariance, allowing clustering to rely on directional similarity in the feature space rather than embedding magnitude. The representation implicitly captures the footprint impression's global shape, local structure, and texture patterns without using explicit outlines or predefined landmarks. This normalization is performed as follows:

$$\hat{\mathbf{v}}_k = \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|_2}. \tag{8}$$

#### 4.2.4. Assessing Axis-Aligned and Oriented Bounding Boxes for Footprint Localization and Orientation Representation

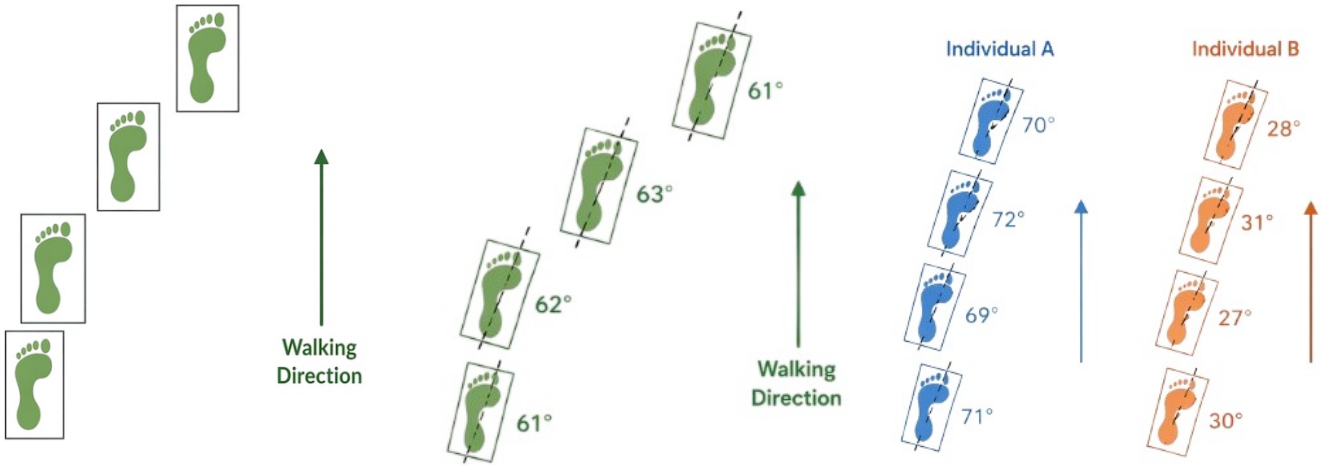
Footprint orientation provides critical information for individual identification. As illustrated in Figure 3a, conventional axis-aligned bounding boxes are limited in that they do not encode rotational information, making it difficult to distinguish between footprints with similar shapes but different orientations.

To address this limitation, we investigate oriented bounding boxes (OBB), which explicitly capture the angular orientation of each footprint. Therefore, as illustrated in Figure 3b, when using OBB, the footprints of the same individual typically maintain consistent orientation angles, while Figure 3c shows that different individuals exhibit different orientation angles.

#### 4.2.5. Motivation for Using DINOv2 Vision Transformer and ESAM for Explainability

DINOv2 was selected as the feature backbone due to its strong out-of-the-box representation capability under a frozen self-supervised setting. Previous studies have shown that frozen DINOv2 representations can be used effectively without fine-tuning for downstream tasks [44–46]. The pretrained embeddings preserve fine-grained structural and spatial information without requiring supervised adaptation or task-specific heads, making them suitable for similarity-based forensic footprint analysis in scenarios where labeled data are unavailable. These properties are particularly important for barefoot footprint analysis,

where identity cues are subtle, shape-driven, and distributed across multiple impressions affected by soil deformation.




(**a**) BBox, axis aligned (**b**) OBB, same individual (**c**) OBB, different individuals

**Figure 3.** Images of human barefoot prints showing (**a**) axis aligned BBox does not encode angularity, (**b**) OBB indicates that same individual maintain almost similar orientation angle, and (**c**) different individuals exhibits distinct orientation angles.

The final transformer block outputs a compact latent embedding of 768-D unit vector, where D denotes the dimensionality of the characteristic. These embeddings encode fine-grained soil-impression texture, geometric footprint contours, and overall morphological structure. Subsequently, similarity metrics are computed between normalized embeddings generated by the projection head to measure relationships between footprint impressions for Trail-ID grouping.

To improve interpretability, we introduce the Embedding Spatial Attribution Module (ESAM), which constitutes the primary methodological contribution of this study. ESAM transforms frozen DINOv2 feature embeddings into spatial attribution heatmaps by computing token-level relevance between local patch embeddings and the global footprint representation. The resulting patch tokens are reshaped onto the original DINOv2 spatial grid, enabling spatial interpretation of discriminative footprint regions through heatmaps that highlight the areas contributing most strongly to matching footprints belonging to the same individual.

Figure 4 illustrates the proposed Vision Transformer pipeline consisting of: (1) trail image crop processing, (2) patch embedding, (3) frozen DINOv2 feature extraction, (4) projection head, (5) Embedding Attribution Module (ESAM), (6) attribution heatmap, (7) similarity-based Trail-ID grouping, and finally (8) forensic interpretation and explainability.

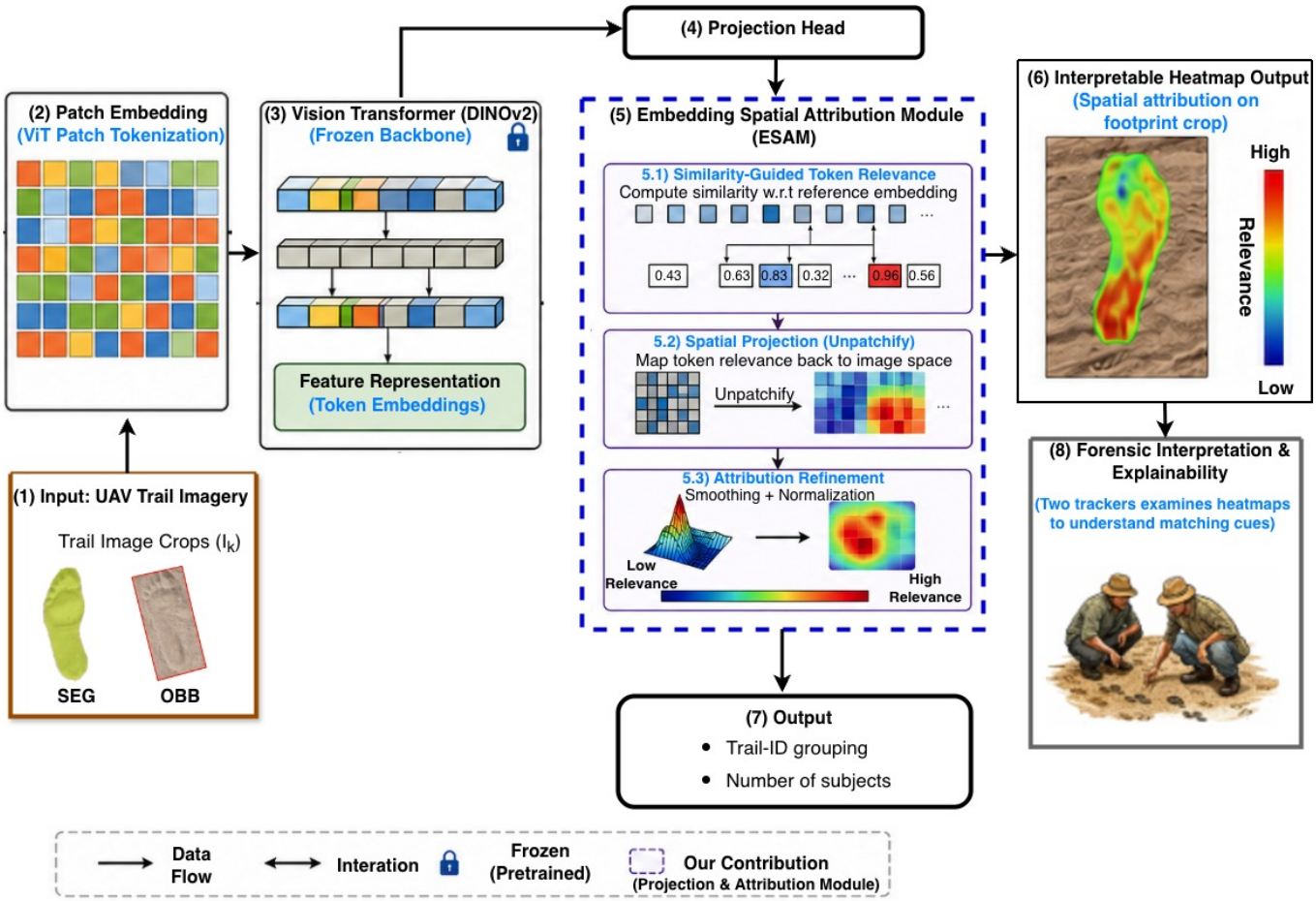



**Figure 4.** ViT-based DINOv2 workflow for barefoot trail analysis. Cropped footprints obtained using OBB and SEG are patch-encoded and transformed into feature embeddings. These embeddings are processed through ESAM to produce spatially grounded activation heatmaps highlighting discriminative footprint regions. The resulting heatmaps are qualitatively evaluated by expert human trackers, who verify whether the high relevance regions correspond to meaningful footprint structures relevant for trail pattern interpretation and subject-level consistency.

## 5. Performance Evaluation

### *5.1. Similarity Metrics: Comparative Evaluation*

In this work, we investigate three distinct approaches for quantifying similarity between footprint features extracted by DINOv2.

To evaluate the choice of similarity measure, we compare cosine similarity, Spearman correlation, and Pearson correlation coefficient as ablation baseline. While cosine similarity operates directly on L2-normalized embeddings and captures angular similarity in feature space, Pearson correlation introduces an additional mean-centering step, while Spearman correlation assesses the monotonic relationship between ranked feature values, making it robust to non-linear but monotonic patterns.

The inclusion of Pearson correlation allows us to explicitly assess the effect of mean-centering on similarity measurement. In particular, this comparison helps determine whether footprint grouping benefits more from preserving the global structure of embedding vectors or from emphasizing relative feature variations independent of their absolute magnitude. This provides a principled basis for selecting an appropriate similarity metric for the task. All three metrics are in the range $[-1, 1]$, where 1 indicates perfect alignment, 0 indicates no linear relationship (orthogonality), and $-1$ indicates perfect opposition.

5.1.1. Similarity Metric 1: Cosine Similarity

Cosine similarity measures the angular alignment between feature vectors in the embedding space. Given two feature vectors $\mathbf{v}_k$ and $\mathbf{v}_l$, we first L2-normalize them:

$$\hat{\mathbf{v}}_k = \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|_2}, \quad \hat{\mathbf{v}}_l = \frac{\mathbf{v}_l}{\|\mathbf{v}_l\|_2}.$$

The cosine similarity is then computed as the dot product of the normalized vectors:

$$\text{CosSim}(\mathbf{v}_k, \mathbf{v}_l) = \hat{\mathbf{v}}_k \cdot \hat{\mathbf{v}}_l \tag{9}$$

5.1.2. Similarity Metric 2: Pearson Correlation Coefficient

Pearson correlation measures the linear relationship between two feature vectors after mean-centering. For feature vectors $\mathbf{v}_k, \mathbf{v}_l \in \mathbb{R}^d$, it is defined as

$$\rho(\mathbf{v}_k, \mathbf{v}_l) = \frac{\sum_{i=1}^{d}(v_{k,i} - \bar{v}_k)(v_{l,i} - \bar{v}_l)}{\sqrt{\sum_{i=1}^{d}(v_{k,i} - \bar{v}_k)^2}\sqrt{\sum_{i=1}^{d}(v_{l,i} - \bar{v}_l)^2}} \tag{10}$$

where

$$\bar{v}_k = \frac{1}{d}\sum_{i=1}^{d} v_{k,i}, \quad \bar{v}_l = \frac{1}{d}\sum_{i=1}^{d} v_{l,i}.$$

denote the mean values of the feature vectors $\mathbf{v}_k$ and $\mathbf{v}_l$, respectively.

5.1.3. Similarity Metric 3: Spearman Rank Correlation

The Spearman rank correlation assesses monotonic relationships between barefoot feature vectors, providing robustness to non-linear or curved variations in DINOv2 embeddings. For feature vectors $\mathbf{v}_k, \mathbf{v}_l \in \mathbb{R}^d$, define:

$$r_{k,i} : \text{ rank of } v_{k,i} \text{ within } \mathbf{v}_k, \quad r_{l,i} : \text{ rank of } v_{l,i} \text{ within } \mathbf{v}_l.$$

The Spearman rank correlation is defined as:

$$\rho_s(\mathbf{v}_k, \mathbf{v}_l) = \frac{\sum_{i=1}^{d}(r_{k,i} - \bar{r}_k)(r_{l,i} - \bar{r}_l)}{\sqrt{\sum_{i=1}^{d}(r_{k,i} - \bar{r}_k)^2}\sqrt{\sum_{i=1}^{d}(r_{l,i} - \bar{r}_l)^2}} \tag{11}$$

where

$$\bar{r}_k = \frac{1}{d}\sum_{i=1}^{d} r_{k,i}, \quad \bar{r}_l = \frac{1}{d}\sum_{i=1}^{d} r_{l,i}.$$

5.1.4. Grouping Evaluation Protocol

For each group of footprints identified through our proposed Traid-ID framework, we compute similarity metrics.

To report predictions as percentage scores on images, similarity values in $[-1, 1]$ are rescaled to the $[0, 1]$ range using

- Similarity rescaling:

$$\text{Rescaled similarity (\%)} = \left(\frac{\text{Similarity} + 1}{2}\right) \times 100 \tag{12}$$

and for a group $G$ containing $n \geq 2$ barefoot prints, we calculate:

- intragroup cosine similarity:

$$\text{CosSim}(G) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{CosSim}(\mathbf{v}_i, \mathbf{v}_j) \tag{13}$$

- Intragroup Pearson Correlation:

$$r_{\text{intra}}(G) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} r(\mathbf{v}_i, \mathbf{v}_j) \tag{14}$$

- intragroup Spearman Correlation:

$$\rho_{\text{intra}}(G) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \rho(\mathbf{v}_i, \mathbf{v}_j) \tag{15}$$

- Group Ranking is defined as follows:
  High consistency: No mixed groups.
  Moderate consistency: A few mixed groups.
  Low consistency: Many mixed groups.

- Discriminability Index:

$$D = \frac{\mu_{\text{inter}} - \mu_{\text{intra}}}{\sqrt{\frac{1}{2}\left(\sigma_{\text{inter}}^2 + \sigma_{\text{intra}}^2\right)}} \tag{16}$$

  where $\mu_{\text{inter}}$ and $\sigma_{\text{inter}}$ denote the mean and standard deviation of inter-class similarity scores, and $\mu_{\text{intra}}$ and $\sigma_{\text{intra}}$ denote the mean and standard deviation of intra-class similarity scores. $D$ is a unitless discriminability index, where higher values indicate improved separability between classes.

## 6. Results and Analysis

We conduct a systematic comparison of cosine similarity, Pearson correlation, and Spearman correlation for footprint grouping on the proposed XTrail-ID methods. Table 4 presents the quantitative comparison across the dataset.

**Table 4.** Comparison of Similarity Metrics for Footprint Identification Across Localization Methods.

| Localization | Metric | n | Intra-Group | Inter-Group | Discriminability | Accuracy |
|---|---|---|---|---|---|---|
| OBB-XTrail-ID | **Cosine** | 40 | $\mathbf{0.80 \pm 0.04}$ | $\mathbf{0.35 \pm 0.07}$ | **2.54** | **91.5%** |
| | Pearson | 40 | $0.70 \pm 0.03$ | $0.32 \pm 0.08$ | 2.06 | 87.3% |
| | Spearman | 40 | $0.64 \pm 0.03$ | $0.33 \pm 0.07$ | 1.82 | 85.1% |
| SEG-XTrail-ID | **Cosine** | 40 | $\mathbf{0.91 \pm 0.05}$ | $\mathbf{0.36 \pm 0.06}$ | **3.21** | **94.2%** |
| | Pearson | 40 | $0.86 \pm 0.03$ | $0.33 \pm 0.07$ | 2.92 | 90.8% |
| | Spearman | 40 | $0.85 \pm 0.04$ | $0.34 \pm 0.08$ | 2.53 | 89.9% |

Note: Cosine similarity consistently outperforms Pearson and Spearman correlations across both localization methods. Bold values indicate the best-performing metric.

A total of 40 images were used for evaluation, with each image providing paired similarity measurements across all three metrics (Cosine, Pearson, and Spearman). This sample size is sufficient for paired statistical testing, as it exceeds commonly recommended minimum thresholds ($n \geq 30$) and supports reliable estimation of mean differences and

significance levels [47]. The paired design further improves statistical power by controlling for inter-image variability.

To ensure evaluation integrity, the 40-image subset was selected from temporally separated UAV frames. Frames were sampled such that no two images originated from the same continuous walking sequence, ensuring temporal independence. In addition, participant-level separation was enforced to prevent any overlap between evaluation samples and data used in model development, thereby avoiding subject-level data leakage.

Figure 5 presents a comparison of intra-group similarity scores and overall accuracy across the three metrics for both localization methods. Cosine similarity achieves the highest intra-group similarity for OBB-XTrail ID ($0.80 \pm 0.04$) and the highest accuracy at 91.5%, followed by Pearson ($0.70 \pm 0.03$) and Spearman ($0.64 \pm 0.03$). Similarly, for SEG-XTrail ID, cosine similarity again outperforms both Pearson ($0.86 \pm 0.03$) and Spearman ($0.85 \pm 0.04$) with a score of $0.91 \pm 0.05$ and the highest overall accuracy of 94.2%. The superior intra-group similarity of cosine similarity indicates that footprints from the same individual exhibit stronger directional alignment in their DINOv2 feature representations than they do linear or rank-based correlations.

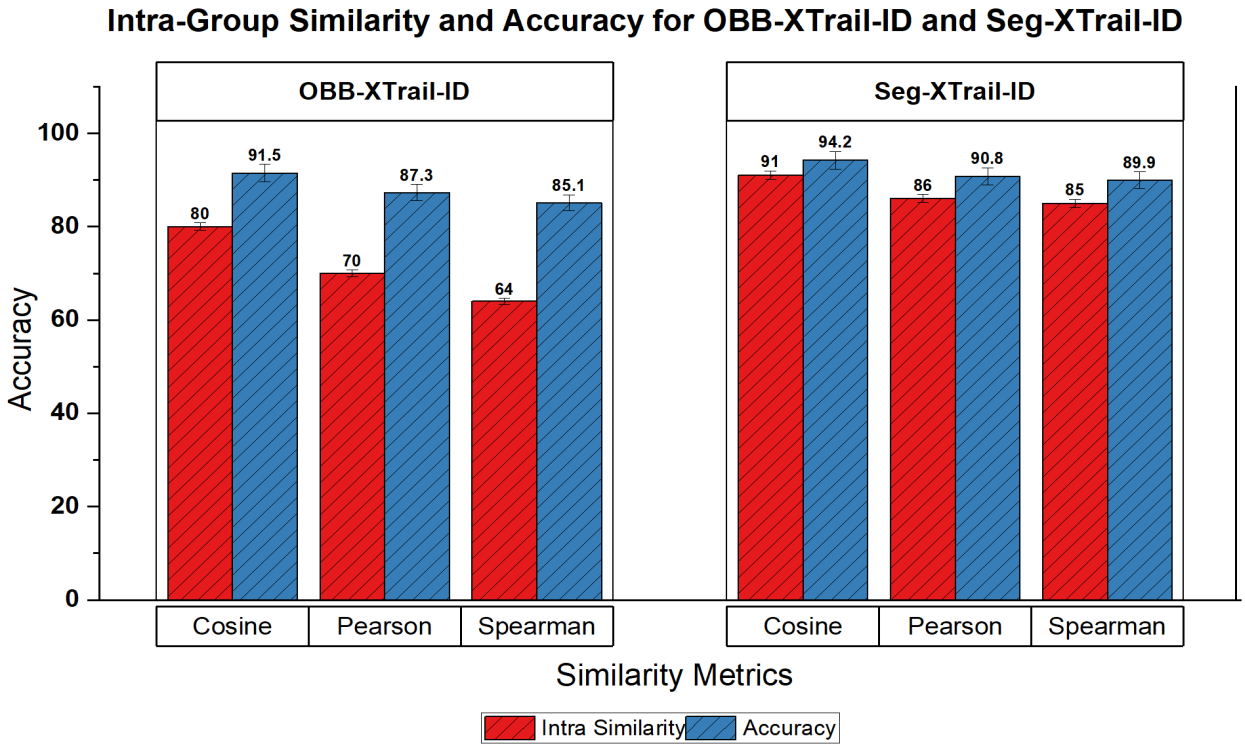



**Figure 5.** Performance Evaluation of XTrail-ID Models for Correlating Barefoot Prints on Soil Substrates.

### *6.1. Statistical Analysis*

We conduct paired $t$-tests to verify the statistical significance of the performance differences between the three metrics and presents the results in Table 5.

The difference between Cosine and Pearson is statistically significant ($p < 0.0167$, paired $t$-test), as is the difference between Cosine and Spearman ($p < 0.0167$). All significant results remain significant after Bonferroni correction ($\alpha = 0.0167$). However, Pearson and Spearman are not statistically significant ($p > 0.05$) for either localization. The statistically significant differences confirm that cosine similarity provides a meaningful advantage over both Pearson and Spearman correlations for footprint grouping with DINOv2 features.

**Table 5.** Paired *t*-test comparisons between similarity metrics.

| Localization | Comparison | t-Value | *p*-Value | Significant |
|---|---|---|---|---|
| OBB-XTrail-ID | Cosine vs. Pearson | 4.52 | 0.0001 | Yes |
| | Cosine vs. Spearman | 5.83 | 0.00001 | Yes |
| | Pearson vs. Spearman | 1.52 | 0.14 | No |
| SEG-XTrail-ID | Cosine vs. Pearson | 4.27 | 0.0001 | Yes |
| | Cosine vs. Spearman | 5.12 | 0.00001 | Yes |
| | Pearson vs. Spearman | 0.89 | 0.38 | No |

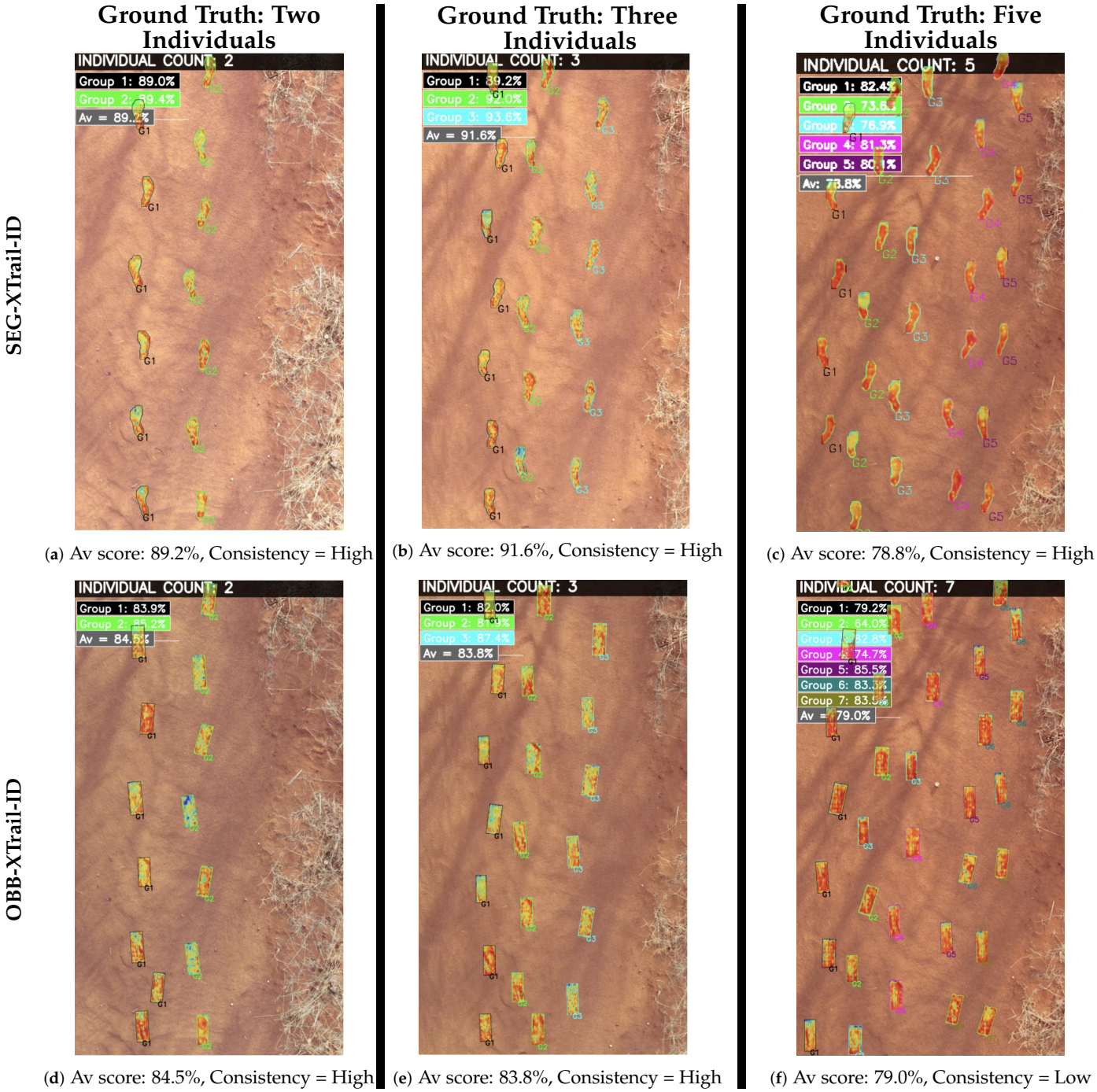
Note: Paired *t*-tests were performed on 40 identical image pairs. Bonferroni-corrected significance threshold: $\alpha = 0.05/3 = 0.0167$.

### 6.2. Grouping Accuracy

Using cosine similarity as the similarity metric leads to the highest grouping accuracy (OBB: 91.5%; SEG: 94.2%) compared to Pearson (OBB: 87.3%; Seg: 90.8%) and Spearman (OBB: 85.1%; SEG: 89.9%. These results validate our choice of cosine similarity for the final footprint grouping system.

### 6.3. Qualitative Results

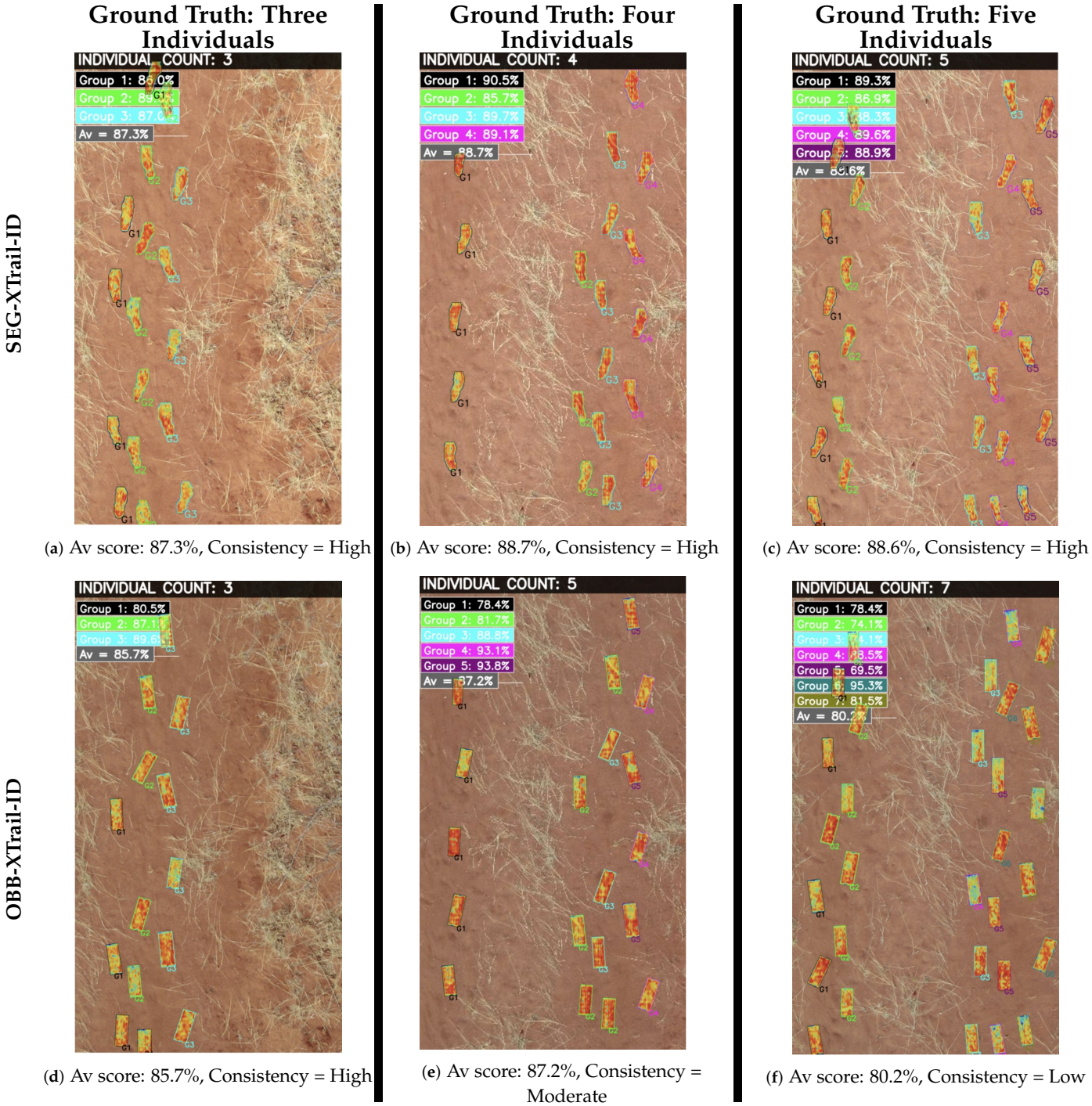
Figures 6–8 demonstrate the effectiveness of our SEG-XTrail-ID method. Based on the quantitative comparison in Table 4, we selected cosine similarity as the optimal similarity metric, as it consistently outperformed both Pearson correlation and Spearman correlation across all evaluation criteria (intra-group similarity, discriminability, and grouping accuracy) During qualitative evaluation, the automatically generated footprint groups were compared against independently verified ground-truth trail information to assess whether the predicted groupings corresponded to the known number of individuals present in each scene. In addition, two experienced trackers independently examined the grouped barefoot trails and interpreted the grouping outcomes based on footprint morphology, trail continuity, and spatial heatmaps. Their assessments were used to qualitatively evaluate whether the footprints assigned to the same group exhibited consistent morphological characteristics and whether the resulting trail reconstructions were consistent with expert tracking interpretations. The trackers also provided explanations for correctly grouped and misgrouped footprints, offering domain specific insights into the strengths and limitations of the proposed framework under field conditions.




(**a**) Av score: 89.2%, Consistency = High (**b**) Av score: 91.6%, Consistency = High (**c**) Av score: 78.8%, Consistency = High

(**d**) Av score: 84.5%, Consistency = High (**e**) Av score: 83.8%, Consistency = High (**f**) Av score: 79.0%, Consistency = Low

**Figure 6.** Qualitative results of barefoot print clustering on clear soft soil. (**a**–**c**) SEG Trail-ID: high consistency in grouping footprints. (**d**,**e**) OBB Trail-ID: effective grouping for two and three individuals. (**f**) OBB Trail-ID: low consistency as number of individuals increases to five.




(**a**) Av score: 87.3%, Consistency = High

(**b**) Av score: 88.7%, Consistency = High

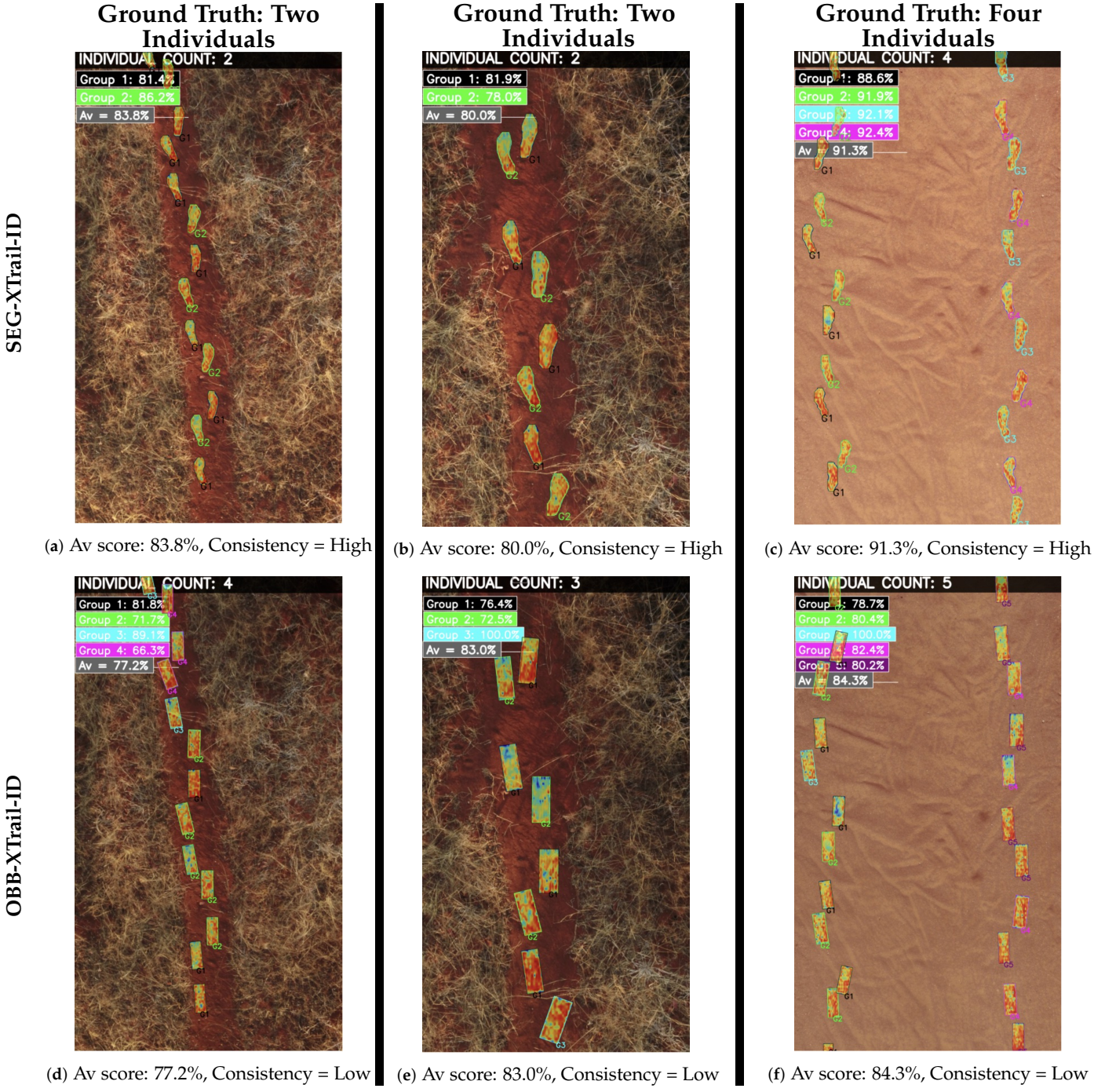
(**c**) Av score: 88.6%, Consistency = High

(**d**) Av score: 85.7%, Consistency = High

(**e**) Av score: 87.2%, Consistency = Moderate

(**f**) Av score: 80.2%, Consistency = Low

**Figure 7.** Qualitative results of barefoot print clustering on partially occluded environment. (**a**–**c**) SEG Trail-ID: high consistency in grouping footprints. (**d**) OBB Trail-ID: effective for three individuals. (**e**,**f**) OBB Trail-ID: low consistency and wrong clustering for four and five individuals.




(**a**) Av score: 83.8%, Consistency = High (**b**) Av score: 80.0%, Consistency = High (**c**) Av score: 91.3%, Consistency = High

(**d**) Av score: 77.2%, Consistency = Low (**e**) Av score: 83.0%, Consistency = Low (**f**) Av score: 84.3%, Consistency = Low

**Figure 8.** Qualitative results of barefoot print clustering on a singular path walking. (**a**–**c**) SEG Trail-ID: high consistency in grouping footprints. (**e**,**f**) OBB Trail-ID: low consistency and wrong clustering in all scenarios.

## 7. Interpretation and Explainability of Spatial Attribution Heatmaps on OBB-XTrail-ID and SEG-XTrail-ID

### *7.1. Explaination of SEG-XTrail-ID Method on Parallel Walking Dataset on Clear Soft Soil*

As illustrated in Figure 6a, the ground truth indicates the presence of two individuals. The spatial attribution heatmaps for G1 are highly concentrated along the principal axis of the barefoot print, with a cosine similarity of 89%, indicating a stable and consistent gait pattern. In contrast, G2 exhibits more spatially dispersed activation heatmaps, but these remain consistent between barefoot prints, as evidenced by a cosine similarity of 89.4%. Moreover, the model precisely estimates the number of individuals, with predictions

aligned with the ground truth. As shown in Figure 6b, the ground truth reveals the presence of three individuals. The heatmaps for G1 exhibit a bias toward the left and right regions of the barefoot prints, achieving a cosine similarity of 89.2%, potentially indicative of a gait that preferentially engages the regions of the lateral feet during locomotion.

For G2, the heatmaps reveal concentrated hotspots in the heel and along the lateral foot regions, with 92% similarity, suggesting pronounced consistency in initial heel contact succeeded by weight transfer along the outer foot. In contrast to G1 and G2, the heatmaps for G3 display prominent warm colors (yellow, green, and light red) extending into the mid-foot region at 93.6%, indicating a more distributed pressure pattern. Furthermore, the model accurately predicts the number of individuals with estimates that are consistent with the ground truth. As shown in Figure 6c, the ground truth reveals the presence of five individuals. The heatmaps for G1 exhibit a relatively uniform red distribution from heel to forefoot, with a cosine similarity of 82.4%, denoting enhanced consistency and a balanced gait pattern.

For G2, the heatmaps reveal a pronounced hollow center in the left barefoot prints and a partially uniform distribution on the right, corresponding to a moderate similarity score of 73.6%, which is lower than that of G1. G3 displays a mildly uniform activation pattern from heel to toe, with a cosine similarity of 76.9%, suggesting a relatively neutral walking pattern. In G4, the heatmaps are more pronounced from heel to midfoot, with a similarity of 81.3%, indicating a gait pattern that emphasizes initial heel strike rather than a fully uniform heel-to-toe transition. G5 demonstrates heel-dominant activation patterns, with a cosine similarity of 78.8%, implying that this individual predominantly charges the heel during walking, with comparatively less activation in the midfoot and forefoot regions. Furthermore, the model accurately predicts the number of individuals present, with the estimated count matching the ground truth.

### *7.2. Explaination of OBB-XTrail-ID Method on Parallel Walking Dataset on Clear Soft Soil*

As demonstrated in Figure 6d, the OBB in G1 indicates that the red zones are isolated and concentrated, with a very deep strike in the heel and a distinct strike in the forefoot. The cosine similarity of 89% indicates a high degree of consistency in the features of the footprint in all steps. This configuration suggests that the individual has a high arch or employs a relatively light midfoot strike. In G2, the heatmaps exhibit greater continuity and elongation, indicative of a flatter foot strike. The cosine similarity of 91.6% further underscores the high consistency of this footprint pattern across steps.

In Figure 6e, G1 exhibits solid activation heatmaps extending from the heel to the toes, signifying a firm and consistent gait. This consistency is further supported by a high cosine similarity of 82% between activation patterns in successive steps. G2, by contrast, exhibits fragmented heatmaps around the central region with noticeable shifts toward the lateral edges, denoting irregular activation patterns, and a reduced cosine similarity of 81% compared to G1. G3 demonstrates the most pronounced blue regions in the center, accompanied by a cosine similarity of 87.3%. This profile likely reflects a gap in the arch region of the barefoot print as a prominent characteristic. Figure 6f reveals that the ground truth comprises footprints from five individuals. Nevertheless, the heatmaps exhibit a reduced separability of characteristics, as the model delineates seven distinct groups (G1–G7). This implies that several groups display fragmented and overlapping activation patterns, thereby impeding the model's capacity to reliably associate footprints from the same individual. In particular, the model accurately clusters footprints in G1, where activation patterns are more compact and consistent between successive steps. Consequently, the model achieves an average cosine similarity of 79.0%, indicative of

decreased consistency in the correlating footprints of the same individual amid increased grouping complexity.

### 7.3. Explanation of SEG-XTrail-ID Method on Parallel Walking Dataset on Partially Occluded Environment

Figure 7a indicates that the ground truth encompasses three individuals, which the model accurately predicts through three different groups (G1–G3). G1 exhibits highly compact and localized activation patterns, predominantly concentrated in the heel and midfoot regions. G2 exhibits moderately distributed activations that extend from the heel toward the forefoot, with slightly greater spread than in G1. G3, meanwhile, displays heatmaps that are more distinguishable from G1 and G2, with increased attention to the heel. In general, the model achieves a high average cosine similarity of 87.3%, demonstrating robust consistency in the association of footprints of the same individual. Figure 7b ground truth data shows four individuals present. The model accurately linked all barefoot impressions. G1 displays more concentrated and localized activations, suggesting heavier footfalls; G2 reveals moderately stretched activations; G3 exhibits robust, continuous heat along the footprint's axis, denoting a clearly defined shape. G4 shows the most vivid and extended activation, indicative of stronger or deeper imprints. The cosine similarity reaches 88.7%, reflecting strong consistency between the footprints.

Figure 7c ground truth data reveal five individuals (G1–G5). G1 produces high-intensity heatmaps from heel to toes, implying a flatter foot; G2 displays strong heatmaps on the heel and toes; G3 shows lower-intensity heatmaps along the foot's axis. G4 exhibits a uniform distribution, suggesting full barefoot contact with the ground; The G5 activation maps are intense on the heel, sides, and toes but weaker centrally. The cosine similarity is 86.6%, which signals high consistency.

### 7.4. Explanation of OBB-XTrail-ID Method on Parallel Walking Dataset on Partially Occluded Environment

Figure 7d ground truth data uncover three individuals, correctly predicted. G1 shows the strongest heat on the OBB's vertical edges, indicating a stable, wide-based gait; G2 consistently displays redder tops than the heel, suggesting forward-driven or faster movement. G3 reveals red at the bottom with a cooler center, indicating firm heel planting. The cosine similarity is 85.7%, demonstrating robust discriminability of features among a simpler grouping. Figure 7e ground truth data shows four individuals, yet the OBB prediction identifies five, highlighting the diminished discriminability with more people. G1 features intense red on the right vertical walls through the center; G2 exhibits a more scattered and balanced thermal signature, indicative of lighter footfalls. G3, G4 and G5 suffer misgrouping from overlaps, although cosine similarity remains elevated (87.2%) due to embedding overlaps. Figure 7f ground truth data indicates five individuals, but the model predicts seven owing to overlapping instances. G1 shows strong activation heatmaps on the heel, midfoot, and sides, suggesting a stable walker; G2 displays less prominent heatmaps, indicating a lighter person. G3–G7 are misgrouped due to overlaps.

### 7.5. Explanation of SEG-XTrail-ID Method on Single File Path Dataset on Occluded and Clear Environment

As illustrated in Figure 8a, the ground truth indicates that only two people are identified and follow a single-file trajectory. G1 displays a highly concentrated dark red intensity at the heel, which indicates a pronounced heel strike. G2 appears to be fainter and more dispersed, possibly indicating a lighter gait from someone landing with a flatter foot rather than a forceful heel impact. The model correctly categorizes these two clusters, achieving a cosine similarity of 83.8%. In Figure 8b, the ground truth confirms two individuals,

accurately detected by the model. G1 exhibits a more forceful heel strike, while G2 reveals a gentler, more tentative walker with evenly distributed and scattered heatmaps.

The ground truth of Figure 8c reveals four individuals—two following a single path on the left and two on the right—all properly classified. G1 show intense red along the outer edge of the barefoot, suggesting a gait that favors weight on the exterior. G2 displays predominantly yellow tones with isolated red spots, characteristic of a softer stepper. G3 presents subtler heatmaps with a faint red heel, while G4 exhibits tighter, more condensed patterns indicative of flat-footed walking. The similarity matrix reaches 91.3%, reflecting strong alignment across the footprints.

### *7.6. Explaination of OBB-XTrail-ID Method on Single File Path Dataset Single File Path Dataset on Occluded and Clear Environment*

Figure 8d presents the ground truth as two individuals; however, the model incorrectly associates overlapping and visually similar footprints, leading to an overestimation of four individuals. A comparable pattern is observed in Figure 8e, where the model predicts three individuals instead of two, and in Figure 8f, where five individuals are predicted instead of four. These results highlight a limitation of the OBB approach, particularly in scenarios where individuals traverse a common path, resulting in overlapping footprint patterns that degrade detection accuracy.

Figure 8d depicts ground truth for two individuals; yet, the model erroneously merges overlapping, visually similar footprints, resulting in a miscount of four. A similar issue arises in Figure 8e, with three predicted instead of two, and in Figure 8f, where five are identified rather than four. These results underscore a key drawback of the OBB method, especially when people share a path, causing footprint overlaps that impair detection precision.

## 8. Discussion

In this study, we evaluated two barefoot-print localization methods on soil substrates namely, OBBs and segmentation masks. We used frozen DINOv2 patch-level embeddings for feature extraction and subsequent clustering of footprint trails belonging to the same individual. Based on these methods, the assessment of three similarity matrices demonstrates that cosine similarity consistently outperformed the alternative metrics (see, Table 4), achieving the highest accuracy for both OBB and segmentation in correlating footprints with similar feature embeddings. The proposed methods were qualitatively and quantitatively evaluated in our diverse datasets to assess their practical forensic effectiveness. The OBB-XTrail-ID with cosine similarity, the best-performing correlation metric, achieved a promising accuracy of 91.5%. Despite this promising accuracy, a key limitation observed for the OBB-XTrail-ID approach is a reduction in the consistency of the grouping when more than three individuals are present, as shown in Figures 6f and 7e,f.

The OBB-XTrail-ID method provides compact object representations and reduces unnecessary background inclusion compared with conventional horizontal bounding boxes, particularly for arbitrarily oriented objects such as barefoot prints. However, they do not accurately capture the fine-grained morphometrics of barefoot prints. This limitation is consistent with findings in the oriented object detection literature, where OBBs improve localization and provide a more compact representation but still approximate object geometry rather than explicitly modeling contours or boundaries [48,49]. Therefore, when footprints are detected as OBBs, the Dinov2 vision transformer decomposes the footprint region into uniform patches that may cover both the footprint and minimal surrounding soil substrate. As a result, patch-level features may contain information from both relevant

morphological regions (toes, arch, heel) and irrelevant surrounding soil, potentially diluting the discriminative power of the global embedding.

However, we hypothesize that minimal background influence on the global feature representation is expected to be limited when footprint regions dominate the OBB area. This is evident because the visualizations of the X-Trail-ID attribution heatmap demonstrate that ViT is still able to focus on prominent footprint regions, with stronger activations concentrated around prominent anatomical structures [50,51].

In contrast, SEG-XTrail-ID improves performance by isolating footprint regions using segmentation masks, enabling more object-centric feature extraction and reducing background noise. When combined with cosine similarity-based matching, this results in a more robust and practically applicable framework for correlating barefoot feature embeddings in forensic contexts. Consequently, SEG-XTrail-ID consistently outperformed all variants of OBB-XTrail-ID in our experimental setup, achieving an average accuracy of 94.2%. This improvement may be attributed to the ability of the SEG-XTrail-ID method to guide the Vision Transformer to focus on relevant footprint regions, which is consistent with prior works on transformer attention [52]. The observed performance further supports the motivation that our proposed method is more suitable for barefoot impression analysis on natural soil substrates, where classical image-processing methods have been reported to be limited by noise, low contrast, and variability in footprint appearance [53,54].

Therefore, our approach aligns well with the observations in [55–57] that deep learning-based approaches enhance the robustness of feature learning from challenging datasets while providing greater potential for interpretable and discriminative representations. As shown in Figure 6a–c, on soft clear soil, the visualization of the XTrail-ID heatmaps indicate a clear focus on critical morphological regions, leading to embeddings that capture individual-specific characteristics more accurately.

The aggregation of patch-level embeddings into a global footprint descriptor produces a more robust and discriminative representation for similarity computation and clustering, even with increasing numbers of individuals and environmental disturbances such as partial grass occlusion. This is evident in Figure 7a–c, where the model correctly associated footprints belonging to an individual under challenging conditions. This is consistent with previous work showing that combining segmentation with Vision Transformer-based architectures improves feature quality by suppressing background information and improving object-centric representation learning [58–60].

## 9. Study Limitations

This study presents several limitations that must be acknowledged. First, the dataset comprises barefoot print images collected from 10 male and 10 female participants. Although the dataset was sufficient to evaluate the XTrail-ID framework, a larger and more diverse participant population would provide a stronger basis to assess generalizability in broader demographic groups. Second, currently there is no widely adopted benchmark dataset for barefoot print trail grouping from UAV soil imagery. Consequently, the evaluation was conducted using a custom dataset collected for this study. Although this dataset allowed for assessment of the proposed framework, the absence of standardized benchmark datasets limits direct comparison with existing approaches. Third, although the dataset includes variations in compact and loose soft-soil substrates, illumination conditions, and natural grass occlusions, it does not cover the full range of environmental conditions that may be encountered in different operational settings. Factors such as severe footprint degradation, adverse weather, dense vegetation, and additional substrate types were not investigated. Therefore, further evaluation may be necessary under different settings.

## 10. Conclusions and Future Direction

This study presented XTrail-ID, an unsupervised framework for grouping barefoot footprint trails from UAV imagery using self-supervised vision transformer embeddings and a cosine similarity-based feature association. The results demonstrated that cosine similarity applied to normalized feature embeddings effectively groups footprints belonging to the same individual, with SEG-XTrail-ID consistently outperforming OBB-XTrail-ID due to improved footprint localization.

In addition, the proposed Embedding Spatial Attribution Module (ESAM) enhances interpretability by transforming frozen Vision Transformer feature embeddings into spatial attribution heatmaps that highlight discriminative footprint regions, addressing the black-box nature of similarity-based grouping in forensic applications. Future work will focus on evaluating the framework in larger, more diverse datasets and soil substrates, improving robustness under broader environmental conditions to further establish generalizability.

**Author Contributions:** Conceptualization, data curation, methodology, software, validation, writing—original draft preparation, formal analysis, W.M., R.S.J.J., Z.C.J., S.K.A., O.M. and T.P.; funding acquisition, investigation, resources, R.S.J.J.; editing, visualization, Z.C.J.; project administration, review and editing T.P.; supervision, R.S.J.J., O.M. and S.K.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** Research was sponsored by the U.S Army Research Office under Grant W911NF-23-1-0293. The views and conclusions are those of the authors and do not represent official policies of the Army Research Office or the U.S Government.

**Institutional Review Board Statement:** On behalf of the Human Ethics Research Committee, I hereby give ethical approval in respect of the undertakings contained in the above-mentioned project and research instrument(s). Should any other instruments be used, these require separate authorization. The researcher may therefore commence with the research as from the date of this certificate, using the reference number HREC-12. The study was approved by Human Ethics Research Committee of Botswana International University of Science and Technology (protocol code: HREC-12, approval date: 10 August 2023) for studies involving humans. Data collection on the wild was also approved by the Ministry of Environment and Tourism (Permit No ENT 8/36/4 LVII (7)).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The authors will liaise with the Human Ethics Research Committee for guidance on the ethical implications and best practices for providing the research data. Once the manuscript is published, the data may be provided ONLY upon request. The research community will be provided with a license agreement to sign, stating that they will not in any manner distribute or use the data for commercial purposes. Upon aproval the research data will be found https://github.com/wacco1/XTrail-ID accessed on 10 June 2026.

**Acknowledgments:** The authors acknowledge the BIUST Human Research Ethics Committee for granting the required permits to conduct the human experiments. We are also grateful to the Ministry of Environment and Tourism for issuing us permits to collect data in wildlife areas. We are grateful to the Botswana Defence Force for providing the resources necessary to carry out this research, and to the U.S. Army Research Office for funding support under Grant W911NF-23-1-0293, which enabled the development of this study. Finally, we sincerely thank all the volunteers who participated in the barefoot print data collection on soil substrates.

**Conflicts of Interest:** Author Zoe C. Jewell and Sky K. Alibhai was employed by the company WildTrack Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References


1. Mukhra, R.; Krishan, K.; Kanchan, T. Bare footprint metric analysis methods for comparison and identification in forensic examinations: A review of literature. *J. Forensic Leg. Med.* **2018**, *58*, 101–112. [CrossRef] [PubMed]
2. Seal, S.; Soták, M.; Švábová, P.; Krishan, K.; Beňuš, R. Dynamic footprints in person identification: A review of forensic applications and future prospects. *Egypt. J. Forensic Sci.* **2026**, *16*, 17. [CrossRef]
3. Ardhianto, P.; Liau, B.Y.; Jan, Y.K.; Tsai, J.Y.; Akhyar, F.; Lin, C.Y.; Subiakto, R.B.R.; Lung, C.W. Deep learning in left and right footprint image detection based on plantar pressure. *Appl. Sci.* **2022**, *12*, 8885. [CrossRef]
4. Singh, B.; Krishan, K.; Kaur, K.; Kanchan, T. Estimation of body weight from the base of gait and the area swept in one stride—forensic implications. *Egypt. J. Forensic Sci.* **2018**, *8*, 48. [CrossRef]
5. Burrow, J.G.; Kelly, H.D.; Francis, B.E. Forensic podiatry-an overview. *J. Forensic Sci. Crim. Investig.* **2017**, *5*, 555666. [CrossRef]
6. Müller-Franzes, G.; Khader, F.; Siepmann, R.; Han, T.; Kather, J.N.; Nebelung, S.; Truhn, D. Medical slice transformer for improved diagnosis and explainability on 3D medical images with DINOv2. *Sci. Rep.* **2025**, *15*, 23979. [CrossRef] [PubMed]
7. Sams, F.R.; Supti, S.K.; Hamid, S.B.; Junayed, R.; Bari, K.F.A.; Ali, M.J.; Gani, R.; Shams, K.; Rashid, M.R.A.; Islam, R.U. Real-Time Sunflower Detection Using Semi-Supervised and Self-Supervised Deep Learning for Precision Agriculture. *Smart Agric. Technol.* **2025**, *13*, 101684.
8. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
9. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
10. Gunn, N. Old and new methods of evaluating footprint impressions by a forensic podiatrist. *Br. J. Podiatr. Med. Surg.* **1991**, *3*, 8–11.
11. Smerecki, C.; Lovejoy, C. Identification via pedal morphology. *Identif. News* **1985**, *35*, 5–12.
12. Hu, A.; Arnold, J.B.; Causby, R.; Jones, S. The identification and reliability of static and dynamic barefoot impression measurements: A systematic review. *Forensic Sci. Int.* **2018**, *289*, 156–164. [CrossRef] [PubMed]
13. Robbins, L.M.; Gantt, R. *Footprints: Collection, Analysis, and Interpretation*; Charles C Thomas Publisher: Springfield, IL, USA, 1985.
14. Reel, S.M.L. Development and Evaluation of a Valid and Reliable Footprint Measurement Approach in Forensic Identification. Ph.D. Thesis, University of Leeds, Leeds, UK, 2012.
15. Osisanwo, F.; Adetunmbi, A.; Alese, B. Forensic analysis and the inimitability of human footprints. *Int. J. Intell. Comput. Res.* **2015**, *6*, 527–530. [CrossRef]
16. Švábová, P.; Falbová, D.; Kozáková, Z.; Chovancová, M.; Vorobeľová, L.; Beňuš, R. Analysis of Footstep/Stride Length from Gait Patterns of Dynamic Footprints as a Parameter for Biological Profiling—A Preliminary Study. *Forensic Sci.* **2025**, *5*, 29. [CrossRef]
17. Domjanic, J.; Fieder, M.; Seidler, H.; Mitteroecker, P. Geometric morphometric footprint analysis of young women. *J. Foot Ankle Res.* **2013**, *6*, 27. [CrossRef] [PubMed]
18. Liu, H.; Ding, N.; Lin, S.; Lv, H.; Liu, X. Research on the measurement method of barefoot footprint similarity. In *Proceedings of the International Conference on Statistics, Data Science, and Computational Intelligence (CSDSCI 2022)*; SPIE: Bellingham, WA, USA, 2023; Volume 12510, pp. 112–118.
19. Kennedy, R.B.; Pressman, I.S.; Chen, S.; Petersen, P.H.; Pressman, A.E. Statistical analysis of barefoot impressions. *J. Forensic Sci.* **2003**, *48*, JFS2001337. [CrossRef]
20. Sun, K.; Luo, Y. Reliability study of the Chinese method for bare footprint linear measurement. *Forensic Sci. Res.* **2025**, *10*, owaf049. [CrossRef] [PubMed]
21. Zhou, X.; Zhang, Y.; Fan, X.; Wang, N.; Bao, W. MFID-200: A Multimodal Footprint Dataset for Identification. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5740876 (accessed on 11 February 2026).
22. Zhou, B.; Singh, M.S.; Doda, S.; Yildirim, M.; Cheng, J.; Lukowicz, P. The Carpet Knows: Identifying People in a Smart Environment from a Single Step. In *Proceedings of the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*; IEEE: New York, NY, USA, 2017; pp. 527–532.
23. Zhang, C.; Pan, S.; Qi, Y.; Yang, Y. A Footprint Extraction and Recognition Algorithm Based on Plantar Pressure. *Trait. Du Signal* **2019**, *36*, 419–424. [CrossRef]
24. Xiang, L.; Gu, Y.; Mei, Q.; Wang, A.; Shim, V.; Fernandez, J. Automatic classification of barefoot and shod populations based on the foot metrics and plantar pressure patterns. *Front. Bioeng. Biotechnol.* **2022**, *10*, 843204. [CrossRef] [PubMed]
25. Bao, W.; Xu, Q.; Wang, N.; Yang, X. Barefoot Pressure Image Recognition and Analysis in Natural Walking State Based on Selective Fusion Attention Network. In *Proceedings of the International Conference on Mechanisms and Robotics (ICMAR 2022)*; SPIE: Bellingham, WA, USA, 2022. [CrossRef]
26. Abuqadumah, M.M.A.; Ali, M.A.M.; Abd Almisreb, A.; Durakovic, B. Deep Transfer Learning for Human Identification Based on Footprint: A Comparative Study. *Period. Eng. Nat. Sci.* **2019**, *7*, 1300–1307. [CrossRef]

27. Hild, M. Person identification based on barefoot 3D sole shape. In *Proceedings of the International Workshop on Computational Forensics*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 84–95.
28. Mmereki, W.; Jamisola, R.S., Jr.; Jewell, Z.C.; Petso, T.; Matsebe, O.; Alibhai, S.K. Investigating Bounding Box, Landmark, and Segmentation Approaches for Automatic Human Barefoot Print Classification on Soil Substrates Using Deep Learning. *Forensic Sci.* **2025**, *5*, 56. [CrossRef]
29. Shen, Y.; Jiang, X.; Zhao, Y.; Xie, W. Barefoot Footprint Detection Algorithm Based on YOLOv8-StarNet. *Sensors* **2025**, *25*, 4578. [CrossRef] [PubMed]
30. Jin, Y. Algorithm of personal recognition based on multi-scale features from barefoot footprint image. *Forensic Sci. Technol.* **2022**, *47*, 587–592.
31. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
32. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 16000–16009.
33. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
34. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 9729–9738.
35. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 9650–9660.
36. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
37. Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; Kong, T. ibot: Image bert pre-training with online tokenizer. *arXiv* **2021**, arXiv:2111.07832.
38. Li, F.F.; Fergus, R.; Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop*; IEEE: New York, NY, USA, 2004; p. 178.
39. Ding, A. DINO Features for Salient Object Detection. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2025.
40. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 38–55.
41. Mtasher, A.K.; Al-wakel, E.H.J. Custom Object Detection Using Transfer Learning with Pretrained Models for Improved Detection Techniques. *J. Multiapp* **2024**, *5*, 10–18. [CrossRef]
42. Farzmahdi, A.; Zarco, W.; Freiwald, W.A.; Kriegeskorte, N.; Golan, T. Emergence of brain-like mirror-symmetric viewpoint tuning in convolutional neural networks. *eLife* **2024**, *13*, e90256. [CrossRef] [PubMed]
43. Brachmann, A.; Redies, C. Using Convolutional Neural Network Filters to Measure Left-Right Mirror Symmetry in Images. *Symmetry* **2016**, *8*, 144. [CrossRef]
44. Bianco, S.; Buzzelli, M.; Ciocca, G.; Piccoli, F.; Schettini, R. A study on the generalization of DINOv2 features for food recognition tasks: A unified evaluation framework. *Intell. Syst. Appl.* **2026**, *29*, 200632. [CrossRef]
45. Darcet, T.; Oquab, M.; Mairal, J.; Bojanowski, P. Vision transformers need registers. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; Volume 2024, pp. 2632–2652.
46. Zhang, B.; Chen, Y.; Bai, L.; Zhao, Y.; Sun, Y.; Yuan, Y.; Zhang, J.; Ren, H. Learning to adapt foundation model dinov2 for capsule endoscopy diagnosis. *Procedia Comput. Sci.* **2024**, *250*, 188–194. [CrossRef]
47. Lin, L.; Hedayat, A.S.; Sinha, B.; Yang, M. Tests for assessment of agreement using probability criteria. *J. Stat. Plan. Inference* **2006**, *136*, 207–224.
48. Wang, Y.; Wang, L.; Lu, H.; He, Y. Segmentation based rotated bounding boxes prediction and image synthesizing for object detection of high resolution aerial images. *Neurocomputing* **2020**, *388*, 202–211. [CrossRef]
49. Zand, M.; Etemad, A.; Greenspan, M. Oriented bounding boxes for small and freely rotated objects. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
50. Chefer, H.; Gur, S.; Wolf, L. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 782–791.
51. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359.
52. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2961–2969.

53. Song, J.; Jiao, W.; Lankowicz, K.; Cai, Z.; Bi, H. A two-stage adaptive thresholding segmentation for noisy low-contrast images. *Ecol. Inform.* **2022**, *69*, 101632. [CrossRef]
54. Yu, Y.; Wang, C.; Fu, Q.; Kou, R.; Huang, F.; Yang, B.; Yang, T.; Gao, M. Techniques and challenges of image segmentation: A review. *Electronics* **2023**, *12*, 1199. [CrossRef]
55. Reddy, N.L.; Gopinath, M. Advanced deep learning framework for soil texture classification. *Sci. Rep.* **2025**, *15*, 34407. [CrossRef] [PubMed]
56. He, C.; Liu, Y.; Wang, D.; Liu, S.; Yu, L.; Ren, Y. Automatic extraction of bare soil land from high-resolution remote sensing images based on semantic segmentation with deep learning. *Remote Sens.* **2023**, *15*, 1646. [CrossRef]
57. Romero, A.; Gatta, C.; Camps-Valls, G. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 1349–1362. [CrossRef]
58. Hussien, A.; Elkhateb, A.; Saeed, M.; Elsabawy, N.M.; Elnakeeb, A.E.; Elrashidy, N. Explainable self-supervised learning for medical image diagnosis based on DINO V2 model and semantic search. *Sci. Rep.* **2025**, *15*, 32174. [CrossRef] [PubMed]
59. Hatamizadeh, A.; Xu, Z.; Yang, D.; Li, W.; Roth, H.; Xu, D. Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation. *arXiv* **2022**, arXiv:2204.00631.
60. Sagar, A. Vitbis: Vision transformer for biomedical image segmentation. In *Proceedings of the MICCAI Workshop on Distributed and Collaborative Learning*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 34–45.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.