Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18469-5

1236: EXPLAINABLE ARTIFICIAL INTELLIGENCE SOLUTIONS FOR IN-THE-WILD HUMAN BEHAVIOR ANALYSIS

Exploring biometric domain adaptation in human action recognition models for unconstrained environments

David Freire-Obregón[1] · Paola Barra[2] · Modesto Castrillón-Santana[1] · Maria De Marsico[3]

Received: 29 September 2023 / Revised: 22 December 2023 / Accepted: 29 January 2024

Abstract

In conventional machine learning (ML), a fundamental assumption is that the training and test sets share identical feature distributions, a reasonable premise drawn from the same dataset. However, real-world scenarios often defy this assumption, as data may originate from diverse sources, causing disparities between training and test data distributions. This leads to a domain shift, where variations emerge between the source and target domains. This study delves into human action recognition (HAR) models within an unconstrained, real-world setting, scrutinizing the impact of input data variations related to contextual information and video encoding. The objective is to highlight the intricacies of model performance and interpretability in this context. Additionally, the study explores the domain adaptability of HAR models, specifically focusing on their potential for re-identifying individuals within uncontrolled environments. The experiments involve seven pre-trained backbone models and introduce a novel analytical approach by linking domain-related (HAR) and domain-unrelated (re-identification (re-ID)) tasks. Two key analyses addressing contextual information and encoding strategies reveal that maintaining the same encoding approach during training results in high task correlation while incorporating richer contextual information enhances performance. A notable outcome of this study is the comprehensive evaluation of a novel transformer-based architecture driven by a HAR backbone, which achieves a robust re-ID performance superior to state-of-the-art (SOTA). However, it faces challenges when other encoding schemes are applied, highlighting the role of the HAR classifier in performance variations.

Keywords Human action recognition · Biometrics · Transformers · Domain adaptation

✉ David Freire-Obregón
david.freire@ulpgc.es

[1] SIANI, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

[2] Università di Napoli Parthenope, Naples, Italy

[3] Sapienza Università di Roma, Rome, Italy

Published online: 16 February 2024

1 Introduction

The goal of ML is to use a set of training samples and a suitable objective function to train a model that minimizes misclassifications when applied to unseen test data. However, this process usually relies on a fundamental assumption: the training and test data originate from the same distribution and share similar joint probability distributions. In the real world, such an assumption often crumbles as training and test sets can stem from distinct feature spaces or distributions [9]. Challenges arise when classifying new instances that do not match the training data properties, dimensions, and distribution. This situation can occur due to various factors, such as collecting new samples from diverse sources, leading to a domain shift. For instance, this is typical when embedding and deploying a trained model in a real-world application, where data generally comes from a less controlled environment. However, when the training data does not accurately reflect the distribution of the test data, the performance of the trained model is likely to suffer during testing and even more during real operation and deployment. To address this challenge, researchers have introduced a field in ML known as domain adaptation. In this context, the training and test sets are called the source and target domains, respectively. Domain adaptation endeavors to develop a model from labeled source data that can be effectively applied to a target domain by minimizing the dissimilarity between their respective data distributions.

Recent advancements in domain adaptation for deep learning (DL) models have addressed essential challenges. Researchers have explored techniques to adapt pre-trained models to varying domains efficiently, enhancing domain alignment and performance [28]. Additionally, strategies have been introduced to mitigate domain shift issues and reduce overfitting when applying pre-trained models to new domains [16, 18]. Interpretability in DL models has also been addressed due to the growing concerns of adversarial attacks, biases stemming from contaminated training data, and the legal demand for explanation in intelligent decision-making systems [3], especially in forensic applications. The applicability and effectiveness of transfer learning (TL) have been further examined in the context of specialized or niche computer vision domains [23]. These developments collectively contribute to the ongoing progress in domain adaptation for different DL models.

In general, TL re-uses knowledge (models) learned from a task to perform well on a related task with possibly less training. “Domain adaptation is a subcategory of TL. In domain adaptation, the source and target domains all have the same feature space (but different distributions); in contrast, TL includes cases where the target domain’s feature space is different from the source feature space or spaces” [38]. Domain adaptation is often referred to as domain shift or distributional shift.

Labels play a key role in defining and distinguishing the context and purpose of a dataset within a specific domain. They serve as the identifiers that categorize data points and guide the application of ML algorithms. However, it is essential to recognize the type of domain the dataset belongs to; the same set of features can find application in diverse domains, each with distinct objectives. Features, such as numerical measurements, text, or images, possess inherent qualities and patterns agnostic to the domain. As a result, these features may be reinterpreted to serve different goals across various fields of study. This demonstrates the versatility of data-driven techniques and their potential to uncover valuable insights beyond their original context.

It is possible to further distinguish closed-set domain adaptation from open-set domain adaptation. When the images of the source and target domain represent the same set of categories, this entails a closed set domain adaptation. However, for most realistic applications of the adaptation strategy, the assumption that the target domain contains only images of the same categories as the source domain is too restrictive. For most applications, the target domain dataset contains many images that do not belong to the classes of interest. For this reason, [25] proposes open set domain adaptation, which avoids the above unrealistic assumption.

In this regard, partial domain adaptation arises when the label set in the target domain is a subset of the labels existing in the source domain [4, 48]. In this context, the source domain encompasses many classes, while the target domain encompasses only a subset of these labels, representing fewer classes (domain-related domains/tasks). On the other hand, when we have a dataset with feature spaces different from those in the target dataset, it falls under the category of heterogeneous TL (domain-unrelated domains/tasks). It is interesting to make an example of the role of labels and features when used in different domains. Labels characterize the kind of domain task, while the same features may be used for different tasks with different goals. For instance, MEL spectrograms (features) may be used in speaker recognition (label=identity), emotion recognition (label=emotion), or speech recognition (label=utterance).

In light of this, the paper examines how HAR models can be reused in the context of a biometrics-related task. The presented study evaluates various HAR models under different configurations using unconstrained inputs. As shown in Fig. 1, this entails two experiments with these inputs:

1. A partial domain adaptation experiment, namely domain-related, where the label assigned to the video footage (jogging) is a subset of the labels present in the source domain (which includes 400 distinct actions [21]).
2. A heterogeneous TL experiment, i.e., a domain-unrelated one, explores HAR models for the task of athlete re-ID, which is rather related to biometrics.

The results of these experiments come from both a HAR classifier and a transformer-based re-ID classifier fed with features extracted from a HAR backbone. The primary goal is to scrutinize the correlation (indicated as r from now on) between these outcomes, shedding light on how domain shift impacts the performance of these models in biometrics-related applications. The main contributions of this work can be summarized as follows:

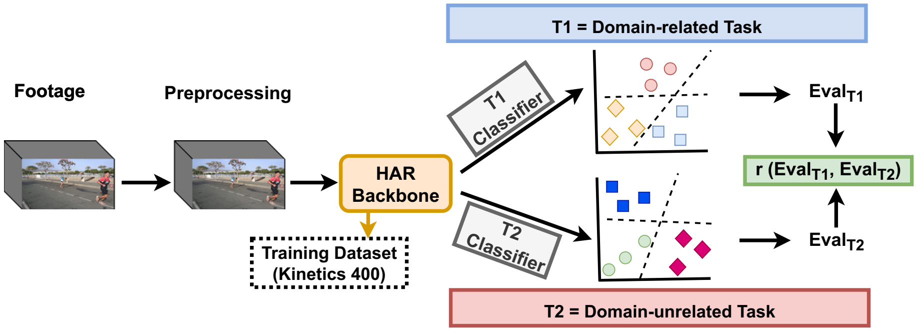

Fig. 1 Domain-shift correlation (in the following indicated as r). We conducted two experiments using unconstrained inputs. One experiment is related to the same source domain, while the other is not. We employed seven pre-trained HAR models (backbones) in these experiments to assess how domain shift affects performance and correlations of architectures using their extracted features. To achieve this, we modified the contextual information or video encoding inputs

- Exploration of task interrelationships. The paper introduces an analysis method to assess domain shift adaptation by correlating two tasks, namely HAR and re-ID, using the same input data. This approach provides insights into how different tasks are affected by domain-related factors.
- The development of a novel transformer-based classifier. We devise a novel transformer classifier specifically tailored to address the re-ID problem while utilizing pre-trained models for the HAR task. This classifier design significantly improves the solution for the re-ID problem. In this case, it works in the context of domain shift adaptation. However, the transformer can be applied to domain-specific tasks, too. In other words, it has no unique elements to make it "domain-shift" specialized, but it performs remarkably in this case.
- Analysis on the influence of Contextual Information. The experiments explore how contextual information impacts the evaluation of both HAR and re-ID tasks and reveal that the richer it is, the more favorable evaluation outcomes it allows to achieve for both tasks, achieving superior performance compared to the current state-of-the-art (SOTA) in the re-ID task with the proposed transformer-based architecture when considering HAR embeddings as input.
- Analysis on the influence of Encoding. Appearance privacy is important as it safeguards individuals from unwarranted scrutiny and protects their personal autonomy. In this regard, the results highlight significant challenges in achieving acceptable evaluation rates when considering the encoding experiment. HAR evaluation rates decline when using specific encoding schemes. However, a depth encoding approach (MiDaS) performs well on re-ID while poorly on HAR. This finding suggests that the performance of the HAR classifier is significantly negatively affected when using specific encoding schemes. The results emphasize the importance of selecting appropriate encoding methods for different tasks.

2 Related work

DL often exploits pre-trained HAR models to TL in tasks like sign language recognition [17], violence detection [13, 34] or person re-ID [15]. This approach may seem advantageous because it generally requires less data to train the final model. However, it comes with several notable disadvantages that must be carefully considered.

Pre-trained models are often tailored to specific domains and perform exceptionally well in domain-related tasks [43]. However, the overfitting risk on these tasks when applying closed-set, open-set, or partial domain adaptation becomes significant. The HAR pre-trained models are typically trained on large datasets encompassing various actions and scenarios. Fine-tuning them on a specific task like fight detection may lead to the model memorizing too general characteristics of the source domain data, hindering a sufficient specialization to the target domain [12]. Consequently, the model might struggle to capture the intricate patterns and nuances of fight-related actions, leading to sub-optimal performance and the inability to effectively detect fights in real-world scenarios. In this regard, several works in literature have dealt with domain-related tasks. Some papers have proposed violence recognition algorithms using HAR-based models like SlowFast networks [8, 45], or I3D [13]. In healthcare, this architecture has proven beneficial by employing sensor TL from a deep SlowFast network on video data to capture physiological data and movement in individuals with spinal cord injuries [2]. In addition, SlowFast backbones have been used to predict stimming behaviors in children with autism spectrum disorder from uncontrolled video recordings [31]. Additional HAR domain-related applications include educational ones, like the student engagement recognition network based on I3D presented in [47], as well as the use for anomalous event detection [22].

Similarly, using HAR pre-trained models for domain-unrelated tasks (heterogeneous TL) may present challenges [7]. These pre-trained models are designed to recognize a broad spectrum of actions, making them less specific to the intricacies of other tasks like deepfakes detection [45]. Consequently, these models may not accurately distinguish between similar individuals in complex scenarios, leading to misclassification.

Last, domain adaptation poses a significant issue when applying pre-trained models to target tasks with different environmental settings and camera perspectives. The pre-trained models' source domain may differ significantly from the target domain regarding illumination conditions, camera angles, and demographics [9]. This domain shift can adversely affect the model's ability to adapt, resulting in decreased performance and reduced reliability in domain-related and domain-unrelated TL tasks.

In this context, efforts have been dedicated to employ HAR-based networks for various purposes, including detecting manipulated videos to combat the proliferation of fake information [45], person re-ID in videos [15], hand gesture recognition [46], wildlife detection [40], and sporting assessment analysis [14].

In summary, although pre-trained models for HAR provide a promising starting point for TL, it is crucial to consider the drawbacks when applying them to domain-related and domain-unrelated tasks. Potential issues such as overfitting, a lack of task-specificity, challenges in domain adaptation, interpretability issues, biases, and ethical concerns can all pose obstacles that may negatively impact the effectiveness and reliability of pre-trained models in these tasks. Our approach aims to analyze a domain-specific and a non-domain-specific task using the same inputs, evaluating their performance to explore potential correlations between these two forms of TL. Firstly, we evaluate over a domain-specific task involving input footage, such as HAR. Secondly, we test a non-domain-specific task, such as re-ID, using the same input footage.

3 Overview of the reported analysis

This section focuses on the two pivotal pipelines designed for our experiments. These pipelines are necessary in our study, each serving distinct purposes within video analysis. Figure 2 illustrates them. The upper one is highlighted in blue and represents the HAR classifier used for processing HAR backbone features. The following pipeline, distinguished in red and fed by the same features, is dedicated to re-ID and incorporates transformer encoder blocks tailored for this task, which generate embeddings trained with triplets. The upcoming subsections will explore both pipelines and footage pre-processing in greater detail.

Pre-processing To enhance the quality of the backbone's embeddings, the input footage provided to the action recognition networks must be clean and devoid of extraneous elements [14]. In the case of the dataset used for the experiments presented herein, they are collected videos from an ultra-distance running competition (see Section 4.1), these extraneous elements include other athletes, race staff, and moving vehicles. Since they are irrelevant in the HAR context, an initial pre-processing block prepares the raw input data by isolating the runner of interest. For this purpose, ByteTrack [50], a multi-object tracking network, accurately tracks the runner within each footage. Subsequently, context-constrained pre-processing techniques create a suitable scenario for the experiments.

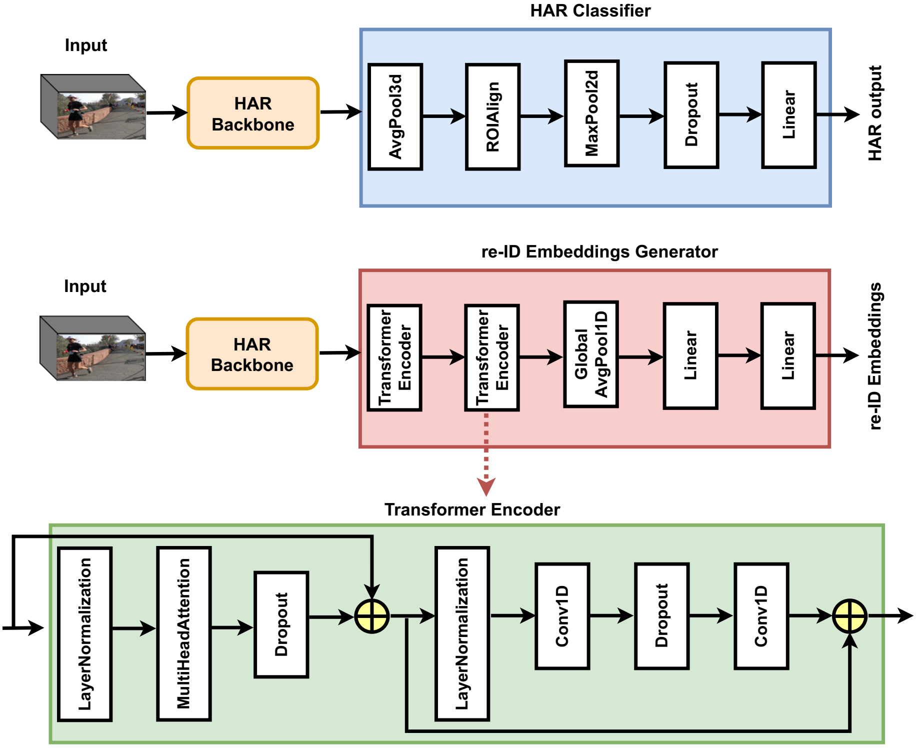

Fig. 2 Pipelines overview. The experiments rely on two distinct pipelines. The first pipeline (the upper one, highlighted in blue) specifies the HAR classifier, commonly employed in video analysis models to process features extracted from the HAR backbone. The second pipeline (the one below, highlighted in red) processes features extracted from the same HAR backbone, is tailored for re-ID and incorporates transformer encoder blocks explicitly designed for this task. It produces embeddings trained with triplets. The bottom part of the figure depicts a single transformer encoder block (highlighted in green)

It was necessary to acquire context-constrained footage frames for a specific runner, denoted as i, at a given time t within an interval $[0, T]$ and recorded from a Recording Point RP (a point where a video camera is positioned) within a range $[0, P]$. For each runner i, the Bounding Box $BB_i(t, RP)$ encompasses the area covered by the runner's i body in the frame recorded at time t from the point RP. Given an original frame $F_i(t, RP)$, two key factors in this step are the bounding box area of the runner $BB_i(t, RP)$, and the average number of frames denoted as $\tau(RP)$, required to create a static background against which the single runner i is situated in the pre-processed footage. The resulting pre-processed footage frame, denoted as $F_i'(t, RP)$, is obtained through a process expressed by the following:

$$F_i'[RP] = BB_i(t, RP) \cup \tau(RP) \tag{1}$$

where $\cup$ denotes the operation that aligns and overlays the BB of runner i to the average of the selected number of $\tau(RP)$ frames (see Fig. 3). The new footage is obtained by the sequence of pre-processed frames.

Using the average frame is beneficial to capture a clean action recognition pattern and mitigate the influence of extraneous moving elements. Assuming a static camera, this process

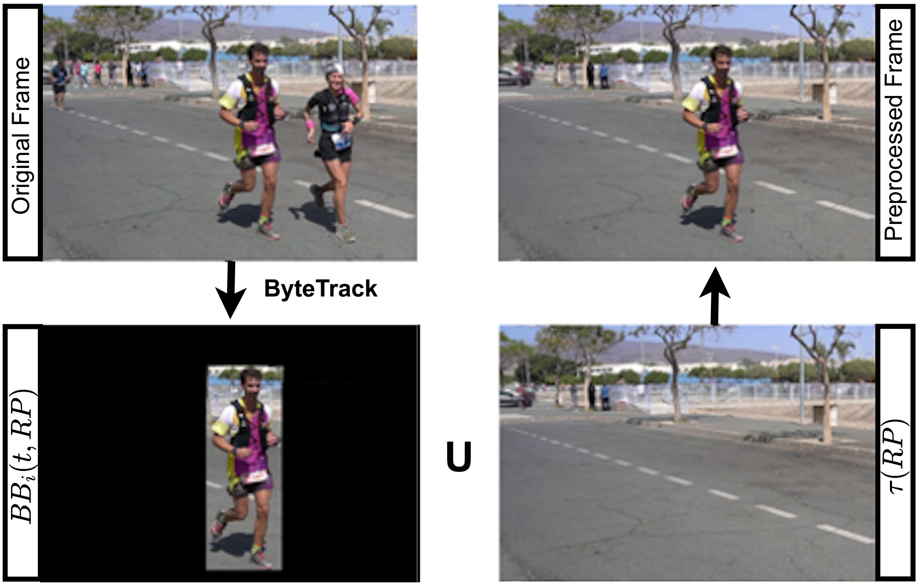

Fig. 3 Creation of context-constrained footage for the used dataset. For each frame at time t and Recording Point RP, the procedure extracts the bounding box of runner i, which is aligned and overlayed on a clean static background obtained as the average of $\tau(RP)$ frames

helps to isolate the person of interest and facilitates a focused analysis of actions without interference from other dynamic elements in the scene.

In scenes where more than one athlete is present, the runner of interest is chosen based on the specific requirements of the task at hand, namely person re-identification (re-ID). The utilized dataset provides bounding boxes for the runners, and even if there are multiple runners in a given scene, the focus is on capturing the action recognition patterns of a specific individual. Consequently, for the purpose of the re-ID task, the experiments isolate the runner of interest from the bounding boxes provided by the dataset.

This targeted selection ensures that the subsequent analysis centers on the identified individual, accurately allowing the study and characterization of the action recognition patterns.

3.1 Backbones

The transformed input footage, comprising a total of n frames, undergoes a two-step process. It is first subjected to downsampling and subsequently divided into m video clips denoted as $v_1, ..., v_m$, with each clip encompassing a sequence of q consecutive frames that encapsulate a snapshot of the activity. In practice, the m clips partially overlap since each clip is one frame apart from the previous one. These video clips then traverse a pre-trained HAR encoder (backbone), yielding r-dimensional feature vectors. Notably, these encoder models have been previously trained on the Kinetics 400 dataset [21], encompassing a wide range of 400 action categories. Once the feature vectors for all n video clips are obtained, an average pooling layer ensures equitable contribution from all clips.

Utilizing HAR models in the initial phase to extract features is a widely employed technique, as elaborated in the SOTA section. Recently, Chen et al. conducted a study on unsupervised domain adaptation, as detailed in [6], where they employed backbones such as SlowFast in the initial stage. Similarly, our approach leverages various backbones for this purpose. We have explored seven distinct backbone models, all categorized into five architectures:

1. **C2D** (Convolutional 2D). The C2D model is tailored to process 2D spatial images, representing individual frames within video clips [37]. It employs a convolutional neural network (CNN) architecture similar to those in image classification tasks. Typically, the C2D model incorporates multiple convolutional layers followed by pooling layers, enabling it to extract progressively intricate features from the input frames. These convolutional layers apply learned filters to the input frames, yielding feature maps that capture spatial information. Subsequently, the pooling layers downsample the feature maps, reducing spatial resolution while preserving the most salient features. The resulting feature maps are then flattened into a feature vector.
2. **I3D** (Inflated 3D ConvNet). Since it operates on short video clips represented as 3D spatiotemporal volumes, the I3D model is a pivotal component of our selection [5]. It employs a two-stream approach, where one stream processes RGB images while the other processes optical flow images, effectively capturing appearance and motion cues. The RGB stream is initialized with weights pre-trained on extensive image classification datasets such as ImageNet. In contrast, the flow stream starts with random initialization and is fine-tuned with the RGB stream. The ultimate output of the I3D model is a feature vector that encapsulates appearance and motion information extracted from the input video clip.
3. **I3D NLN** (I3D Non-local Network). I3D NLN, a modified iteration of the I3D model, incorporates non-local operations to enhance the modeling of spatiotemporal dependencies in videos [44]. Like its predecessor, I3D NLN operates on 3D spatiotemporal volumes and adopts a two-stream architecture comprising RGB and optical flow streams. However, instead of the Inception module, I3D NLN integrates non-local blocks that facilitate learning long-range dependencies between any two positions within the input feature maps. These non-local blocks compute a weighted sum of input features from all positions based on similarity, allowing for capturing global context information and an improved representation of temporal dynamics.
4. **Slow**. The Slow model employs a two-stream architecture to capture short-term and long-term temporal dynamics within videos [11]. Like the C2D model, Slow processes high-resolution frames at a reduced frame rate, adding a temporal-downsampling layer to capture extended temporal dynamics. The reported experiments assessed two adaptations of Slow, denoted as Slow8x8 (henceforth referred to as S8x8) and Slow4x16 (henceforth referred to as S4x16). The key difference between them is the number of frames used for prediction and the respective sampling rates. S8x8 involves the consideration of 8 frames with a sampling rate of 8, while the S4x16 configuration employs 4 frames with a sampling rate of 16.
5. **SlowFast**. This model comprises a slow pathway designed to process high-resolution frames at a lower frame rate, enabling it to capture spatial information and long-term temporal structure [10]. Additionally, it features a fast pathway that processes low-resolution frames at a faster frame rate, capturing fine-grained motion information and short-term temporal structure. The slow pathway employs a deep 3D CNN, processing each frame in a video sequence with a temporal stride of 16 frames, while the fast pathway consists of a shallower 3D CNN, processing every other frame with a temporal stride of 2 frames. The final video-level representation is obtained by combining the outputs of these two pathways through a fusion module that employs a weighted sum of the features. Similarly

to Slow, the experiments used two variations of SlowFast, with the primary distinction in the fusion kernel size: 5 for SlowFast4x16 (hereafter SF4x16) and 7 for SlowFast8x8 (hereafter SF8x8).

These diverse backbone models present an outstanding opportunity to offer a comprehensive view of our proposal on a larger scale.

3.2 Task-specific classifiers

As anticipated in Fig. 2, while the two tasks use the same backbone trained for HAR, the pipelines differ for the classifier used afterward according to the kind of task. The domain-related task still deals with human action (same domain of the backbone) but tackles the classification of a subset of the actions considered to train the backbone. For this task, a HAR classifier is used again. The re-id task rather deals with a different kind of problem (not strictly related to the backbone-related domain). Therefore, a classifier focusing on different feature patterns is used. The following paragraphs provide further details.

HAR classifier It is used with various video analysis models. Each is designed to process features extracted from video data, typically for action recognition tasks. On the one hand, all the models, except for the SlowFast, employ average pooling and ROIAlign for temporal and spatial feature pooling, respectively. They also include max-pooling operations. A dropout layer, usually with a 0.5 dropout rate, is applied for regularization. The projection layer reduces the feature dimension to 400, followed by a softmax activation for classification. On the other hand, SlowFast includes both slow and fast pathways for feature processing. It utilizes two sets of temporal and spatial pooling operations, one with a kernel size of 8 and the other with a kernel size of 32. ROIAlign, max-pooling, dropout, and dimension reduction are applied to features from both pathways. The projection layer reduces the feature dimension to 400 before softmax activation for classification. In Slow and SlowFast architectures, the temporal pooling is performed using a larger kernel size (8) than the others (4). These classifiers are part of the component in video action recognition models. Their backbones process spatial and temporal features extracted from video clips and prepare them for classification tasks. In practice, all HAR models share common classifier properties. Specifically, the classifier component highlighted in blue within the upper pipeline in Fig. 2 maintains a consistent architecture across all HAR models, except for the SlowFast model, as elaborated in this section. While the architectural structure remains the same, it is crucial to emphasize that the training weights for each model's classifier differ. Each model's classifier was trained in conjunction with its respective backbone on the mentioned Kinetics 400 dataset.

The re-ID embeddings generator As illustrated in Fig. 2 with the highlighted portions in red, the processing of action embeddings involves a sequence of transformations. Specifically, these embeddings pass through two transformer encoders, followed by a global 1D average pooling layer, and finally by two fully connected layers. A more detailed depiction of the employed transformer encoder can be found in the lower section of Fig. 2. This transformer architecture is applied to a feature vector, as introduced by Vaswani et al. [41]. Subsequently, the global 1D average pooling layer condenses the output tensor from the Transformer encoder into a feature vector for each data point within the current batch. Following this, a fully connected layer extracts the pertinent features, and its output is the input for the ultimate dense layer in the classification head.

Transformers excel in capturing extensive dependencies and contextual information across sequences of variable lengths. Contrary to convolutional methods [15], Transformers inherently possess positional sensitivity, acknowledging the significance of the order of elements within sequences-an essential factor in tasks like the analysis of actions or features spanning multiple frames, and in particular in the analysis of behavioral biometric traits. Their flexibility in accommodating diverse sequence lengths, the ability to construct hierarchical feature representations, and the capacity to mitigate overfitting collectively contribute to the robust modeling of intricate spatial and temporal relationships.

The resultant embeddings are purposefully crafted to distinguish between different identities and are compelled to inhabit a d-dimensional hypersphere via L2 normalization of the final output. The computation of the distance between HAR embeddings employs the L2 distance function, proposed initially by Schroff et al. [36]. This function calculates the squared difference between feature vectors ($u_!$ and u_2) and aggregates these differences (see (2)). This distance metric is relevant in computing the loss function, which relies on the distances between embeddings of multiple input samples, contingent on the selected loss function. In this context, we have evaluated the triplet loss, which entails a comparison of three samples: an anchor sample, a positive sample (with the same identity as the anchor), and a negative sample (representing a distinct identity from the anchor) [36]. The formal definition of the triplet loss function is provided below:

$$L_{triplet}(D_1, D_2) = \max(D_1^2 - D_2^2 + m_1, 0) \tag{2}$$

Here, D_1 represents the distance between the anchor and positive samples, and D_2 represents the distance between the anchor and negative samples. The margin parameter is denoted as m_1. The objective is to minimize the distance between the anchor and positive samples while maximizing the distance between the anchor and negative samples.

4 Experimental setup

4.1 Dataset

The experiments reported in this paper were performed on a portion of the dataset initially presented by Penate et al. [27]. This dataset was gathered during the Transgrancanaria (TGC) 2020 ultra-distance running competition, encompassing multiple race distances for participants to undertake, totaling up to six variations.

Traditionally, HAR benchmark datasets, while intricate, aim to replicate real-life human activities within various scenarios. Their primary objective is to represent human behavior as faithfully as possible in diverse settings. Therefore, the paramount consideration in evaluating a dataset lies in its fidelity to reality, as a close alignment significantly enhances HAR. In real everyday activities scenes, it is normal to encounter considerable variations in illumination, scene characteristics, occlusions, and background activities. However, it is noteworthy that several existing datasets do not prioritize addressing these challenges and are instead recorded in controlled environments [33]. For this reason, the present study exploits that annotated dataset collected during the TGC (Trans Grand Canaria) Classic race, where runners are tasked with completing a grueling 128-kilometer course within a 30-hour time frame. Using the chosen ultra-distance running competition dataset for the experiment reported here provides several compelling advantages, since it inherently encapsulates the intricate and dynamic aspects of real-world physical activity. These competitions develop in uncontrolled, natural environments, showcasing genuine diversity in light conditions, terrain variations, and participant interactions. Furthermore, unlike numerous actor-centric HAR datasets, ultra-distance running competitions feature activities performed by many individuals, making them a closer reflection of real-world scenarios. Last but not least, our selection of this dataset is strategic as it can be employed for both HAR and re-ID tasks, enhancing its versatility and relevance to our research objectives.

In our domain-related task, we aim not to scrutinize all 400 actions that HAR models can recognize; instead, we aim to focus on a single action-specifically, jogging-and evaluate the classifiers' performance in this subtask. In our investigation, jogging serves as the reference action, representing the most relevant and widespread activity within the sporting environment under examination. Metrics such as jogging frequency and confidence are utilized to assess both the prevalence of jogging actions and the accuracy of the classifiers in identifying this specific activity.

In the context of our domain-unrelated task, the original dataset includes annotations for nearly 600 participants across six RPs. The experimental setup described in this paper primarily focused on the final three RPs. To enhance readability, these RPs are referred to as RP1, RP2, and RP3, which correspond to RP3, RP4, and RP5 in the original dataset. These RPs encompass data recorded beyond the 84-kilometer mark. The unused RPs in the original dataset align with nighttime captures because the race typically commences around midnight. In the present experiments, we opted not to factor in this additional covariate to better concentrate on context and encoding. Figure 4 illustrates samples of these three selected RPs. It is possible to observe that RP1 captures the runners navigating through an environment characterized by rocky terrain (see upper left image in Fig. 4), RP2 captures images with a road parapet on the background (middle left image in Fig. 4), and finally, RP3 portrays runners traversing a conventional road (bottom left image in Fig. 4). The mentioned RP selection allowed evaluating the performance of the considered models in the latter phases of the race, involving 214 runners who are the ones considered for the presented experiments. Having three RPs, it is possible to consider two race stages: stage 1 from RP1 to RP2, and stage 2 from RP2 to RP3. Stages allow getting valuable insight into video sequences. Notably, 111 runners appear in RP1, RP2, and RP3, consistently contributing to 333 videos allowing to analyze transitions across stage 1 and across stage 2. Further 18 runners only appear in RP2 and RP3 (transition across stage 2, 36 videos) while one runner appears in RP1 and RP2 (transition across stage 1, 2 videos). Additionally, a single runner is present in both RP1 and RP3 (2 videos). It is worth pointing out that this specific runner is not considered for sequence transitions for the re-ID task described in the following, but only as a negative instance for the triplet loss metric. Lastly, 83 runners only appear in a single RP video, either RP1, RP2, or RP3, providing 83 videos. Also these samples are negative instances for the triplet loss metric in the re-ID task evaluation. In summary, the analysis involved all 214 runners, collectively contributing to the comprehensive dataset of 456 videos, and it is noteworthy that these figures align with the counts reported by the original paper [27].

4.2 Experimental scenario

Figure 4 illustrates the two analyses conducted in this study: the context analysis and the encoding analysis.

Context analysis This analysis aims to emphasize the pivotal role of context information. The experiments investigate predictions from seven cutting-edge action recognition models across various context scenarios. The goal is to discern how different levels of contextual awareness impact the prediction accuracy, guiding the analysis from the runner's immediate Region of Interest (ROI) to the broader context of the entire footage, as shown in the left column of Fig. 4.

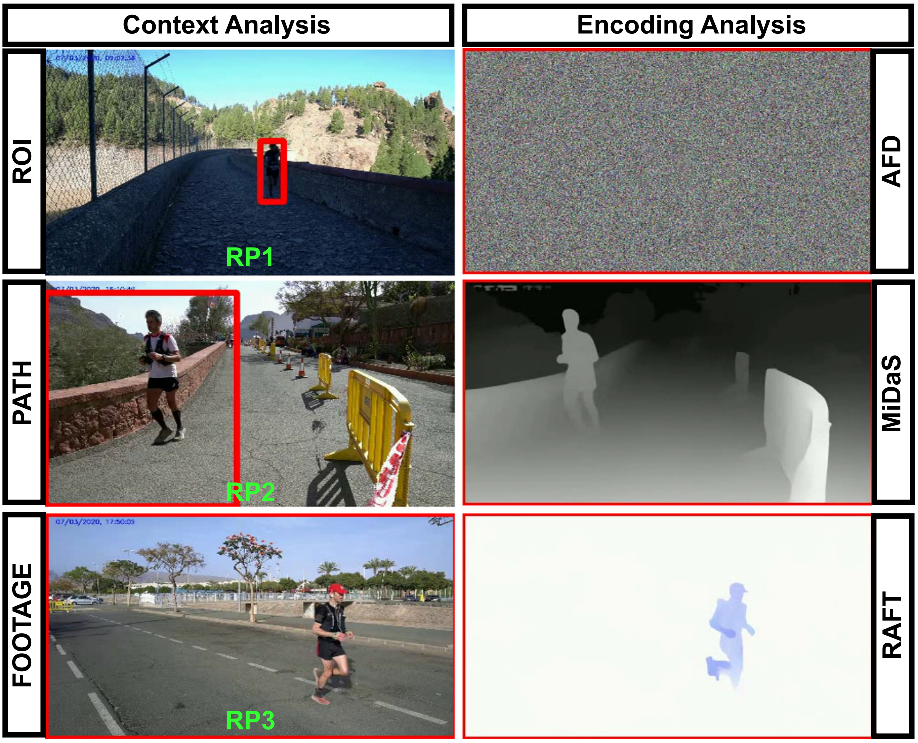

Fig. 4 Experimental scenario. The figure summarizes the visual characteristics of the input video clips or portions of clips used in each experiment. The images correspond to frames captured from each considered RP after the prep-processing steps detailed in Section 3. All three RPs appear in both experimental setups. The left column shows the context experiment configuration, where the input for the HAR model is highlighted in red and either comprises the runner's Region of Interest (ROI), the runner's PATH, or the complete footage. The right column presents the variations of the video encoding inputs used in the second experiment: MiDaS, RAFT, and AFD, respectively. These video encoding inputs are placed alongside with the traditional RGB input for comparison. In this second experiment, the complete footage is always used, as underlined by the red highlighting

The context analysis begins with exploring the restrictions associated with contextual information. Initially, this examination centers exclusively on scrutinizing predictions generated from the runner's ROI. This approach offers valuable insights into the immediate environment surrounding the runner.

Continuing our exploration, the attention shifts to a comprehensive exploration of the runner's trajectory. Predictions stemming from the runner's Path Region of Interest (PATH) provide an intricate insight into actions occurring along the route. PATH is computed from the sequence of locations traversed by the runner's BB (Bounding Box) in a pre-processed video, i.e., from the sequence of BB coordinates. In coding terms, it encompasses the entire area defined by the coordinates between (minROI.x, minROI.y) and (maxROI.x, maxROI.y). Here, minROI.x and maxROI.x, as well as minROI.y and maxROI.y, represent the minimum and maximum x and y coordinates traversed by the moving BBs.

The exploration culminates in thoroughly examining the entire visual scene, the complete footage. Here, the analysis considers all video elements, providing the richest context information. This approach encompasses the runner's actions within the broader context, considering interactions with the environment, potential obstacles, and intricate component dynamics. Predictions from this comprehensive analysis yield a profound understanding of the runner's actions and interactions, offering a comprehensive view of their dynamic engagement with the environment.

Encoding analysis The experiments investigated three SOTA approaches for video encoding, comparing them to the standard RGB input: AFD, MiDaS, and RAFT.

AFD, designed for scenarios like camouflage detection, showcased the models' capability to function without static appearance information [19]. AFD was constructed from motion data extracted from the UCF101 dataset through advanced optical flow estimation. This approach emphasizes the importance of temporal encoding over static details. It highlights the interpretability of the exploited models, underlining that optical flow or similar representations are not intrinsic to the tested networks [19]. This encoding method proves valuable when individuals' identities must remain concealed to protect their privacy. Despite being a relatively recent development, AFD has already found applications in video self-supervised learning [35].

MiDaS, developed for obtaining accurate depth data at scale, calculates relative inverse depth from individual images. It encompasses a range of models, including compact, high-speed options to highly accurate, large-scale variants [29]. These models were trained on ten diverse datasets, utilizing multi-objective optimization to ensure exceptional performance across a wide range of input scenarios. Additionally, invariant loss functions were employed to address compatibility challenges between datasets, enabling training with data from diverse sensing modalities [29]. In recent years, this encoding technique has seen extensive adoption in various domains, including monocular depth estimation [49], image segmentation [32], and video anomaly detection [1] that can be used in turn to fed video-surveillance, HAR, and re-ID.

Finally, the evaluation of input encoding for action recognition models includes RAFT, short for Recurrent All-Pairs Field Transforms. It introduces an innovative deep network architecture for optical flow estimation [39]. RAFT's unique approach comprises three key components: a feature encoder, a correlation layer, and a recurrent GRU-based update operator. The feature encoder extracts feature vectors for each pixel, the correlation layer generates a 4D correlation volume for pixel pairs, and the recurrent operator iteratively refines the flow field [39]. RAFT excels in accuracy across diverse datasets, exhibits robust cross-dataset generalization, and impresses with computational efficiency [39]. While many HAR techniques incorporate flow information to enhance their robustness, it is worth noting that this encoding method has also gained widespread utilization across different domains, such as motion aggregation [20] and ball trajectory prediction [26] whose applications intersect HAR and re-ID.

In order to ensure a fair comparison, all the results presented here are derived from the training and testing conducted according to the same specified procedure, as detailed in Section 4.3. Additionally, given the intended focus on examining architectural behavior in the context of video encoding, we have refrained from retraining the backbones, primarily due to the absence of resources like AFD, MiDaS, or RAFT for Kinetics. This is reasonable since our aim is not to showcase the best performance method but to highlight the performance differences among various networks when exposed exclusively to dynamic information and to explore their correlations with domain-specific and non-domain-specific tasks.

4.3 Metrics

This section describes the protocol used for both tasks in the reported study to evaluate the performance of various HAR and re-ID models on a large-scale dataset. The evaluation metrics differ due to the nature of the experiments concerning different kinds of tasks.

Domain-related task Concerning HAR classification, the experiments involve using pretrained full model, which includes one of the backbones described in Section 3.1, along with the classifier components. Since the task develops in a related domain, there is no need for further training, and it is possible to perform inference on the specified inputs. The analysis focusing on action recognition requires utilizing various statistical measures to glean insights from a dataset. These measures include the mode, mode frequency, jogging frequency, and jogging confidence, being the latter two related to the dataset at hand. Among these measures, one prominent indicator is the mode, representing the action that emerges with the highest frequency within the dataset. In essence, the mode is the action that appears most frequently, serving as a central reference point for understanding the prevailing actions observed across videos. Furthermore, the mode frequency (denoted as *Mode Freq.*) provides a quantitative assessment of the prevalence of the mode within the entire dataset. Expressed as a percentage, it denotes the proportion of videos in which the mode action is present. This measure offers a vital glimpse into the distribution of actions, highlighting the dominance of the most frequent action and its significance within the analysis context. In the present exploration, the action of "jogging" represents the baseline, deeming it the most fitting action for the sporting environment under consideration. Consequently, the jogging frequency (denoted as *Jogg. Freq.*) assumes significance. This metric signifies the percentage of videos in which the jogging action is detected. As the baseline action, jogging frequency provides a benchmark against which other actions can be compared, facilitating the assessment of the prevalence of actions beyond the mode. Jogging confidence (denoted as *Jogg. Conf.*) is another statistical metric contributing to this analysis. This measure represents the mean confidence level each classifier assigns to the jogging action. It quantifies the classifiers' certainty in identifying jogging actions within the dataset. This parameter offers valuable insights into the accuracy and reliability of the classifiers in recognizing the jogging action, shedding light on their performance and the overall consistency of their predictions. In summary, the analysis involves integrating these statistical measures to better understand the actions in the dataset. By assessing these parameters, we gain a comprehensive view of the prevalence, significance, and reliability of specific actions within the sporting environment, ultimately enhancing our ability to interpret the dynamics of action recognition in the context of interest. In particular, the jogging confidence score was used to compute the correlation with the performance achieved for the second task.

Domain-unrelated task In the context of re-ID classification, the chosen evaluation metric is Mean Average Precision (mAP), which is widely employed for re-ID tasks. mAP is derived by first calculating the Average Precision (AP) individually for each class and subsequently determining the mean of these AP values across all classes. AP is the area beneath the Precision-Recall curve (PR curve) corresponding to each class.

Let N denote the total number of classes (i.e., identities). The average precision for class i AP_i can be computed as:

$$AP_i = \frac{1}{R}\sum_{k=1}^{R} P_i(k) \cdot rel(k) \tag{3}$$

where R is the total number of ground truth positives, $P(k)$ is the precision at cutoff k and $rel(k)$ is a relevance function. This relevance function is an indicator function with a value of 1 if the ID at rank k is relevant and a value of 0 otherwise.

The mAP can be calculated using the following equation:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

Since there is no inherent connection between the training (HAR) and testing (re-ID) tasks, it becomes necessary to train the classifier using the backbone-returned embeddings to adapt it to the characteristics of the new domain. To ensure the robustness of findings, the experiments followed a 10-fold cross-validation strategy for training the classifier. This method partitions the dataset into ten equal-sized folds, each containing an equivalent number of samples. Within each fold, one subset of the data serves as the test set, while the remaining nine act as the training set. This process is iterated ten times, each subset serving as the test set once. The performance metric is averaged across all folds, yielding the final evaluation scores. In addition, due to the often incompatible characteristics of the produced embeddings, the cross-folding was performed separately for each backbone. It is worth pointing out that, for instance, the SlowFast-crossfold-trained classifier cannot be directly applied to test on C2D-produced embeddings. Although this 10-fold cross-validation approach has limitations due to the dataset's constrained number of test samples per fold-resulting from its inherent characteristics- it remains a valuable evaluation protocol to assess the re-ID model.

The experiments with the re-ID task will exploit the introduced concept of stage in handling the used dataset videos. Each race stage provides 11 to 12 test runners per fold. In detail the set of 112 runners appearing in stage 1 divided in 10 folds contributes 11 test runners per fold (in the approximation the training set is privileged), while the set of 120 runners appearing in stage 2 divided in 10 folds contributes 12 test runners per fold. The training set comprises the images of the remaining 202 or 201 runners, which amounts to 202 images when considering a total of 214 runners. Because the data split is based on individual runner IDs, we ensure that the training and testing sets do not include the same runner to maintain our evaluation's integrity. The 111 positive pairs for the first stage and 18 for the second stage, for 129 positive video pairs, may not appear sufficient for generalizable results. Despite this, the 10-fold cross-validation strategy comprehensively assesses the model's performance on diverse dataset subsets, ensuring unbiased evaluation results [30]. In addition, this 10-fold cross-validation process was repeated ten times, with different random data splits into folds. This repetition helps mitigate the impact of data variability and provides a more stable estimate of the model's performance.

Domain shift correlation *(r)* Computing the correlation between both task metrics, the mAP and the jogging confidence scores, is a valuable analytical step in assessing the models' overall performance and reliability. As aforementioned, the mAP provides an insightful measure of the model's ability to accurately rank and retrieve relevant images, thus capturing the quality of predictions. On the other hand, using output confidence scores measures the models' certainty in their predictions. It is possible to better understand domain shift by calculating the correlation between these two metrics. A positive correlation suggests that the models tend to have higher confidence in their predictions as the mAP improves (indicating better ranking performance). Conversely, a negative correlation might indicate situations where high confidence scores are associated with lower mAP, highlighting cases where the models may be overconfident but less accurate. This analysis can help to identify scenarios where the models' predictions align or diverge.

5 Experiments and results

5.1 Context analysis

Domain-related experiment Tables 1, 2, and 3 report the experimental results achieved at the respective RP. These are in terms of the statistical measures introduced in Section 4.3 for HAR classification using the backbones described in Section 3.1. Each table further encompasses the different extents of contextual information entailed by the setup for this task. The results reveal profound insights into the dynamics of the presented HAR analysis. This considers the trajectory followed by the runner and the dynamic elements that encompass it within the frame. By subjecting these statistical measures to scrutiny, we thoroughly comprehend the predictive behavior of HAR models across diverse contexts. Central to this analysis is the careful examination of how the trajectory of the runner and the contextual elements converge to shape the predictions made by the HAR models.

Each RP contributes a distinct part to the analysis. For instance, since RP1 captures the runners navigating through an environment characterized by rocky terrain (see upper left image in Fig. 4), this sets the stage for a unique pattern of predictions. Conversely, RP3 portrays runners traversing a conventional road, presenting a different layer of complexity in the predictive dynamics (bottom left image in Fig. 4). The runners' trajectory further accentuates the intricacies of the predictions derived from the HAR models. In RP2, the runners' trajectory moves from the upper segment of the frame to the lower portion, a scenario that inherently challenges the predictive accuracy of the models. On the contrary, RP3 demonstrates runners moving horizontally from left to right, manifesting a trajectory that is fundamentally more intuitive and aligns with conventional expectations (refer to the images in the left column of Fig. 4). These disparate trajectories across different RPs contribute to the diversification of predictions, providing invaluable insights into how HAR models interpret varying motion patterns and adapt their predictions accordingly.

The tables testify that a ROI limited to the runner bounding box fails to activate the relevant jogging action at any RP. Unlike specific actions like smoking, jogging inherently requires a broader context for comprehension. Running implies movement within a space, and as a result, a ROI limited to the runner's body does not account for this contextual aspect, leading to action triggers that are dependent on the specific gestures of the runners. Tables 1, 2, and 3 for the respective RPs demonstrate that for most classifiers, the actions detected are flying a kite, motorcycling, and checking tires, respectively.

Including the PATH significantly expands the scope of the ROI, offering classifiers a more comprehensive context. In the case of RP1, characterized by a rugged terrain and a non-horizontal trajectory from top to bottom, most classifiers struggle to accurately identify jogging as the action when the PATH is considered. Instead, rock climbing emerges as the predominantly selected action across various HAR classifiers, as evident in Table 1. Conversely, the scenarios in RP2 and RP3 prove more intelligible to the HAR models, with jogging being the most frequently inferred action. Just as when considering only the ROI, contextual factors influence the classifier's decision-making. For instance, RP3 focuses on a paved road, leading PATH to trigger the recognition of alternative actions such as motorcycling and pushing a car, as Table 3 shows.

Lastly, the model gains a comprehensive contextual perspective when analyzing the FOOTAGE. This is evident in Tables 1, 2, and 3, where jogging emerges as the predominant action. Furthermore, there is a notable increase in jogging confidence for each classifier compared to the consideration of the PATH alone. The sole exception is the I3D NLN classifier when applied to RP2, which struggles to identify this action across all contexts accurately.

Table 1 HAR models inference on RP1 for context analysis

	Stats.	C2D	I3D	I3D NLN	S4x16	S8x8	SF4x16	SF8x8
ROI	Mode	Flying Kite	Flying Kite	Beatbox.	Clean. Gutters	Rock Climb.	Flying Kite	Throw. Axe
	Mode Freq.	54.2%	36.0%	46.3%	40.9%	44.3%	46.3%	45.8%
	Jogg. Freq.	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%
	Jogg. Conf.	0.4%	0.5%	0.5%	0.5%	0.7%	0.5%	0.5%
PATH	Mode	Rock Climb.	Rock Climb.	Rock Climb.	Rock Climb.	Clean. Gutters	Hit. Base.	Throw. Axe
	Mode Freq.	23.6%	43.3%	38.9%	23.2%	43.3%	22.7%	47.3%
	Jogg. Freq.	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	3.9%
	Jogg. Conf.	0.6%	0.8%	3.3%	1.7%	3.7%	1.9%	1.3%
FOOTAGE	Mode	Jogg.	Jogg.	Jogg.	Jogg.	Jogg.	Jogg.	Jogg.
	Mode Freq.	96.6%	91.6%	51.2%	73.9%	93.6%	76.8%	82.8%
	Jogg. Freq.	–	–	–	–	–	–	–
	Jogg. Conf.	80.7%	73.9%	43.9%	56.4%	76.5%	60.1%	61.0%

Every horizontal block represents a context configuration; within these configurations, each column reports the value of a specific variable for each assessed model. The variables under analysis include the mode, mode frequency (labeled as “Mode Freq.”), jogging frequency (labeled as “Jogg. Freq.”), and jogging confidence (labeled as “Jogg. Conf.”)

Table 2 HAR models inference on RP2 for context analysis

	Stats.	C2D	I3D	I3D NLN	S4x16	S8x8	SF4x16	SF8x8
ROI	Mode	Motorcy.	Ridyng Unicy.	Archery	Bee Keep.	Blow Leaves	Garba. Collec.	Blow Leaves
	Mode Freq.	28.8%	13.7%	16.5%	13.7%	20.1%	26.6%	15.1%
	Jogg. Freq.	0.0%	0.0%	0.0%	0.7%	0.0%	0.7%	2.2%
	Jogg. Conf.	7.4%	6.6%	3.2%	8.5%	0.2%	10.4%	9.0%
PATH	Mode	Jogg.	Jogg.	Faceplant.	Jogg.	Jogg.	Jogg.	Jogg.
	Mode Freq.	55.4%	78.4%	38.8%	77.7%	76.3%	93.5%	89.2%
	Jogg. Freq.	–	–	12.0%	–	–	–	–
	Jogg. Conf.	45.9%	66.4%	12.0%	66.9%	53.5%	79.0%	73.3%
FOOTAGE	Mode	Jogg.	Jogg.	Faceplant.	Jogg.	Jogg.	Jogg.	Jogg.
	Mode Freq.	63.3%	76.3%	30.2%	84.9%	72.7%	96.4%	92.1%
	Jogg. Freq.	–	–	8.6%	–	–	–	–
	Jogg. Conf.	51.3%	69.7%	12.1%	74.2%	52.3%	80.9%	75.0%

Every horizontal block represents a context configuration; within these configurations, each column reports the value of a specific variable for each assessed model. The variables under analysis include the mode, mode frequency (labeled as "Mode Freq."), jogging frequency (labeled as "Jogg. Freq."), and jogging confidence (labeled as "Jogg. Conf.")

Table 3 HAR models inference on RP3 for context analysis

	Stats.	C2D	I3D	I3D NLN	S4x16	S8x8	SF4x16	SF8x8
ROI	Mode	Motorcy.	Hopscot.	Hopscot.	Check. Tires	Check. Tires	Texting	Check. Tires
	Mode Freq.	31.6%	46.4%	50.0%	25.4%	35.1%	53.5%	63.2%
	Jogg. Freq.	1.8%	4.4%	5.3%	1.8%	1.8%	8.8%	17.5%
	Jogg. Conf.	2.5%	6.7%	9.3%	3.9%	6.2%	11.7%	19.3%
PATH	Mode	Motorcy.	Jogg.	Jogg.	Push. Car	Jogg.	Jogg.	Jogg.
	Mode Freq.	36.0%	43.9%	28.1%	44.7%	32.5%	55.3%	70.2%
	Jogg. Freq.	16.7%	–	–	28.1%	–	–	–
	Jogg. Conf.	26.4%	43.3%	26.7%	35.6%	33.0%	43.5%	63.0%
FOOTAGE	Mode	Jogg.	Jogg.	Jogg.	Jogg.	Jogg.	Jogg.	Jogg.
	Mode Freq.	85.1%	90.4%	48.2%	78.1%	94.7%	97.4%	99.1%
	Jogg. Freq.	–	–	–	–	–	–	–
	Jogg. Conf.	77.0%	81.8%	42.4%	63.5%	77.6%	89.4%	93.8%

Every horizontal block represents a context configuration; within these configurations, each column reports the value of a specific variable for each assessed model. The variables under analysis include the mode, mode frequency (labeled as “Mode Freq.”), jogging frequency (labeled as “Jogg. Freq.”), and jogging confidence (labeled as “Jogg. Conf.”)

Domain-unrelated experiment Table 4 offers a comprehensive summary of the performance exhibited by the evaluated backbone models starting from various contexts. Each sub-table encompasses seven rows corresponding to the HAR backbones, as detailed in Section 3.1. Each row presents the mAP values for the triplet loss, introduced in Section 4.3. The first column lists the backbone models, while the second column specifies the number of frames the model utilizes for generating HAR embeddings. It is important to emphasize that each transformer classifier was trained with its respective backbone using default settings, including the number of frames per prediction, sampling rate, backbone depth, and more. Maintaining these settings is crucial, as they are those benchmarked against the Kinetics dataset. In the following columns, the mAP values for each model can be seen at the two different race stages under consideration to tackle this problem [15]. The notation is as fol-

Table 4 mAPs achieved by each considered context experiment

Backbone	#Frames	mAP		
		RP1→RP2	RP2→RP3	Average
ROI				
C2D	8	60.1%	63.4%	61.7%
I3D	8	62.7%	64.8%	63.7%
I3D NLN	8	63.7%	70.5%	**67.1%**
S4x16	4	**64.4%**	66.4%	65.4%
S8x8	8	62.3%	66.6%	64.4%
SF4x16	32	62.7%	68.0%	65.3%
SF8x8	32	62.3%	**71.3%**	66.8%
RPs Average		63.7%	66.6%	65.1%
PATH				
C2D	8	59.2%	67.2%	63.2%
I3D	8	61.3%	66.1%	63.7%
I3D NLN	8	64.7%	70.6%	67.7%
S4x16	4	**68.5%**	69.7%	**69.1%**
S8x8	8	62.3%	66.6%	64.4%
SF4x16	32	66.0%	67.8%	66.9%
SF8x8	32	62.3%	**71.3%**	66.8%
RPs Average		63.4%	68.5%	66.0%
FOOTAGE				
C2D	8	63.8%	73.4%	68.6%
I3D	8	66.7%	**80.7%**	73.7%
I3D NLN	8	71.0%	74.4%	72.7%
S4x16	4	**73.6%**	75.3%	**74.4%**
S8x8	8	68.9%	76.2%	72.6%
SF4x16	32	71.2%	72.8%	71.9%
SF8x8	32	70.4%	73.6%	72.0%
RPs Average		69.4%	75.2%	72.3%

The tables are organized according to the backbones. The second column shows the number of frames the backbone requires to make a prediction (see Section 3.1). Moreover, two competition stages are analyzed using scheme A→B, where A stands as the RP considered for the gallery (reference video) and B as the probe. The average columns show the mean mAP for each backbone, whereas the last row shows the mean mAP at each stage lows: the first term before the → symbol represents the gallery, and the second term after the → symbol represents the probe. Additionally, the table includes the average mAP calculated across these reference stages.

The results underline that a broader context also positively influences the re-ID performance in a significant way. Specifically, a more extensive context consistently leads to enhanced performance. In the initial race stage, RP1 to RP2, the Slow 4x16 model consistently outperforms all the other HAR models. During the subsequent race stage, RP2 to RP3, the SlowFast 8x8 model excels in a constrained context. However, it is surpassed by the Slow models (S8x8 and S4x16) and the I3D ConvNet when a richer context is considered. The presented table also reveals that the number of frames incorporated into the HAR model does not substantially influence the model's performance. The SlowFast (SF) models, which utilize more frames, perform worse than the Slow (S) models. Consequently, the model's architectural design, rather than the quantity of input frames, plays a more relevant role in determining performance outcomes.

Finally, the last row in the table presents the mAP values, averaged across all models for each reference stage. These mAP values shed light on the performance of the athlete re-ID system, revealing that the system performs more effectively during the second stage (RP2 to RP3) than the initial stage (RP1 to RP2). Moreover, it is noteworthy that employing a more extensive context input (FOOTAGE) results in a 7.3% higher mAP than using the ROI input. These findings suggest that the system accurately identifies athletes as they progress through the race and performs better when provided with more input information. In the context of the SOTA comparison, the proposed transformer approach demonstrates a remarkable 11.1% enhancement compared to the results reported in [15]. This mentioned investigation relied on a classifier featuring two 512 dense layers separated by a batch normalization layer. The findings presented there identified the Slow4x16 backbone as the top performer, with a 63.3% mAP. In contrast, the transformer-based classifier presented here fed by the Slow4x16 (S4x16) backbone achieves a significantly higher mAP of 74.4%. It is worth pointing out that, as stated before, the aim of the present study is not to showcase the method for achieving the best performance. Instead, the aim is to highlight the performance differences among various networks when exposed exclusively to dynamic information and to explore the correlation that they are able to maintain between domain-specific and non-domain-specific tasks. Therefore, further comparisons would not be significant since, to the best of our knowledge, no other re-ID approach incorporates HAR techniques. Using seven different backbones for the analysis should provide a comprehensive basis for comparison.

5.2 Encoding analysis

Domain-related experiment Similarly to the context analysis, Tables 5, 6, and 7 delve into the statistical data at each RP, offering insights into the intricacies of the performed HAR analysis. This group of experiments exclusively relies on the FOOTAGE configuration, which was encoded using AFD, MiDaS, and RAFT. This choice can be reasonably motivated by the remarkable performance of the FOOTAGE configuration, surpassing all other configurations, particularly when evaluating the jogging class (the shared analysis baseline).

What is particularly intriguing is that the results in this section exhibit a stark departure from those in the previous section. Curiously, no jogging action emerges as the predominant pattern among video clips at any RP, regardless of the classifier or encoding method employed. This observation introduces a new perspective to the analysis. This further enhances the overall comprehension of the underlying dynamics within the domain-adaptation context explored in this paper.

Table 5 HAR models inference on RP1 for encoding analysis

	Stats.	C2D	I3D	I3D NLN	S4x16	S8x8	SF4x16	SF8x8
AFD	Mode	Blow Leaves	Dunk. Basket.	Tying Knot	Sing.	Jugg. Balls	Jugg. Balls	Blow Leaves
	Mode Freq.	71.4%	50.7%	55.2%	88.2%	86.2%	58.6%	82.3%
	Jogg. Freq.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%
	Jogg. Conf.	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	7.2%
MiDaS	Mode	Snow-board.	Tying Knot	Sharp. Pencil	Weld.	Tying Knot	Writing	Drawing
	Mode Freq.	51.7%	95.1%	35.5%	22.7%	42.9%	33.5%	59.6%
	Jogg. Freq.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%
	Jogg. Conf.	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	2.8%
RAFT	Mode	Flying Kite	Snowkit.	Strech. Leg	Flying Kite	Flying Kite	Flying Kite	Walking Dog
	Mode Freq.	79.2%	90.6%	54.0%	76.2%	88.1%	82.7%	22.8%
	Jogg. Freq.	0.0%	0.5%	0.5%	3.0%	2.0%	0.0	10.9%
	Jogg. Conf.	2.5%	6.5%	11.1%	12.1%	9.8%	2.5	12.7%

Every block represents a video encoding scheme applied to the FOOTAGE context; within these configurations, each table column reports the values of specific variables for one of the models. The variables under analysis include the mode, mode frequency (labeled as "Mode Freq."), jogging frequency (labeled as "Jogg. Freq."), and jogging confidence (labeled as "Jogg. Conf.")

Table 6 HAR models inference on RP2 for encoding analysis

	Stats.	C2D	I3D	I3D NLN	S4x16	S8x8	SF4x16	SF8x8
AFD	Mode	Blow Leaves	Bench Press.	Exer. Arm	Sing.	Jugg. Balls	Play. Trump.	Blow Leaves
	Mode Freq.	76.3%	59.0%	61.2%	90.6%	82.7%	58.3%	52.5%
	Jogg. Freq.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Jogg. Conf.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.7%
MiDaS	Mode	Tying Bow	Bend. Metal	Tying Bow	Fold. Napk.	Smok.	Ripp. Paper	Fold. Paper
	Mode Freq.	31.7%	28.8%	67.6%	86.3%	31.7%	49.6%	27.3%
	Jogg. Freq.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%
	Jogg. Conf.	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	5.4%
RAFT	Mode	Flying Kite	Strech. Arm	Danc. Ballet	Danc. Ballet	Sharp. Pencil	Danc. Ballet	Strech. Arm
	Mode Freq.	85.4%	41.6%	47.4%	57.7%	43.8%	47.4%	43.8%
	Jogg. Freq.	0.0%	6.6%	0.0%	10.2%	13.9%	2.2%	16.1%
	Jogg. Conf.	5.9%	9.3%	3.8%	15.1%	14.4%	5.8%	17.9%

Every block represents a video encoding scheme applied to the FOOTAGE context; within these configurations, each table column reports the values of specific variables for one of the models. The variables under analysis include the mode, mode frequency (labeled as “Mode Freq.”), jogging frequency (labeled as “Jogg. Freq.”), and jogging confidence (labeled as “Jogg. Conf.”)

Table 7 HAR models inference on RP3 for encoding analysis

	Stats.	C2D	I3D	I3D NLN	S4x16	S8x8	SF4x16	SF8x8
AFD	Mode	Blow Leaves	Bench Press.	Tying Knot	Sing.	Jugg. Balls	Play. Trump.	Blow Leaves
	Mode Freq.	78.9%	61.4%	55.3%	85.1%	85.1%	62.3%	78.9%
	Jogg. Freq.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Jogg. Conf.	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.9%
MiDaS	Mode	Danc. Ballet	Mak. Snowm.	Side Kick	Tai Chi	Mak. Snowm.	Writting	Strech. Arm
	Mode Freq.	51.8%	50.9%	80.7%	44.7%	25.4%	36.0%	70.2%
	Jogg. Freq.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	11.4%
	Jogg. Conf.	0.2%	0.0%	0.3%	0.0%	0.1%	0.0%	10.7%
RAFT	Mode	Flying Kite	Snow-kitt.	Danc. Ballet	Flying Kite	Flying Kite	Flying Kite	Walking Dog
	Mode Freq.	94.6%	26.8%	65.2%	40.2%	58.0%	87.5%	33.0%
	Jogg. Freq.	0.0%	24.1%	11.6%	8.9%	28.6%	3.6%	17.9%
	Jogg. Conf.	6.1%	12.6%	13.9%	14.4%	18.5%	6.9%	15.5%

Every block represents a video encoding scheme applied to the FOOTAGE context; within these configurations, each table column reports the values of specific variables for one of the models. The variables under analysis include the mode, mode frequency (labeled as “Mode Freq.”), jogging frequency (labeled as “Jogg. Freq.”), and jogging confidence (labeled as “Jogg. Conf.”)

Among the various encoding approaches, AFD stands out as the one that most significantly alters the predictions based on training RGB images. Interestingly, the I3D classifier is the model that shifts its mode value prediction at each RP. In contrast, the C2D, S8x8, and SF4x16 models maintain a consistent mode value throughout all RPs. What is particularly interesting is that the inferred actions, such as "blow leaves", exhibit no discernible connection to jogging. This represents a distinct scenario where the backbone models fail to activate the relevant signals for jogging. Intriguingly, the confidence in jogging detection is lightly pronounced only in the case of SF8x8 at RP2.

MiDaS yields results comparable to those obtained with AFD. The depth inputs, however, prove to be excessively noisy for the backbone models. As evident in Fig. 4, the encoding approaches in this section strip away a significant amount of contextual information. As expected, this is consistently reflected in the outcomes. Like AFD, MiDaS does not succeed in identifying jogging as the dominant action (mode) for any classifier. Nevertheless, despite the backbones falling short, the mode frequency is lower than that observed with AFD, indicating higher uncertainty in the selected actions. In contrast to AFD, there is one backbone that consistently triggers jogging: SF8x8. Its jogging confidence increases from 2.8% at RP1 to 5.4% at RP2 and 10.7% at RP3. This observation is interesting, particularly considering that, in Section 5.1, the baseline jogging achieves higher confidence as the RP moves further in the FOOTAGE context.

Finally, it is worth noting that RAFT consistently delivers superior jogging confidence results. This outcome aligns with expectations since these models are designed to handle flow information (as explained in Section 3.1). Interestingly, "Flying Kite" emerges as the action mode most frequently obtained by these models. However, in the absence of scene information and relying solely on the runner's flow, the model's inference can occasionally lead to confusion. Notably, during the context experiments with ROI as the context, some models also inferred "Flying Kite". In general, both Slow models (S8x8 and S4x16) consistently outperform others in jogging confidence across all three RPs. Furthermore, jogging ranks as the third most frequent action for both Slow models at RP1 (refer to Table 5), the third most frequent action for S8X8 and SF8x8 at RP2 (refer to Table 6), and the second most frequent action for I3D, I3D NLN, and S8x8, with S4x16 and SF8x8 placing third at RP3 (refer to Table 7).

Unrelated-domain experiment Table 8 presents an overview of the performance of the considered backbone models across various encoding information scenarios. For the sake of comparison, we have included the RGB approach in the same FOOTAGE context setting used in Table 4. In fact, all considered encoding methods of this part of the analysis utilize the FOOTAGE context setting. Each sub-table consists of seven rows corresponding to the HAR backbones, as detailed in Section 3.1. The mAP values for the triplet loss are provided within each row. The first column enumerates the backbone models, while the second column specifies the number of frames utilized for generating HAR embeddings. Subsequent columns display the mAP values for each model at two distinct race stages, $RP1 \rightarrow RP2$ and $RP2 \rightarrow RP3$, addressing the problem as outlined in [15]. The average mAP across these reference stages is also included for each model.

In contrast to the context experiment, the encoding information in this analysis exhibits a distinct lack of incremental relation. The selected encoding methods are fundamentally disparate. To elaborate, it is worth reminding that RAFT generates flow video clips where only moving elements are discernible, and MiDaS generates depth video clips, retaining the silhouettes of objects near the camera. AFD produces a more challenging-to-interpret flow-encoded video sequence, even for human observers. The results offer intriguing insights. AFD performs notably poorly in the re-ID task to the extent that, for the first time, the first

Table 8 mAP achieved by each considered encoding experiment

Backbone	#Frames	mAP		
		RP1→RP2	RP2→RP3	Average
AFD				
C2D	8	34.2%	35.6%	34.9%
I3D	8	**40.1%**	32.0%	36.1%
I3D NLN	8	37.3%	35.4%	36.3%
S4x16	4	34.6%	35.7%	35.1%
S8x8	8	29.9%	35.4%	32.7%
SF4x16	32	36.2%	**40.5%**	**38.3%**
SF8x8	32	38.8%	36.9%	37.8%
RPs Average		35.9%	35.9%	35.9%
MiDaS				
C2D	8	63.8%	68.0%	65.9%
I3D	8	60.3%	68.7%	64.5%
I3D NLN	8	**67.3%**	**69.8%**	**68.6%**
S4x16	4	67.2%	69.0%	68.1%
S8x8	8	61.0%	69.7%	65.4%
SF4x16	32	62.4%	67.6%	65.0%
SF8x8	32	61.4%	61.4%	61.4%
RPs Average		63.6%	67.5%	65.6%
RAFT				
C2D	8	**54.4%**	45.9%	**50.1%**
I3D	8	52.3%	46.0%	49.1%
I3D NLN	8	51.2%	45.7%	48.4%
S4x16	4	51.8%	**48.4%**	**50.1%**
S8x8	8	51.2%	45.5%	48.4%
SF4x16	32	44.4%	43.0%	43.7%
SF8x8	32	45.9%	42.3%	44.1%
RPs Average		50.2%	45.3%	47.8%
RGB				
C2D	8	63.8%	73.4%	68.6%
I3D	8	66.7%	**80.7%**	73.7%
I3D NLN	8	71.0%	74.4%	72.7%
S4x16	4	**73.6%**	75.3%	**74.4%**
S8x8	8	68.9%	76.2%	72.6%
SF4x16	32	71.2%	72.8%	71.9%
SF8x8	32	70.4%	73.6%	72.0%
RPs Average		69.4%	75.2%	72.3%

The tables are organized according to backbones. The second column shows the number of frames the backbone requires to make a prediction (see Section 3.1). Moreover, two competition stages are analyzed using scheme A→B, where A stands as the RP considered for the gallery and B as the probe. The average columns show the mean mAP for each backbone, whereas the last row shows the mean mAP at each stage race stage yields superior mAP rates compared to the second stage. This is intriguing as RAFT, another flow-related encoding approach, encounters a similar issue. The generated features for the second race stage are positioned further in the feature space due to the flow encoding approach. Among AFD approaches, SF4x16 fares the best with a 38.3% mAP.

RAFT outperforms AFD, but its mAP generally remains below 50%. Remarkably, all models exhibit strikingly similar performance, which is particularly noteworthy considering that all HAR classifiers in the previous experiment share the same subset of mode actions, namely, "Flying Kite" and "Dancing Ballet". This limited diversity in modes may contribute to the consistent mAP results, suggesting that these models potentially lose crucial information for their predictions. Within the RAFT encoding approach, S4x16 attains the highest accuracy.

Lastly, MiDaS surpasses all other encoding approaches. Furthermore, MiDaS demonstrates competitive results compared to the ROI and PATH context configurations in Table 4. In this encoding setting, the second race stage outperforms the first. The most effective approach uses I3D NLN, a variant of I3D designed to address the covariance shift problem in the batch normalization layer. Additionally, S4x16 reports intriguing mAP rates.

Finally, the last row of the table presents the mAP values averaged across both reference stages for each model. These mAP values provide insights into the performance of the athlete re-ID system, revealing that the system operates more effectively during the second stage (RP2 to RP3) than the initial stage (RP1 to RP2) using the best encoding approach. Notably, when employing the more informative MiDaS encoding, the results are approximately 15% and nearly 30% better than the two flow-based approaches, RAFT and AFD, respectively.

5.3 Domain shift correlation

In the following, the Pearson correlation coefficient will be used to express the correlation between the two distinct tasks, HAR and re-ID, while utilizing seven different HAR backbones. This procedure offers a valuable analytical approach.

The Pearson correlation between two variables X and Y can be computed using the following equation:

$$\rho_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{5}$$

where $cov(X, Y)$ is the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively.

In the first place, Pearson correlation provides a robust and easily interpretable measure of linear association, enabling a quantitative assessment of how changes in performance on the domain-related task of HAR correspond to changes in the unrelated-domain task of athlete re-ID across various backbone models. This allows investigating whether improvements in one task may lead to improvements in the other, shedding light on potential synergies or dependencies between the considered tasks. Furthermore, using a statistical measure like this ensures the reliability and consistency of the assessment, offering a clear and intuitive interpretation of the task correlations.

This can be pivotal for researchers and practitioners seeking to optimize model performance and gain insights into the relationships between distinct but interrelated tasks. Of course, the same analytical approach could be used for other pairs of tasks with similar sharing of underlying features.

Table 9 illustrates the correlation of the tasks in different contexts. Each row corresponds to a specific backbone model. The second column represents the average jogging confidence for each model when considering different contexts: ROI, PATH, and FOOTAGE. The re-ID

Table 9 Context-based analysis of tasks correlation

Backbone	Domain-related task $\mu_{ROI}, \mu_{PATH}, \mu_{FOOTAGE}$	Domain-unrelated task	*Pearson* Correlation *(r)*
C2D	3.4%, 24.3%, 69.7%	61.7%, 63.2%, 68.6%	**99.5%**
I3D	4.6%, 36.8%, 75.2%	63.7%, 63.7%, 73.7%	88.9%
I3D NLN	4.3%, 14.0%, 32.8%	67.1%, 67.7%, 72.7%	97.1%
S4x16	4.3%, 34.7%, 65.0%	65.4%, 69.1%, 74.4%	**99.5%**
S8x8	2.4%, 30.1%, 68.8%	64.4%, 64.4%, 72.6%	90.1%
SF4x16	7.5%, 41.5%, 76.8%	65.3%, 66.9%, 71.9%	96.2%
SF8x8	9.6%, 45.9%, 76.6%	66.8%, 66.8%, 72.0%	84.1%

The Pearson correlation coefficient (r) is computed for each backbone architecture. In the context of the domain-related task, which entails HAR classification, the computation averages (μ) the jogging confidence score at each RP. For the unrelated-domain task, the computation is based on the average (μ) results of the re-ID mAP

average results for re-ID mAP are displayed in the third column. The last column showcases the Pearson correlation (r) between both tasks.

Remarkably, both tasks exhibit a strong positive correlation in this context-based experiment. Notably, the C2D and S4x16 backbones show the highest correlation between the tasks. Interestingly, the C2D backbone consistently does not deliver the best results individually, but it strikes a balanced performance when bridging the gap between domain-related and domain-unrelated-domain tasks. In contrast, the SF8x8 backbone, widely used in literature for HAR-triggered transfer learning (see Section 2), demonstrates the weakest correlation. It excels primarily in jogging confidence for HAR but struggles to enhance its mAP when transitioning from ROI to PATH context. This challenge is observed in other models like I3D and S8x8, contributing to the limitation in achieving a stronger correlation.

Furthermore, the table underlines how the domain-specific task noticeably improves as the context widens, while the non-domain-specific tasks face challenges. Additionally, the table highlights the absence of a straightforward relationship between HAR classification performance and re-ID evaluation. Some models with subpar HAR performance, such as I3D NLN, achieve moderate to high performance in re-ID. This divergence could be attributed to classifier-related issues, where the HAR head encounters difficulties with these embeddings that the re-ID head does not experience.

Table 10 offers a distinctive perspective on the correlation aspects when the analysis is based on different encodings. In this context, models exhibit suboptimal performance when fed by-products of image encoding. As observed in the second column, the average jogging confidence scores at each RP are noticeably impacted by image encoding, in stark contrast to the RGB-based contextual approaches shown in Table 9. This outcome aligns with expectations, given that the training data primarily consisted of RGB videos.

However, it is important to emphasize that our objective, as outlined in Section 4.2, is not to showcase the method for achieving the best performance. Instead, we aim to elucidate the performance disparities among networks exposed exclusively to dynamic information and explore their correlations when tackling domain-specific and non-domain-specific tasks.

In this context, RAFT emerges as the top performer in the HAR task, likely owing to the inclusion of flow information during the training process, making recognizing actions more familiar to the models, as explained in the previous section. Intriguingly, this behavior does not translate similarly to the re-ID task, where the MiDaS encoding approach outperforms RAFT and AFD. This discrepancy notably impacts the correlation, resulting in a predominantly weak

Table 10 Encoding-based analysis of tasks correlation

Backbone	Domain-related task	Domain-unrelated task	*Pearson* Correlation *(r)*
	$\mu_{AFD}, \mu_{MiDaS}, \mu_{RAFT}$		
C2D	0.0%, 0.6%, 4.8%	34.9%, 65.9%, 50.1%	0.00%
I3D	0.0%, 0.3%, 9.5%	36.1%, 64.5%, 49.1%	−0.05%
I3D NLN	0.6%, 0.2%, 9.6%	36.3%, 68.6%, 48.4%	−0.13%
S4x16	0.0%, 0.0%, 13.9%	35.1%, 68.1%, 50.1%	−0.05%
S8x8	0.0%, 0.3%, 14.2%	32.7%, 65.4%, 48.4%	−0.02%
SF4x16	0.0%, 0.0%, 5.1%	38.3%, 65.0%, 43.7%	−**0.33**%
SF8x8	3.6%, 6.3%, 15.4%	37.8%, 61.4%, 44.1%	−0.04%

The Pearson correlation coefficient (r) is calculated for each backbone architecture. In the context of the domain-related task, which involves HAR classification, the computation involves averaging (μ) the jogging confidence score at each RP. For the domain-unrelated task, the calculation is based on the average (μ) results of the re-ID mAP

negative linear relationship between the evaluations of the two tasks. It is worth underlining that, on the contrary, dynamic features as conveyed by a flow should improve the re-ID based on personal kinematic strategies that underlie individual walk and run patterns [24, 42]. The strongest correlation occurs when SlowFast (SF4x16) is considered, and it is negative, implying that as one task improves its evaluation, the other task tends to decrease. Other correlations are closer to zero, indicating a weak or negligible relationship between the two.

In Fig. 5, we showcase the behavior of the models achieving the highest correlation between the domain-related task (DRT) and the domain-unrelated task (DUT) -specifically, C2D and S4x16 in the context-based analysis, and I3D NLN and SF4x16 in the encoding-based analysis. In the context-based analysis, the positive correlation is evident from the closely aligned performances of the orange and blue lines with their dotted counterparts, reflecting consistent model behavior across varying contextual parameters. Conversely, the encoding-based experiment unveils a notable negative correlation, illustrated by the divergent trends of the purple and green lines in contrast to their dotted counterparts. This disparity emphasizes these models' sensitivity to the encoding task's intricacies and underscores the need for a nuanced understanding of their performance characteristics.

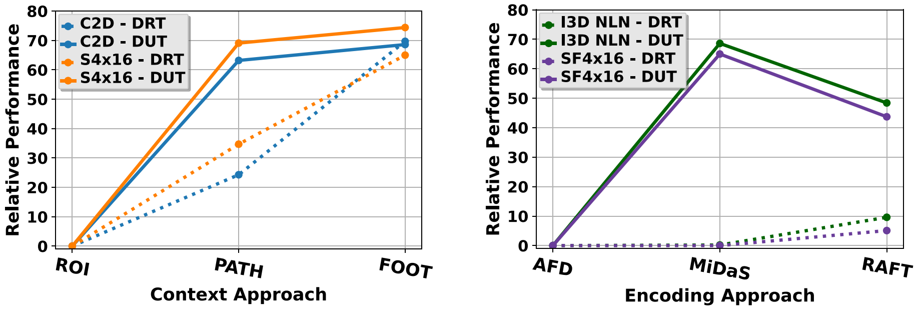

Fig. 5 Correlation Analysis. The figures provide a concise overview of the correlation analysis conducted on the models achieving the highest task correlation (according to Tables 9 and 10) in two distinct experiments: the context-based experiment (depicted on the left) and the encoding-based experiment (depicted on the right). In these visualizations, dotted lines represent outcomes from the domain-unrelated task (DUT), while solid lines illustrate findings from the domain-related task (DRT)

6 Conclusions

This paper introduces an innovative analysis to assess domain shift adaptation by correlating two distinct tasks: a domain-related task (HAR) and a domain-unrelated task (re-ID), both utilizing the same input data. To conduct this correlation analysis, the experiments employed seven different pre-trained backbone models feeding the ad-hoc devised architectures. The best performing among them was a novel transformer classifier specifically tailored to address the re-ID problem while utilizing pre-trained models for the HAR task. The scope of the presented research is delineated by two primary analyses: a contextual information-based analysis and an encoding-based analysis. The former seeks to elucidate how contextual information impacts the evaluation of both tasks, while the latter delves into evaluating these tasks when image encoding comes into play.

The reported findings from the contextual analysis reveal that the correlation between tasks is notably high when the same training encoding scheme (RGB) is employed. Furthermore, it becomes evident that the richer the contextual information, the more favorable the evaluation outcomes for both HAR and re-ID. In this scenario, we have achieved superior performance compared to the current state of the art in the re-ID task within this unconstrained sporting environment. However, our results also highlight the considerable challenges in achieving acceptable evaluation rates when considering the encoding experiment. In this scenario, HAR evaluation rates tend to decline, with only MiDaS (depth images) providing noteworthy results in re-ID. Interestingly, when MiDaS is employed, the HAR task exhibits subpar performance. This underscores an intriguing point, suggesting that the HAR classifier may contribute to this decline in HAR performance, as the features extracted from the identical backbones yield promising results for re-ID. Conversely, this issue is less pronounced when AFD or RAFT encoding schemes are used, as the resulting features are inadequate for either task.

This research offers relevant insights into domain shift adaptation and task correlations using shared inputs. Understanding how tasks influence each other with shared data is essential for robust AI systems. Additionally, it investigates contextual and encoding impacts on performance, benefiting applications like HAR and person re-ID, which are relevant in surveillance and sports analytics. This work advances adaptable and reliable AI models in complex environments, crucial for real-world AI applications.

The conclusions from this research suggest several directions for future work. Firstly, exploring advanced encoding schemes beyond RGB, MiDaS, AFD, and RAFT could yield insights into their impact on task correlations between Human Activity Recognition (HAR) and person re-identification (re-ID). Experimenting with diverse encoding strategies may address challenges in achieving acceptable evaluation rates, particularly in the HAR task.

In particular, investigating the observed decline in HAR evaluation rates with specific encoding schemes, such as MiDaS, requires attention. Understanding the interplay between encoding processes and the HAR classifier's performance may lead to refinements in model architectures or training strategies to mitigate challenges. Moreover, the potential impact of the HAR classifier on performance decline, especially when features from identical backbones yield promising re-ID results, deserves exploration. Fine-tuning the HAR classifier to align with encoding schemes may be a promising avenue for future research.

Furthermore, extending the study to diverse sporting environments could enhance generalizability. Assessing the adaptability of the proposed models across different settings and activities would contribute to the robustness of AI systems in complex, dynamic environments. In summary, future work involves a deeper exploration of encoding strategies, addressing observed declines in HAR evaluation rates, refining the interplay between the

HAR classifier and encoding schemes, and broadening application scenarios to validate model adaptability. These efforts collectively advance the understanding and applicability of domain shift adaptation in AI models.

Acknowledgements This work is partially funded by the Spanish Ministry of Science and Innovation under project PID2021-122402OB-C22 and by the ACIISI-Gobierno de Canarias and European FEDER funds under project ULPGC Facilities Net and Grant EIS 2021 04.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. Baradaran M, Bergevin R (2023) Multi-task learning based video anomaly detection with attention. In: Conference on computer vision and pattern recognition workshops (CVPRW). pp 2886–2896
2. Bensland S, Paul A, Grossmann L, Hoogland IE, Riener R, Paez-Granados D (2023) Healthcare monitoring for sci individuals: learning activities of daily living through a slowfast neural network. In: IEEE/SICE international symposium on system integration (SII). pp 1–7
3. Buhrmester V, Münch D, Arens M (2021) Analysis of explainers of black box deep neural networks for computer vision: a survey. Mach Learn Knowl Extr 3(4):966–989
4. Cao Z, Long M, Wang J, Jordan MI (2017) Partial transfer learning with selective adversarial networks. IEEE/CVF conference on computer vision and pattern recognition. pp 2724–2732
5. Carreira J, Zisserman A (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 4724–4733
6. Cheng Y, Xu Z, Fang F, Lin D, Fan H, Wong Y, Sun Y, Kankanhalli MS (2023) A study on differentiable logic and llms for epic-kitchens-100 unsupervised domain adaptation challenge for action recognition 2023. arXiv:2307.06569
7. Day O, Khoshgoftaar TM (2017) A survey on heterogeneous transfer learning. J Big Data 4(1):29
8. Diao X, Xu Y (2022) A slowfast-based violence recognition method. In: Asian conference on artificial intelligence technology (ACAIT). pp 1–6
9. Farahani A, Voghoei S, Rasheed K, Arabnia HR (2021) A brief review of domain adaptation. In: Advances in data science and information engineering. pp 877–894
10. Feichtenhofer C, Fan H, Malik J, He K (2017) Slowfast networks for video recognition. 2019 IEEE/CVF International conference on computer vision (ICCV). pp 6201–6210
11. Feichtenhofer C, Fan H, Xiong B, Girshick RB, He K (2021) A large-scale study on unsupervised spatiotemporal representation learning. 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 3298–3308
12. Foster DP, Kakade SM, Salakhutdinov R (2011) Domain adaptation: overfitting and small sample statistics. arXiv:1105.0857
13. Freire-Obregón D, Barra P, Castrillón-Santana M (2022) de Marsico M (2022) Inflated 3D ConvNet context analysis for violence detection. Mach Vis Appl 33:15
14. Freire-Obregón D, Lorenzo-Navarro J, Santana OJ, Hernández-Sosa D, Castrillón-Santana M (2022) Towards cumulative race time regression in sports: I3D ConvNet transfer learning in ultra-distance running events. In: International conference on pattern recognition (ICPR). pp 805–811
15. Freire-Obregón D, Lorenzo-Navarro J, Santana OJ, Hernández-Sosa D, Castrillón-Santana M (2023) A large-scale re-identification analysis in sporting scenarios: the Betrayal of Reaching a Critical Point. In: International joint conference on biometrics (IJCB)

16. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V (2016) Domain-adversarial training of neural networks. J Mach Learn Res 17(59):1–35
17. Hassan A, Elgabry A, Hemayed E (2021) Enhanced dynamic sign language recognition using slowfast networks. In: International computer engineering conference (ICENCO). pp 124–128
18. Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros A, Darrell T (2018) CyCADA: Cycle-consistent adversarial domain adaptation. In: Dy J, Krause A (eds) International conference on machine learning, vol. 80. pp 1989–1998
19. Ilic F, Pock T, Wildes RP (2022) Is appearance free action recognition possible? In: European conference on computer vision (ECCV)
20. Jiang S, Campbell D, Lu Y, Li H, Hartley RI (2021) Learning to estimate hidden motions with global motion aggregation. International conference on computer vision (ICCV). pp 9752–9761
21. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A (2017) The Kinetics Human Action Video Dataset. CoRR
22. Koshti D, Kamoji S, Kalnad N, Sreekumar S, Bhujbal S (2020) Video Anomaly Detection using Inflated 3D Convolution Network. In: International conference on inventive computation technologies (ICICT). pp 729–733
23. Long M, Wang J, Ding G, Sun J, Yu PS (2013) Transfer feature learning with joint distribution adaptation. IEEE international conference on computer vision. pp 2200–2207
24. Marsico MD, Mecca A (2019) A survey on gait recognition via wearable sensors. ACM Comput Surv (CSUR) 52(4):1–39
25. Panareda Busto P, Gall J (2017) Open set domain adaptation. In: International conference on computer vision. pp 754–763
26. Patrick M, Campbell D, Asano YM, Metze IMF, Feichtenhofer C, Vedaldi A, Henriques JF (2021) Keeping your eye on the ball: Trajectory attention in video transformers. Neural Inform Process Syst https://api.semanticscholar.org/CorpusID:235390605
27. Penate-Sanchez A, Freire-Obregón D, Lorenzo-Melián A, Lorenzo-Navarro J, Castrillón-Santana M (2020) TGC20ReId: a dataset for sport event re-identification in the wild. Pattern Recogn Lett 138:355–361
28. Qu S, Zou T, Roehrbein F, Lu C, Chen GS, Tao D, Jiang C (2023) Upcycling models under domain and category shift. IEEE/CVF conference on computer vision and pattern recognition. pp 20,019–20,028
29. Ranftl R, Lasinger K, Hafner D, Schindler K, Koltun V (2020) Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence (TPAMI)
30. Raschka S (2020) Model evaluation, model selection, and algorithm selection in machine learning
31. Pandian D, Rajagopalan SS, Jayagopi D (2022) Detecting a child's stimming behaviours for autism spectrum disorder diagnosis using rgbpose-slowfast network. In: IEEE international conference on image processing (ICIP). pp 3356–3360
32. Sakaino H (2023) Panopticvis: Integrated panoptic segmentation for visibility estimation at twilight and night. In: Conference on computer vision and pattern recognition workshops (CVPRW). pp 3385–3398
33. Saleem G, Bajwa UI, Raza RH (2023) Toward human activity recognition: a survey. Neural Comput Appl 35(5):4145–4182
34. Santos F, Durães D, Marcondes FS, Lange S, Machado J, Novais P (2021) Efficient violence detection using transfer learning. In: Highlights in practical applications of agents, multi-agent systems, and social good. The PAAMS Collection. Springer International Publishing, pp 65–75
35. Sarkar P, Beirami A, Etemad A (2023) Uncovering the hidden dynamics of video self-supervised learning under distribution shifts
36. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conf. on computer vision and pattern recognition (CVPR). pp 815–823
37. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. arXiv:1406.2199
38. Sun S, Shi H, Wu Y (2015) A survey of multi-source domain adaptation. Inform Fus 24:84–92 https://doi.org/10.1016/j.inffus.2014.12.003 https://www.sciencedirect.com/science/article/pii/S1566253514001316
39. Teed Z, Deng J (2020) Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision (2020). https://api.semanticscholar.org/CorpusID:214667893
40. Thomas AK, Poovizhi P, Saravanan M, Tharageswari K (2023) Animal intrusion detection using deep learning for agricultural fields. In: International conference on smart systems and inventive technology (ICSSIT). pp 1021–1027
41. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: NIPS

42. Wan C, Wang L, Phoha VV (2018) A survey on gait recognition. ACM Comput Surv (CSUR) 51(5):1–35
43. Wang J, Lan C, Liu C, Ouyang Y, Qin T (2021) Generalizing to unseen domains: A survey on domain generalization. IEEE Trans Knowl Data Eng 35:8052–8072
44. Wang X, Girshick RB, Gupta AK, He K (2017) Non-local neural networks. 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 7794–7803
45. Wang Y, Dantcheva A (2020) A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. In: IEEE international conference on automatic face and gesture recognition (FG 2020). pp 515–519
46. Wang Y, Wang S, Zhou M, Jiang Q, Tian Z (2019) Ts-i3d based hand gesture recognition method with radar sensor. IEEE Access 7:22,902-22,913
47. Zhang H, Xiao X, Huang T, Liu S, Xia Y, Li J (2019) An novel end-to-end network for automatic student engagement recognition. In: International conference on electronics information and emergency communication (ICEIEC). pp 342–345
48. Zhang J, Ding Z, Li W, Ogunbona P (2018) Importance weighted adversarial nets for partial domain adaptation. IEEE/CVF conference on computer vision and pattern recognition. pp 8156–8164
49. Zhang S, Dong J, Chervan A, Kurlovich D, Hou W, Ding M (2023) Reinforcing local structure perception for monocular depth estimation. IEEE Sens J 23(16):18539–18549
50. Zhang Y, Sun P, Jiang Y, Yu D, Yuan Z, Luo P, Liu W, Wang X (2021) ByteTrack: multi-object tracking by associating every detection box. In: European conference on computer vision