electronics

MDPI

Article

Comparison Analysis of Multimodal Fusion for Dangerous Action Recognition in Railway Construction Sites

Otmane Amel [1,*], Xavier Siebert [2] and Sidi Ahmed Mahmoudi [1]

[1] ILIA Lab, Faculty of Engineering, University of Mons, 7000 Mons, Belgium; sidi.mahmoudi@umons.ac.be
[2] Department of Mathematics and Operational Research, University of Mons, 7000 Mons, Belgium; xavier.siebert@umons.ac.be
* Correspondence: otmane.amel@umons.ac.be

Abstract: The growing demand for advanced tools to ensure safety in railway construction projects highlights the need for systems that can smoothly integrate and analyze multiple data modalities, such as multimodal learning algorithms. The latter, inspired by the human brain's ability to integrate many sensory inputs, has emerged as a promising field in artificial intelligence. In light of this, there has been a rise in research on multimodal fusion approaches, which have the potential to outperform standard unimodal solutions. However, the integration of multiple data sources presents significant challenges to be addressed. This work attempts to apply multimodal learning to detect dangerous actions using RGB-D inputs. The key contributions include the evaluation of various fusion strategies and modality encoders, as well as identifying the most effective methods for capturing complex cross-modal interactions. The superior performance of the MultConcat multimodal fusion method was demonstrated, achieving an accuracy of 89.3%. Results also underscore the critical need for robust modality encoders and advanced fusion techniques to outperform unimodal solutions.

Keywords: multimodal fusion; RGB-D dangerous action recognition; deep learning

check for updates

Citation: Amel, O.; Siebert, X.; Mahmoudi, S.A. Comparison Analysis of Multimodal Fusion for Dangerous Action Recognition in Railway Construction Sites. *Electronics* **2024**, *13*, 2294. https://doi.org/10.3390/electronics13122294

Academic Editor: Krzysztof Zboiński

Received: 4 April 2024
Revised: 23 May 2024
Accepted: 31 May 2024
Published: 12 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Construction and transportation are among the industries in the EU with the biggest number of fatal workplace accidents, highlighting the high-risk nature of these industries (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Accidents_at_work_-_statistics_by_economic_activity (accessed on 7 May 2024)). For that matter, it is important to ensure the safety of workers in the field of railway construction with advanced surveillance systems. Recently, artificial intelligence methods such as deep learning models have emerged as powerful solutions that can ensure worker safety by changing the way construction sites are monitored and managed. This is due to the advanced performance of deep learning in the computer vision field, which makes it particularly adapted for identifying patterns of risky behavior and predicting potential accidents [1].

Among deep learning models, research in multimodal deep learning models has known greater interest [2,3] as it was proven for their greater performance in classification tasks [4]. Especially in the era of big data, where information is available in various forms such as text, audio, video, and images, it has become increasingly important to develop deep learning algorithms that can simultaneously process these different types of data. Multimodal learning algorithms attempt to leverage the diversity of data types to improve the performance of existing unimodal deep learning solutions. As the human brain can by design integrate sensory data to interpret the world, multimodal learning also aims to fuse information from different modalities for an improved decision-making process. These methods have demonstrated remarkable potential in many fields [5], including but not limited to computer vision, natural language processing (NLP), healthcare, and surveillance systems.

In this work, multimodal fusion approaches were applied in RGB-D action recognition by combining RGB and depthmap frames, which in theory will increase the overall performance of the model [4]. The motivation of this work is to combine the RGB data information such as color information that captures texture and appearance, with depth data that provide spatial information about the scene's geometry. By aggregating them, a model can exploit both the appearance cues (RGB data) and the spatial cues (depth data), ultimately extracting more comprehensive features which enable one to obtain a rich and complete perception of the scene. The impacts of this research are important, with the potential to drastically reduce accidents and enhance safety standards in high-risk sectors in railway construction sites.

In this work, the following contributions are made:

- Proposing the fusion of unimodal action recognition models as encoders for their ability to capture useful features.
- Providing an analysis of multimodal fusion approaches, highlighting their effectiveness in recognizing dangerous actions in railway construction.
- Analyzing the contribution of depthmaps modality in terms of performance.

The remainder of this paper is organized into four main sections. Section 2 covers the state-of-the-art in multimodal learning and RGB-D action recognition. Section 3 presents the proposed RGB-D dangerous action recognition approach using multimodal fusion methods. Section 4 provides a detailed analysis of the results, where the impact of different design choices are studied, such as the chosen fusion method and modality encoder, as well as the impact of depthmaps modality. Finally, Section 5 concludes the report by summarizing the main findings and takeaways from this study.

2. Literature Review

This section covers the state of the art in multimodal learning, such as its taxonomy, key concepts, and a detailed comparison of multimodal fusion methods. Furthermore, related work in RGB-D action recognition will be discussed and categorized based on the techniques used.

2.1. Multimodal Learning Definitions and Concepts

Multimodal machine learning is a learning paradigm that aims to mimic human behavior interacting with the surrounding world. Compared to the high-level cognitive abilities of the human brain, the goal of this paradigm is to combine multiple modality data sources when involved in capturing cross-modal interactions. However, those modalities bear characteristics that need to be taken into consideration for optimal learning, and in the following subsection, these aspects will be explored in greater detail.

2.1.1. Multimodal Principles Definition

A modality refers to the specific setting or type of signal where an event takes place or is observed [3] depending on the sensory inputs. An observation state can be recorded through multiple sources. For example, an image holds a visual representation of an object along with a caption that has its textual description. Both depict the corresponding object (or scene) but differently and uniquely. These kinds of modalities are often heterogeneous but complementary due to their interconnections. In the literature, these two are defined as follows:

Heterogeneity

As described in [3], the heterogeneity of modalities is a complex and nuanced concept that should be approached as a spectrum rather than a binary choice, with different modalities carrying variety levels of heterogeneity depending on their qualities, structures, and representations. Therefore, it requires careful consideration to define the interconnections between modalities in multimodal learning. For instance, while two images captured using the same camera may share some similarities, they can still differ in lighting condi-

tions, camera settings, and other factors, hence exhibiting some heterogeneity. In contrast, utilizing different sensors (such as depth maps and Lidar sensors) might create more heterogeneity since they record different types of information and have distinct qualities, structures, and representations. Thus, in multimodal learning, modal heterogeneity should be viewed as a subtle complex term that differs based on the modalities involved.

Interconnections

Having two heterogeneous modalities does not exclude the fact of exhibiting complementary information that interacts with each other [3]. This characteristic has at least two types:

- Modality connections: Describe how modalities are often related and share commonalities, such as correspondences between the same concept in language and images or dependencies across spatial and temporal dimensions. Several dimensions of modality connections were outlined based on both statistical (association and dependency) and semantic (correspondence and relationship) perspectives (Figure 1).
- Modality interactions: Modality interactions study how modality elements interact to give rise to new information when integrated for task inference. It is important to highlight a key difference between interaction and connections: the interaction takes place when at least two modalities are involved during the learning process of a multimodal model, which helps devise a new response during inference compared to an unimodal model.

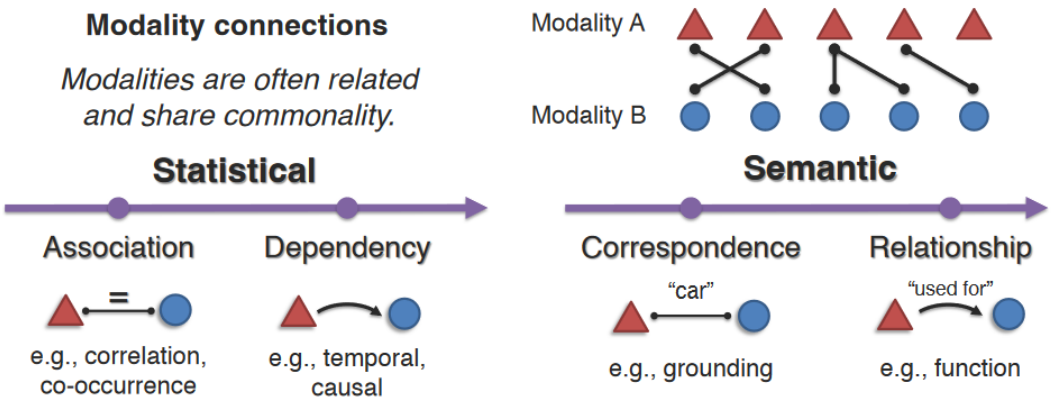


Figure 1. Modality connections [3].

Having established a clear understanding of what constitutes a modality and its fundamental principles, an overview of the main areas in multimodal learning will be provided next.

2.2. Multimodal Learning Taxonomy Overview

Multimodal learning is often divided into five areas: representation, alignment, generation, fusion, and co-learning. In a recent survey [3], quantification was added as an important area that aims to interpret and suggest methods for applying XAI in multimodal architectures. It is worth noting that these areas often overlap when solving a complex task such as co-learning [6].

2.2.1. Representation Learning

The purpose of representation learning is to capture complementary data and exploit these through a proper representation vector that summarizes the most valuable features [2,3,7]. The heterogeneity of multimodal data makes it challenging to construct such representations. For example, language is often symbolic while audio and visual modalities will be represented as signals. While issues with missing data or the absence of some modalities may arise, representations should be adapted to solve the issues as men-

tioned earlier. In the literature, it is often divided into two categories, joint and coordinated representation, as illustrated in Figure 2.

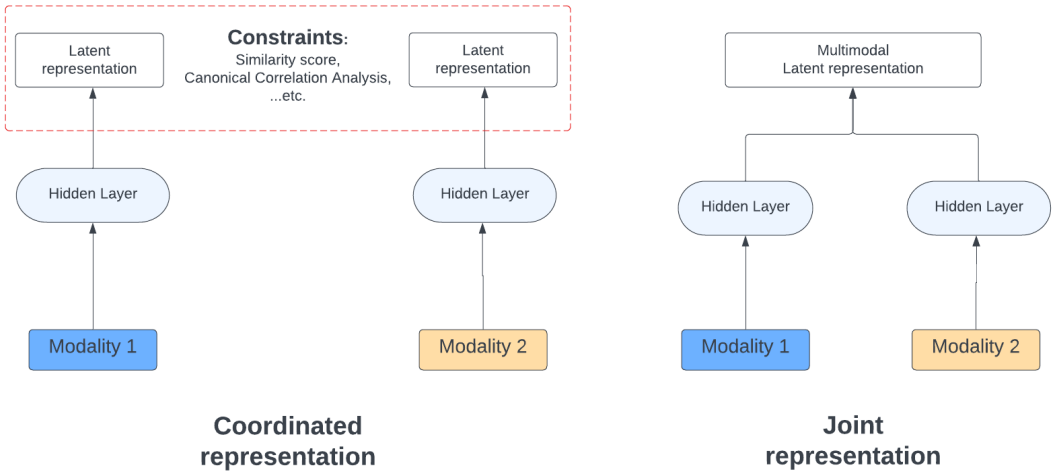


Figure 2. Illustration of joint and coordinated representation learning inspired by [2].

2.2.2. Alignment

The main objective is to find correspondence and alignment between the sub-components of instances that preserve order, it can be divided into at least two categories: explicit and implicit alignment [2]. The first category covers use cases where the model will explicitly align sub-components between modalities, for example, align recipe steps with video instruction [8]. Whether the implicit one is based on an intermediate (latent) representation, (e.g., image retrieval based on text may require an alignment step between words and image regions). For example, [9] consists of RVOS (referring video object segmentation task) task that used multimodal transformers for aligning the region of interest with the most semantically related word.

2.2.3. Generation

Generation (or translation) is referred to in the context of mapping one modality to another. In other words, the purpose is to generate the same entity in a different modality (e.g., image captioning). It can be categorized into two types: example-based and combination-based models [2]. The example-based models are parametric since they depend on a dictionary used for cross-modal retrieval based on the closest element. Whether the combination-based models, they combined them to construct more meaningful and semantic intermediate representations. Generally, they are hand-crafted or based on heuristics. A concrete example of generation models is Dall-E [10] a transformer-based system for generating images from text captions developed by OpenAI. It consists of bridging the gap between vision and text across a broad spectrum of concepts described in natural language. Utilizing a 12-billion parameter variant of the GPT-3 transformer model, it interprets natural language inputs and produces corresponding images based on textual descriptions (see Figure 3).

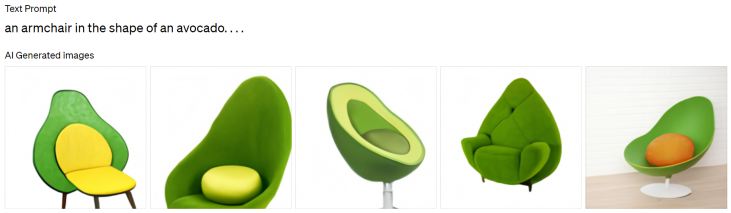


Figure 3. Dall-E image generation example from OpenAI official website.

2.2.4. Co-Learning

Co-learning refers to transferring knowledge from one mode of learning to another [2,3,6]. This is often achieved by incorporating external modalities into the training process, creating a shared representation space, and evaluating how this model performs on tasks that involve only one modality. Examples of this approach include using word embeddings to classify images, knowledge graphs for image classification, and video data for text classification. Co-learning is valuable because it allows us to improve unimodal systems by integrating external data, which is particularly beneficial for low-resource tasks. Moreover, the principles of co-learning can provide insights into other multimodal tasks by guiding the creation of a joint representation space. For instance, in [11], the authors suggested training a hallucination network to transfer depth map information into an RGB stream that allows the model to make inferences even when depth modality is missing without compromising the performance.

2.2.5. Quantification

Quantification is the study that aims to provide a deeper understanding of the heterogeneity, cross-modal interactions, and multimodal learning process through empirical and theoretical analysis. According to [3], quantification can be categorized into three sub-challenges:

- Identifying the dimensions of heterogeneity in multimodal data and their impact on modeling and learning, such as the presence of modality bias and noise.
- Quantifying the types of connections and interactions among different modalities in datasets and trained models.
- Characterizing the learning and optimization challenges involved in heterogeneous data.

It is worth mentioning the importance of studying these sub-challenges in the context of other multimodal challenges. A comprehensive understanding of these core challenges will lead to insights that can enhance the robustness, interpretability, and reliability of multimodal applications in real-world scenarios.

2.2.6. Fusion

Fusion is one of the primary and innovative areas of focus in multimodal learning. This concept entails the integration of various modalities within a single network, aiming to achieve continuous predictions, such as regression, or discrete predictions, such as classification. The existing literature typically categorizes fusion into four distinct approaches: early, late, intermediate, and hybrid fusion [2,12].

- Early fusion: This is the study of learning a joint representation at an early stage (raw modalities), which was one of the first attempts to achieve representation learning [2]. This type of fusion may vary depending on the modality's level of abstraction. In the literature, the most common methods used for multimodal fusion are concatenation, element-wise multiplication, and weighted sum [7,12].
- Intermediate fusion: Also known as feature-level fusion [13]. Contrary to the early fusion, this approach takes features at a higher level after feeding the raw modalities through shallow layers. This method is advantageous for capturing cross-modal interactions and relationships because it allows for the integration of more abstract representations of the data from each modality. It may also be more susceptible to incorporating noisy information [13] or allowing the dominance of one modality to influence the overall model's performance negatively.
- Late fusion: Also known as decision-level fusion, late fusion is a technique where data sources are processed independently and then combined at a later stage for decision-making purposes. This approach is based on ensemble classifiers and is simpler than the early fusion method, especially when dealing with data sources with varying characteristics such as sampling rate, data dimensionality, and measurement

unit. One of the most common approaches is decision aggregation by averaging [14], voting [15], or weighted-sum [16]

- Hybrid fusion: Hybrid fusion is the methodology of merging the attributes of both early and late fusion techniques within a unified architecture. Most prior research focuses on the combined representation created by Early Fusion at the feature level. This integrated representation is then combined with decision-level approaches, such as weighted voting [17], to establish a more effective fusion process.

Frequently, overfitting can be an issue in multimodal data fusion, as data from diverse sources and inconsistent learning rates generalize in distinct ways. To tackle this overfitting issue, adopting a flexible and adaptive fusion strategy, along with a regularization method like gradient blending, can help calculate the ideal combination of modalities according to their overfitting tendencies [18].

2.3. Multimodal Fusion Techniques

Moving forward, state-of-the-art multimodal fusion techniques widely used for various application tasks will be presented. They can be divided into three types: tensor-based fusion, transformers-based fusion, and adaptive fusion.

2.3.1. Tensor-Based Fusion:

- Tensor fusion: The main idea behind this technique is to capture complex interactions and complementary information between modalities' embeddings. To do so, in [19], they suggested performing a 3-fold Cartesian multiplication between all modalities (see Figure 4). The particularity of this technique is that it preserves the unimodal characteristics as well as the trimodal and bimodal ones. Despite the advantages of being non-parametric, it comes with a high computation cost, which increases the model complexity and might not be adapted for real-world applications.

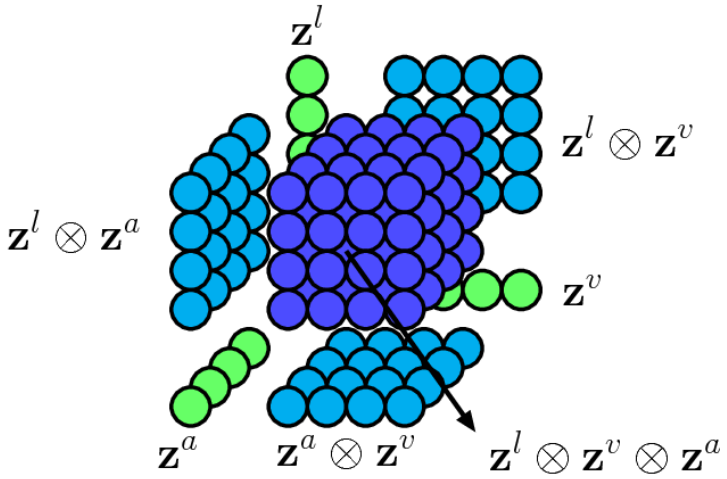


Figure 4. Tensor fusion method [19].

- Low-rank tensor fusion: To mitigate the computation cost without compromising the performance, [20] divides the tensor fusion multiplication into low-factor weights that reduce the number of parameters (see Figure 5). Their work shows prominent results in terms of linear scalability with the number of modalities being used.

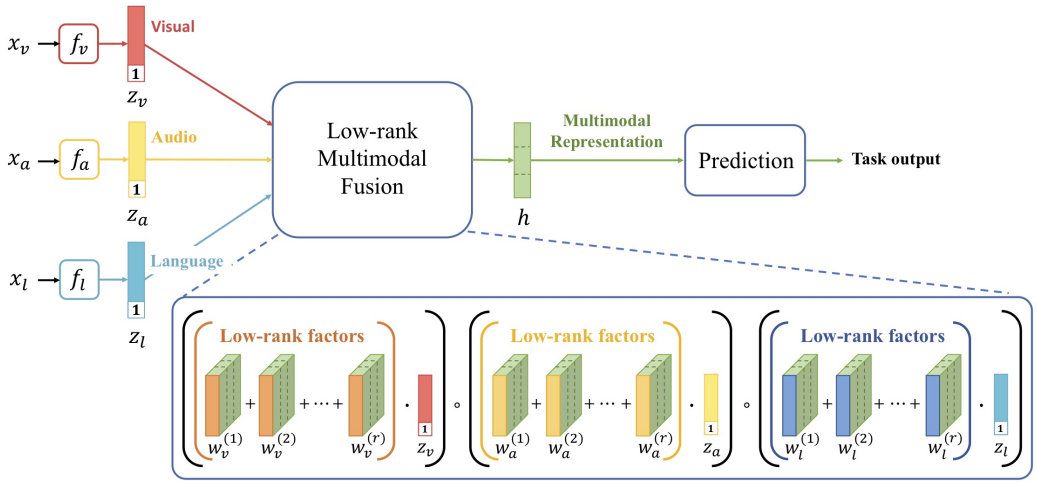


Figure 5. Low-rank multimodal fusion method [20].

2.3.2. Transformers-Based

With the recent success of Transformers [21] with both text and image modalities, a rise in architectures that take both inputs simultaneously for multimodal tasks have been widely used recently. In the literature, and more specifically Bert-like models, they are often divided into two types: single-stream and two-stream models [22]. A single stream suggests combining modalities at the entry level and is then processed using a single module followed by a co-attention layer, whether a two-stream approach takes two modules, each of them will treat a modality separately.

- VisualBert: [23] a single-stream approach that uses self-attention integrated into a transformer layer for an implicit alignment between image and text inputs (see Figure 6). The image features were generated from a ResNeXt-based Faster RCNN pre-trained on Visual Genome [24] and text tokens using a BERT tokenizer [25]. The main contribution of this architecture is the ability to capture cross-modal interactions by aligning visual and textual information; in addition, it can generate a joint representation by leveraging the transformer architecture to learn contextualized embeddings for words and image regions.
- LXMERT: [26] LXMERT (learning cross-modality encoder representations from transformers) is a two-stream approach introduced for understanding and reasoning across both visual and textual modalities. The architecture leverages the powerful Transformer architecture and employs a cross-modality attention mechanism to learn the joint representations of image and text data. The latter are implemented through a series of co-attention layers that consist of three attention sub-modules: self-attention for the visual modality, self-attention for the textual modality, and cross-attention between the two modalities. The self-attention sub-modules focus on learning relationships within each modality, while the cross-attention sub-module learns to align and relate the information across the visual and textual domains. LXMERT is pre-trained on large-scale datasets to extract general features from images and text and is then fine-tuned for specific tasks such as a visual question answering image-caption matching, and visual reasoning.
- Multimodal attention bottleneck: [27] A single-stream approach that comes with a novel way of fusing modalities using attention bottleneck fusion. The intuition behind this technique is to condense and capture the most important complementary information between modalities and omit what is unnecessary. It has been proven that it is computationally efficient and performs better than classical self-attention fusions. As illustrated in Figure 7, They achieved this by adding a set of B extra fusion bottleneck tokens to the input sequence and then used to condense all the cross-modal interaction of modalities through bottleneck tokens.

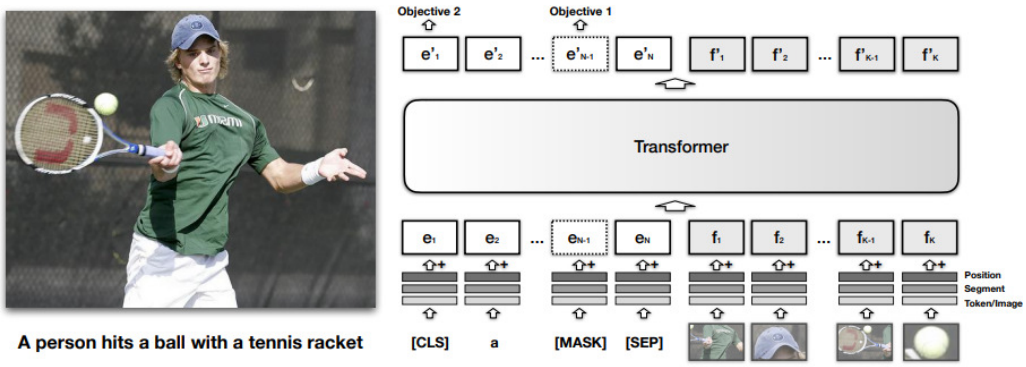


Figure 6. VisualBERT architecture [23].

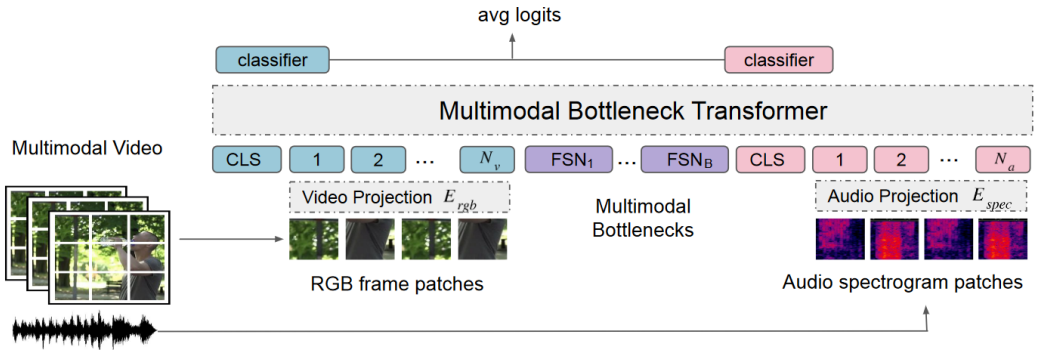


Figure 7. This figure illustrates the integration of bottleneck tokens into the audiovisual input [27].

2.3.3. Adaptive Fusion

- Auto-fusion: In this work [28], an adaptive fusion is proposed that allows for effective multimodal fusion. Instead of using static techniques such as concatenation, it lets the network decide how to combine a given set of multimodal features more effectively. The main idea behind the auto-fusion is to maximize the correlation between multimodal inputs and use a reconstruction strategy to obtain a novel fusion layer that is minimized using an Euclidean distance between the original and reconstructed concatenated vector.
- MFAS: "Multimodal Fusion Architecture Search" [29] is a research study focused on developing a method for automatically discovering optimal fusion architectures for multimodal data. The central idea is to utilize neural architecture search (NAS) [30,31] techniques to search for the best fusion strategy that combines information from multiple modalities. This approach aims to address the challenge of designing effective fusion strategies for multimodal tasks by employing NAS techniques to automatically learn the most suitable architecture for a given task. They suggested picking a modalities representation at the different stages/layers of each encoder and nonlinear transformation to come up with the optimized combination that takes into account the appropriate abstraction level of the modality representation and the fusion technique used.

Next, state-of-the-art fusion techniques used for RGB-D action recognition and their categories will be examined.

2.4. RGB-D Fusion Methods

The domain of action recognition has significantly evolved with the integration of deep learning techniques [32], particularly through the processing of RGB and depth map sequences. This section categorizes the notable advancements into fusion strategies, hybrid models, and machine learning-based solutions. To provide a comprehensive overview,

Table 1 summarizes the key multimodal fusion methods, highlighting their fusion method, input types, encoders, and datasets used.

2.4.1. Fusion Strategies for Action Recognition

A variety of fusion strategies have been explored to enhance the performance of action recognition systems. In [13], the impacts of early, intermediate, and late fusion techniques in multimodal settings are discussed. Similarly, ref. [33] employs a late fusion approach by integrating the CNN streams of dynamic image and depth data. In [34], c-ConvNet is introduced, which operates over constructed visual and depth dynamic images from the raw data RGB and depth maps at the input level, utilizing a decision-level score-product fusion, adaptable for scenarios with missing modalities. Furthermore, ref. [35] adopts a late fusion strategy using the Naive Bayes method for classifier score fusion.

2.4.2. Deep Learning-Based Solutions

Several studies have proposed hybrid networks and advanced models to capture complex spatial and temporal information. The authors in [36] explore a deep learning auto-encoder framework for RGB-D action recognition, focusing on capturing shared information between modalities. The authors in [37] present a hybrid CNN-RNN network, leveraging weighted dynamic images and canonical correlation analysis. The authors in [38] introduce a discriminative model combining 3D-CNN and LSTM with an attention mechanism. The authors in [39] propose a Transformer-based framework for egocentric action recognition, emphasizing inter-frame and mutual-attentional fusion. The authors in [40] contrast previous methods by integrating a SlowFast multimodality compensation block for processing depth and RGB sequences. The authors in [41] propose a dual-stream cross-modality fusion transformer, enhancing feature representations through self-attention and cross-attention mechanisms. The authors in [42] also proposed a dual-stream architecture with a cross-modality compensation block (CMCB) that captures cross-modal interactions. The authors in [39] present a Transformer-based RGB-D egocentric action recognition framework that utilizes a self-attention mechanism to model temporal structures in video data. It consists of an inter-frame attention encoder and a mutual-attentional fusion block for processing and integrating features from RGB and depth modalities. The authors in [43] suggested a multi-modal contextualization unit (MCU), which encodes sequences from one modality with action content features from others, like RGB with depth or IR.

2.4.3. Machine Learning-Based Solutions

Some works suggest that traditional machine learning approaches remain relevant, with [35] modeling action videos using a bag-of-visual-words model and a multi-class SVM classifier. Ref. [44] primarily focuses on depth-based input for human action recognition, employing depth motion maps (DMM) and local binary patterns (LBPs) applied to depth data for feature extraction and representation.

For the further exploration of RGB-D action recognition techniques, the surveys [32,45,46] and the review in [47] offer detailed insights into the field's current state and future directions.

Table 1. Table summarizing RGB-D fusion methods, including the encoders used, fusion input blocks, and experimental datasets.

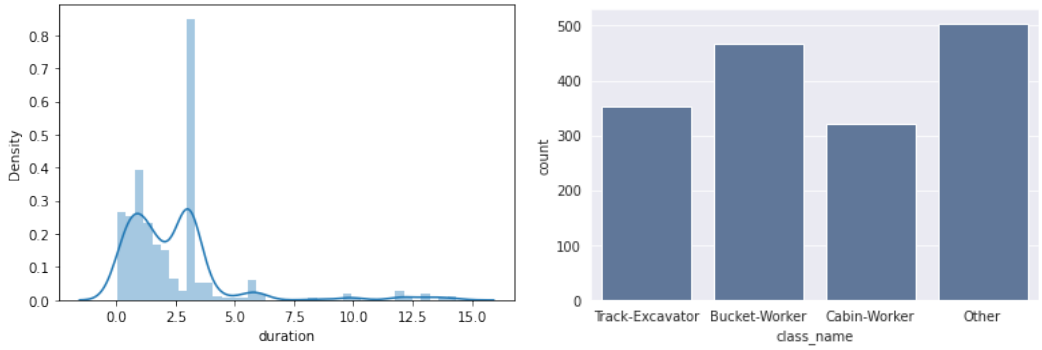
Fusion Methods	Fusion Inputs	Encoders	Datasets
SKPDE-ELM (hybrid) [44]	Depth motion map (DMM), local binary pattern (LBP)	ELM optimized by SKPDE	MSRAction3D [48], MSRDaily Activity3D [49], MSRGesture3D [50], UTD-MHAD [51]
Deep autoencoder based (intermediate) [36]	RGB and depth features	Not specified	Online RGBD action [52], MSRDaily Activity3D [49], NTU RGB+D [53], 3D action pairs [54], RGBD-HuDaAct [55]
Multi-modal contextualization unit (intermediate) [43]	RGB and depth embeddings	ResNet-18	NTU RGB+D 60 [53], NTU RGB+D 120 [56], NW-UCLA [57]
Mutual-attentional fusion block (intermediate) [39]	RGB and depth embeddings	ResNet-34 and Transformer encoder	THU-READ [58], FPHA [59], WCVS [60]
Attention mechanism (Intermediate) [38]	RGB and depth	Densely connected 3D CNN	Real-set, SBU-Kinect, MSR-action-3D
CCA-based feature fusion [37]	RGB and depth	3D ConvLSTM, weighted dynamic images	ChaLearn LAP IsoGD [61], NTU RGB+D [53], multi-modal and multi-view and interactive benchmark [62]
Naive Bayes combination (late fusion) [35]	RGB and depth	SIFT (RGB) and SURF (depth)	UTKinect-Action3D [63], CAD-60 [64,65], LIRIS human activities
Product score fusion (late fusion) [34]	Visual (RGB) and depth dynamic images	c-ConvNet	ChaLearn LAP IsoGD [61], NTU RGB+D [53]
Fusion score (late fusion) [33]	RGB and depth frames	Pre-trained VGG networks	MSRDaily activity 3D [49], UTD-MHAD [51], CAD-60 [64,65]
early, intermediate, and late fusion [13]	RGB, depth, skeleton	I3D and shift-GCN	NTU RGB+D [53], SBU interaction [66]
Cross-modality compensation block (intermediate) [42]	RGB and depth	ResNet and VGG with CMCB	NTU RGB+D 120 [56], THU-READ [58], PKU-MMD [67]
SlowFast multimodality compensation block (intermediate) [40]	RGB & Depth features	Swin transformer	NTU RGB+D 120 [56], NTU RGB+D 60 [53], THU-READ [58], PKU-MMD [67]
Cross-modality fusion transformer (intermediate) [41]	RGB and depth features	Restnet50 feature extractors	NTU RGB+D 120 [56], THU-READ [58], PKU-MMD [67]

3. Proposed Approach for RGB-D Dangerous Action Recognition

This section will present the proposed RGB-D dangerous action recognition approach using multimodal fusion methods. The models were evaluated based on the accuracy metric on a simulated dataset to solve a multi-class classification task. The goal is to assess whether multimodal fusion helps one obtain an improvement in terms of performance for detecting dangerous actions in railway construction to satisfy customers' needs. First, the dataset used is presented. Second, the problem is formulated mathematically. Third, the modality encoders used are defined. Finally, the experimental setup is described.

3.1. Dataset

The railway construction dangerous action recordings were initially introduced in the work [1]. Readers are encouraged to refer to the Figure 8 and Table 2 for the class distribution statistics and a detailed definition of each dangerous action. This dataset consists of dangerous actions generated using the game engine Unity (https://unity.com/ (accessed on 7 May 2024)) which gives enough flexibility to generate a wide range of scenarios within these environments. The scenes consisted of scenes close to real-world scenarios where workers are confronted with dangerous situations. Given that this unimodal dataset was initially provided (RGB sequences only), this work has been extended by generating synthetic depth map sequences given RGB modality. Indeed, multiple works have been proposed for depth estimation [68] and monodepth2 [69] was selected for its superior results on the KITTI benchmark [70]. In Figure 9, an example using monodepth2 on a simulated video sample is demonstrated.



(**a**) (**b**)

Figure 8. Dataset statistics of the simulated dataset, as provided in [1]: (**a**) Action's duration (in seconds) distribution; and (**b**) class distribution of the simulated dangerous actions.

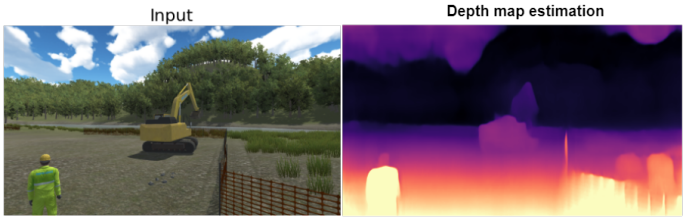


Figure 9. Depth estimation on the simulated dataset. On the left, there is the RGB frame, and on the right, there is the corresponding depth map frame generated using monodepth2's GitHub implementation (https://github.com/nianticlabs/monodepth2 (accessed on 7 May 2024)).

Table 2. Dataset classes' definition as provided in [1].

Actions	Definition
Other	No dangerous actions to be notified
Cabin-Worker	Worker moving too close to the cabin while the excavator is being operated
Bucket-Worker	Worker moving under the bucket, in danger of getting hit, or materials may fall from the bucket
Track-Excavator	The excavator moving forward to the tracks (active railway line or electric wires)

3.2. Problem Formulation

For an easy understanding, this problem can be formulated mathematically as follows: let I_{rgb} be an input sequence of N RGB frames, and I_{depth} be an input sequence of N depth map sequence, such that both inputs are synchronized and captures the same scene simultaneously. Each modality stream is fed into its corresponding modality encoder g, consisting of pre-trained backbone with frozen weights. This encoder is used to extract spatial and temporal features Z_{rgb} and Z_{depth} calculated as follows:

$$g_{rgb}(I_{rgb}) = Z_{rgb} \tag{1}$$

and

$$g_{depth}(I_{depth}) = Z_{depth} \tag{2}$$

The extracted visual features are then fused using a fusion operation denoted as $\oplus$, resulting in a multimodal representation vector Z such that $Z = z_{rgb} \oplus z_{depth}$, where $\oplus$ denotes a fusion operation (e.g., concatenation). This vector is then fed into a one-layer classifier that performs multi-class dangerous action prediction. It is worth mentioning that element-wise operations used in this comparison study (addition, multiplication, max, and average) do not need feature vectors z to be down-spaced since both RGB and depth encoders have the same latent feature dimensions. The training strategy adopted in this work is illustrated in Figure 10.

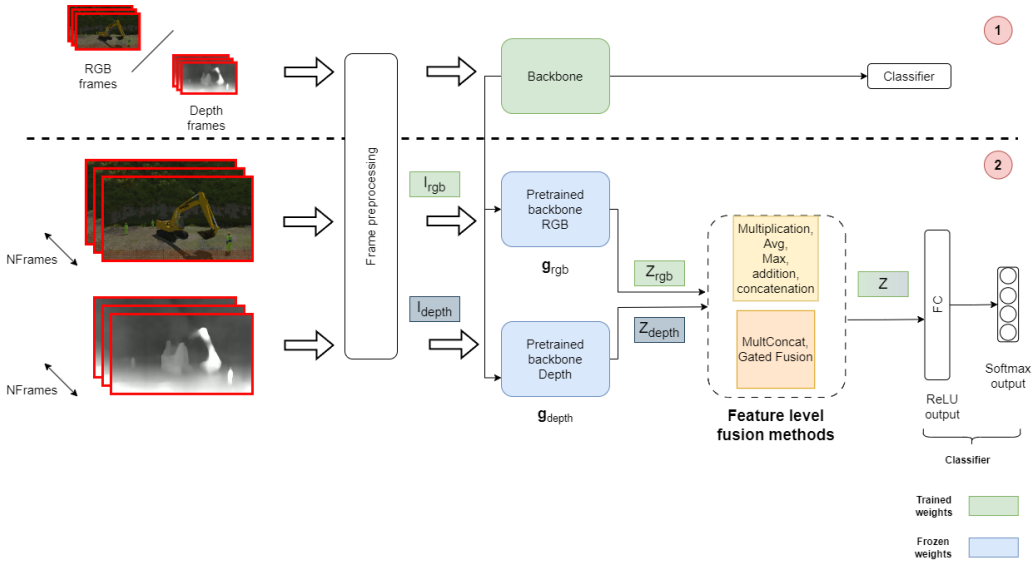


Figure 10. This figure illustrates the complete data flow and components involved in the proposed method, including preprocessing, feature extraction, and classification. The training strategy is divided into two training steps: the first one consists of training the unimodal action recognition classifiers on dangerous action recognition, then the backbone weights are frozen and used as feature extractors in the next step. RGB and depth frames are preprocessed and passed through separate pre-trained backbones (g_{rgb} and g_{depth}) to extract features (Z_{rgb} and Z_{depth}). These features are then fused using various feature-level fusion methods for comparison purposes. The fused features (Z) are fed into a classifier consisting of fully connected layers (FC) with a ReLU activation and Softmax output to perform the final action recognition.

3.3. Modality Encoders

This work compares three architectures which are well known for their superior performance on action recognition tasks:

- Slowfast: [71] SlowFast networks are designed for video recognition using RGB frames. It is divided into two main components: a slow and a fast pathway(see Figure 11). The slow pathway, operating at a low frame rate, allows one to capture spatial semantics, and the Fast pathway, operating at a high frame rate, is used to capture motion at fine temporal resolution. It was tested to achieve strong performance for both action classification and detection in video, large contribution, and state-of-the-art accuracy are reported on major video recognition benchmarks: kinetics [72], charades [73], and AVA [74]
- C3D: introduced by [75], utilizes 3D convolutional layers (3D ConvNets) to analyze video data, capturing spatial and temporal information simultaneously. This model processes video clips by applying filters across three dimensions (width, height, and time), enabling it to extract features from the sequences of frames for action recognition (see Figure 12). The architecture is straightforward, comprising repeated blocks of 3D convolutions followed by pooling layers, designed to work with fixed-length video segments, typically 16 frames. Despite its simplicity, C3D has shown effectiveness in various video analysis tasks compared to 2D ConvNets, benefiting from the ability to be pre-trained on large video datasets and fine-tuned for specific tasks.

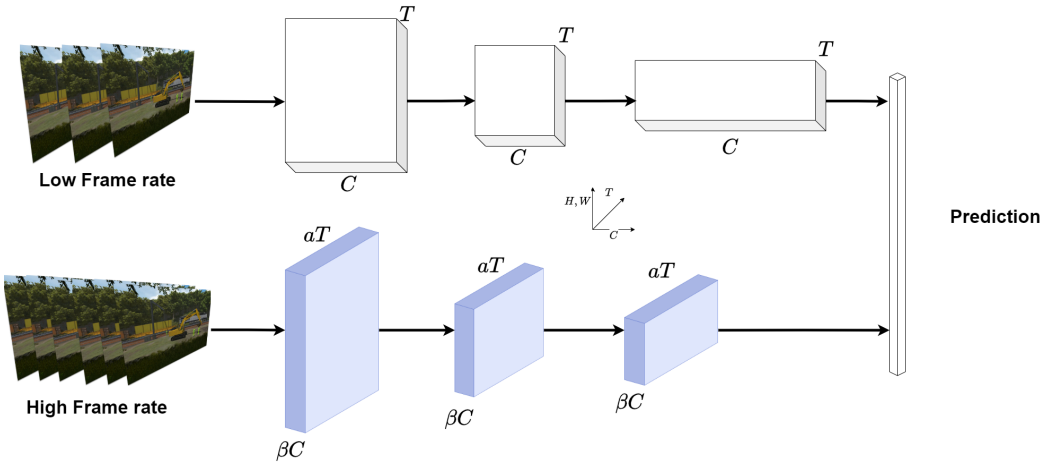


Figure 11. Slowfast architecture sourced from [71].

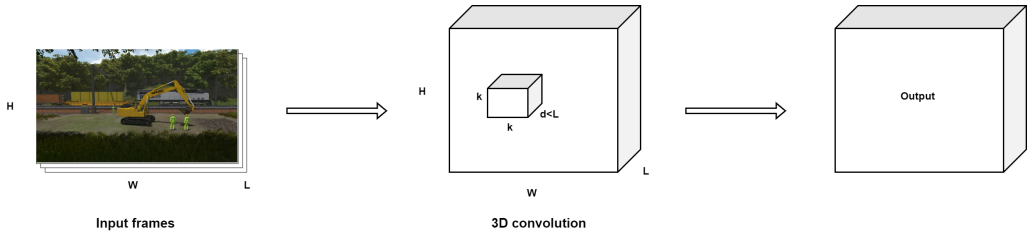


Figure 12. Three-dimensional convolution operations on videos, as initially introduced in [75] operate on video volumes and preserve the temporal information of the input signal.

- TimesFormer: [76] a convolution-free approach for video classification that is built exclusively on self-attention over space and time. This architecture, which is based on Transformers, enables spatiotemporal feature learning directly from frame-level patches(see Figure 13). The authors claim that, despite the radically new design, TimeSformer achieves state-of-the-art results on several action recognition benchmarks such as Kinetics-400 [72] and Kinetics-600 [77].

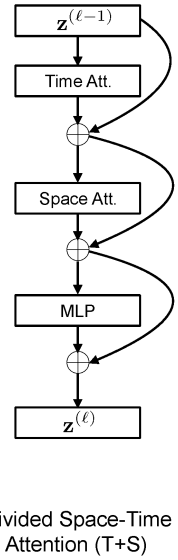


Figure 13. Architecture of timesformer used in this study that employs self-attention to capture spatiotemporal features using residual connections. Source [76].

3.4. Setup

The comparative analysis was performed using mmaction2's [78] public checkpoints for the modality encoders. For instance, slowfast_r50_256p_8x8x1_256e_kinetics400_rgb public checkpoint for Slowfast, and Timesformer (divided space+time attention) [76] public checkpoint divST_8xb8-8x32x1-15e_kinetics400-rgb (see Figure 13), both trained on Kinetics [72]. Finally, *c3d_sports1m-pretrained_8xb30-01-16x1x1-45e_ucf101-rgb* for C3D trained on UCF-101 [79].

The SGD optimizer [80] was used with a 0.001 learning rate for finetuning encoders and 0.01 for training multimodal models, both training steps where operating with a 0.9 momentum and a total of 50 epochs. An early stopping was set to 15 epochs to avoid overfitting. The feature extractors' weights were frozen during training as depicted in Figure 10. The models were trained and evaluated on the GPU of type NVIDIA GeForce GTX 1080 Ti with 11264 Mb of VRAM using the PyTorch framework (https://pytorch.org/ (accessed on 7 May 2024)). Finally, the metric used for evaluation was the accuracy score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In the next section, the results obtained following the methodology and the experimental setup detailed previously will be discussed.

4. Results and Discussion

Table 3 reports comparison results between fusion methods and stream encoders. First, it is noticed that slowfast [71] gives an overall better performance in both multimodal and unimodal training compared to c3d [75] and timesformer [76]. The best result was recorded with MultConcat, a fusion method that was initially introduced in [81] for HS code prediction. This proves its effectiveness in feature-level fusion and compatibility with the slowfast backbone. However, one of the main limitations of this method is its dependency on efficient modality encoders, since it has lower performance than unimodal models and the other fusion strategies when used with c3d and timesformer backbones. It is also important to note that all models perform better with RGB-only compared to depth-only settings. Indeed, Depth maps, while useful for understanding the three-dimensional structure of a scene, represent only one aspect of the visual data (the distance of surfaces from the camera) and lack the detailed appearance information contained in RGB images. Another reason for this decrease in accuracy could be attributed to the pre-training of encoders since they did not operate on similar data modalities.

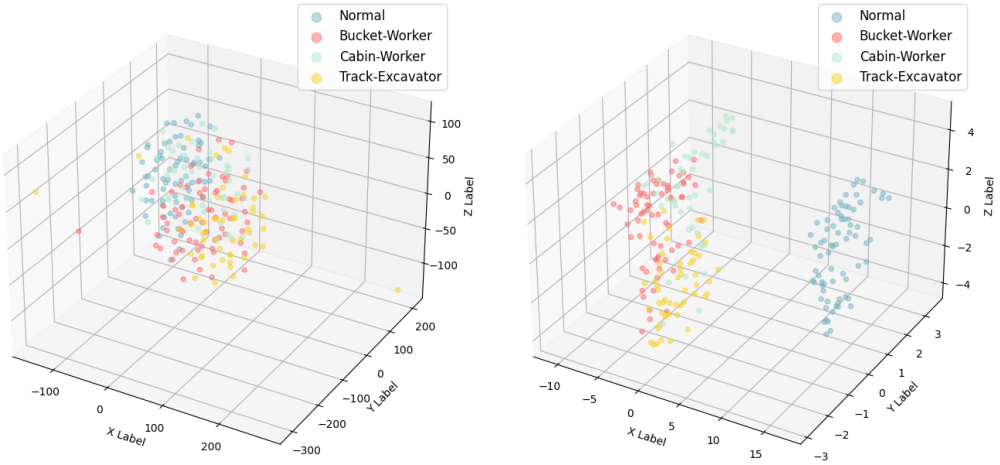
The next fusion methods showing moderate accuracy are addition and average. It is still worth noting that they perform less than unimodal solutions operating on RGB-only frames with c3d and timesformer, which can be attributed to the feature vector incompatibility for multimodal fusion compared to slowfast.

Table 3. Results on the simulated dataset of dangerous actions in railways construction by varying the backbones and fusion methods. Best result are in bold.

Backbone	Fusion Method	Modalities	Test Accuracy
C3d	Addition	RGB-D	0.741
	Concatenation	RGB-D	0.728
	Max	RGB-D	0.663
	Product	RGB-D	0.683
	Average	RGB-D	0.720
	GatedFusion [82]	RGB-D	0.22
	/	rgb only	0.806
	/	depth only	0.786
	MultConcat [81]	RGB-D	0.675
Timesformer	Addition	RGB-D	0.296
	Concatenation	RGB-D	0.296
	Max	RGB-D	0.399
	Product	RGB-D	0.428
	Average	RGB-D	0.383
	GatedFusion [82]	RGB-D	0.3868
	/	depth only	0.7572
	/	rgb only	0.7901
	MultConcat [81]	RGB-D	0.465
Slowfast	Addition	RGB-D	0.868
	Concatenation	RGB-D	0.860
	Max	RGB-D	0.860
	Product	RGB-D	0.860
	Average	RGB-D	0.876
	GatedFusion [82]	RGB-D	0.8477
	/	depth only	0.7984
	/	rgb only	0.8601
	MultConcat [81]	RGB-D	**0.893**

4.1. Embedding Visualization

The embedding space was visualized in Figure 14 using 3D t-SNE of the dangerous action categories including the normal one (no danger). The unimodal embeddings Z_{rgb} show some clustering of the dangerous action categories, but the separation between different classes is not very distinct. Specifically, the clusters for "Bucket-Worker", "Cabin-Worker", and "Track-Excavator" overlap significantly, making it difficult to differentiate between these categories based on RGB data alone.



(**a**) Unimodal embeddings (RGB) (**b**) Multimodal embeddings

Figure 14. Projection of multimodal and unimodal embeddings of test set observations using t-SNE algorithm.

In contrast, a clear pattern is observed with the multimodal embeddings Z: the model could separate the dangerous actions from the normal ones (safe actions). The inclusion of depth maps in addition to RGB frames allows the model to better distinguish between the actions. For instance, the "Bucket-Worker" and "Cabin-Worker" clusters are more distinctly separated in the multimodal embeddings, indicating that the additional spatial information provided by the depth maps helps in accurately identifying these actions. This improved separation in the multimodal embeddings highlights the effectiveness of the MultConcat [81] fusion method, which can better capture the nuances of each action when both RGB and depth information are utilized compared to the unimodal approach (RGB only).

4.2. Modalities Contribution

The contribution of the multimodal model (RGB-D) was quantified compared to the RGB-only model (unimodal). The goal is to isolate only these two modalities to measure the impact of the depthmaps modality. To achieve this, the accuracy scores were calculated for both models on the test set results, followed by computing the accuracy difference between the two. The result of this operation is illustrated in Figure 15. The contribution was calculated as follows:

$$C = Acc_{f:X \mapsto Y}(Depthmaps, rgb) - Acc_{g:X \mapsto Y}(rgb) \quad (4)$$

where $f : X \mapsto Y$ represents the multimodal model taking RGB and depth maps as input simultaneously, and $g : X \mapsto Y$ represents the unimodal model trained on RGB sequences only.

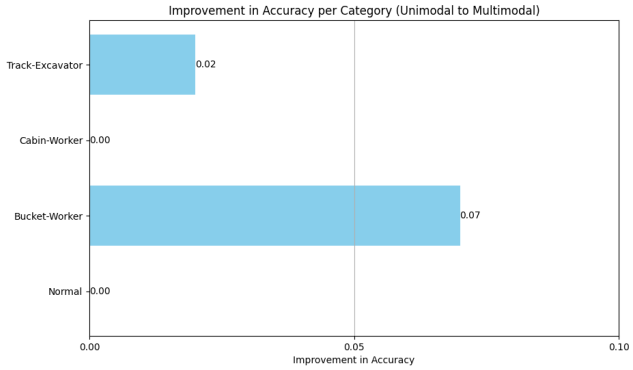


Figure 15. Multimodal contributions per action class.

The results indicate that the addition of depth information alongside RGB sequence contributes to varying degrees of accuracy improvement across action classes. Specifically, the "Bucket-Worker" class showed the most significant improvement with a 7% increase in accuracy, followed by "Track-Excavator" with a 2% increase. The categories "Normal" and "Cabin-Worker", however, showed no improvement. These results suggest that the depth maps modality contributes more significantly to categories where the additional depth information likely provides crucial distinguishing features not available in RGB data alone. Indeed, this information is crucial to perceive how far the worker is from the operating engines to determine whether there is a risk of collision.

5. Conclusions

This work addresses dangerous action recognition under a multimodal fusion framework. The experiments conducted through this study demonstrate the critical need for robust modality encoders, which are key components in extracting pertinent features. Equally crucial is the design of a fusion method that maintains the modality-specific characteristics, while simultaneously exploiting cross-modal interactions. In addition, the success of the fusion method MultConcat initially proposed in [81] was observed, yielding better

results than the other fusion methods tested as well as unimodal solutions operating on rgb frames only. It was also noticed that the depth modality improves the overall performance, enabling the model to easily discriminate dangerous actions from safe ones. Overall, this study highlights the advantage of multimodal fusion techniques in industrial settings. As a final note, the dependence on robust encoders could represent a limitation that merits further investigation in future work.

Author Contributions: Conceptualization, O.A., S.A.M., and X.S.; methodology, O.A., S.A.M., and X.S.; software, O.A.; investigation, O.A.; writing—original draft preparation, O.A.; visualization, O.A.; supervision, S.A.M. and X.S.; project administration, S.A.M. and X.S.; funding acquisition, S.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded through the collaborative expertise project between UMONS and Infrabel called "Project Field Worker Protection with AI". Infrabel manages and maintains Belgium's railway infrastructure. Their mission is to ensure the safe, reliable, and sustainable operation of the Belgian railway network.

Data Availability Statement: The dataset used in this study cannot be published for confidentiality reasons.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mahmoudi, S.A.; Amel, O.; Stassin, S.; Liagre, M.; Benkedadra, M.; Mancas, M. A Review and Comparative Study of Explainable Deep Learning Models Applied on Action Recognition in Real Time. *Electronics* **2023**, *12*, 2027. [CrossRef]
2. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef]
3. Liang, P.P.; Zadeh, A.; Morency, L.P. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv* **2022**, arXiv:2209.03430.
4. Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; Huang, L. What makes multi-modal learning better than single (provably). *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10944–10956.
5. Liang, P.P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M.A.; Zhu, Y.; et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv* **2021**, arXiv:2107.07502.
6. Rahate, A.; Walambe, R.; Ramanna, S.; Kotecha, K. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Inf. Fusion* **2022**, *81*, 203–239. [CrossRef]
7. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 689–696.
8. Lin, A.S.; Rao, S.; Celikyilmaz, A.; Nouri, E.; Brockett, C.; Dey, D.; Dolan, B. A recipe for creating multimodal aligned datasets for sequential tasks. *arXiv* **2020**, arXiv:2005.09606.
9. Botach, A.; Zheltonozhskii, E.; Baskin, C. End-to-end referring video object segmentation with multimodal transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4985–4995.
10. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. In Proceedings of the International Conference on Machine Learning, Virtual, 8–24 July 2021.
11. Garcia, N.C.; Morerio, P.; Murino, V. Modality distillation with multiple stream networks for action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 103–118.
12. Joshi, G.; Walambe, R.; Kotecha, K. A review on explainability in multimodal deep neural nets. *IEEE Access* **2021**, *9*, 59800–59821. [CrossRef]
13. Boulahia, S.Y.; Amamra, A.; Madi, M.R.; Daikh, S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Mach. Vis. Appl.* **2021**, *32*, 121. [CrossRef]
14. Shutova, E.; Kiela, D.; Maillard, J. Black holes and white rabbits: Metaphor identification with visual features. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 160–170.
15. Fränti, P.; Brown, G.; Loog, M.; Escolano, F.; Pelillo, M. *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, 20–22 August 2014*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8621.
16. Li, G.; Li, N. Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network. *Electron. Commer. Res.* **2019**, *19*, 779–800. [CrossRef]
17. Che, C.; Wang, H.; Ni, X.; Lin, R. Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis. *Measurement* **2021**, *173*, 108655. [CrossRef]

18. Wang, W.; Tran, D.; Feiszli, M. What makes training multi-modal classification networks hard? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–16 June 2020; pp. 12695–12705.
19. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
20. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
22. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *arXiv* **2022**, arXiv:2206.06488.
23. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. Visualbert: A simple and performant baseline for vision and language. *arXiv* **2019**, arXiv:1908.03557.
24. Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; Parikh, D. Pythia v0. 1: The winning entry to the vqa challenge 2018. *arXiv* **2018**, arXiv:1807.09956.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
26. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
27. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention Bottlenecks for Multimodal Fusion. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Virtual, 6–14 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: New York, NY, USA, 2021; Volume 34, pp. 14200–14213.
28. Sahu, G.; Vechtomova, O. Adaptive fusion techniques for multimodal data. *arXiv* **2019**, arXiv:1911.03821.
29. Pérez-Rúa, J.M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; Jurie, F. Mfas: Multimodal fusion architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6966–6975.
30. Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.J.; Fei-Fei, L.; Yuille, A.; Huang, J.; Murphy, K. Progressive neural architecture search. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 19–34.
31. Perez-Rua, J.M.; Baccouche, M.; Pateux, S. Efficient progressive neural architecture search. *arXiv* **2018**, arXiv:1808.00391.
32. Morshed, M.G.; Sultana, T.; Alam, A.; Lee, Y.K. Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors* **2023**, *23*, 2182. [CrossRef]
33. Singh, R.; Khurana, R.; Kushwaha, A.K.S.; Srivastava, R. Combining CNN streams of dynamic image and depth data for action recognition. *Multimed. Syst.* **2020**, *26*, 313–322. [CrossRef]
34. Wang, P.; Li, W.; Wan, J.; Ogunbona, P.; Liu, X. Cooperative training of deep aggregation networks for RGB-D action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
35. Avola, D.; Bernardi, M.; Foresti, G.L. Fusing depth and colour information for human action recognition. *Multimed. Tools Appl.* **2019**, *78*, 5919–5939. [CrossRef]
36. Shahroudy, A.; Ng, T.T.; Gong, Y.; Wang, G. Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1045–1058. [CrossRef] [PubMed]
37. Wang, H.; Song, Z.; Li, W.; Wang, P. A hybrid network for large-scale action recognition from rgb and depth modalities. *Sensors* **2020**, *20*, 3305. [CrossRef] [PubMed]
38. Yu, J.; Gao, H.; Yang, W.; Jiang, Y.; Chin, W.; Kubota, N.; Ju, Z. A discriminative deep model with feature fusion and temporal attention for human action recognition. *IEEE Access* **2020**, *8*, 43243–43255. [CrossRef]
39. Li, X.; Hou, Y.; Wang, P.; Gao, Z.; Xu, M.; Li, W. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 246–252. [CrossRef]
40. Xiao, X.; Ren, Z.; Li, H.; Wei, W.; Yang, Z.; Yang, H. SlowFast Multimodality Compensation Fusion Swin Transformer Networks for RGB-D Action Recognition. *Mathematics* **2023**, *11*, 2115. [CrossRef]
41. Liu, Z.; Cheng, J.; Liu, L.; Ren, Z.; Zhang, Q.; Song, C. Dual-stream cross-modality fusion transformer for RGB-D action recognition. *Knowl.-Based Syst.* **2022**, *255*, 109741. [CrossRef]
42. Cheng, J.; Ren, Z.; Zhang, Q.; Gao, X.; Hao, F. Cross-modality compensation convolutional neural networks for RGB-D action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1498–1509. [CrossRef]
43. Lee, S.; Woo, S.; Park, Y.; Nugroho, M.A.; Kim, C. Modality mixer for multi-modal action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 3298–3307.
44. Pareek, P.; Thakkar, A. RGB-D based human action recognition using evolutionary self-adaptive extreme learning machine with knowledge-based control parameters. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 939–957. [CrossRef]
45. Kumar, R.; Kumar, S. Survey on artificial intelligence-based human action recognition in video sequences. *Opt. Eng.* **2023**, *62*, 023102. [CrossRef]
46. Wang, C.; Yan, J. A comprehensive survey of rgb-based and skeleton-based human action recognition. *IEEE Access* **2023**, *11*, 53880–53898. [CrossRef]
47. Shaikh, M.B.; Chai, D. RGB-D Data-Based Action Recognition: A Review. *Sensors* **2021**, *21*, 4246. [CrossRef] [PubMed]

48. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *J. Real-Time Image Process.* **2016**, *12*, 155–163. [CrossRef]
49. Zhang, H.; Parker, L.E. 4-dimensional local spatio-temporal features for human activity recognition. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 2044–2049.
50. Kurakin, A.; Zhang, Z.; Liu, Z. A real time system for dynamic hand gesture recognition with a depth sensor. In Proceedings of the 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Piscataway, NJ, USA, 27–31 August 2012; pp. 1975–1979.
51. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 168–172.
52. Yu, G.; Liu, Z.; Yuan, J. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Proceedings of the Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, 1–5 November 2014*, Revised Selected Papers, Part V 12; Springer: Berlin/Heidelberg, Germany, 2015; pp. 50–65.
53. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
54. Oreifej, O.; Liu, Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 716–723.
55. Ni, B.; Wang, G.; Moulin, P. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1147–1153.
56. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef] [PubMed]
57. Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.C. Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2649–2656.
58. Tang, Y.; Wang, Z.; Lu, J.; Feng, J.; Zhou, J. Multi-stream deep neural networks for rgb-d egocentric action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3001–3015. [CrossRef]
59. Garcia-Hernando, G.; Yuan, S.; Baek, S.; Kim, T.K. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 409–419.
60. Moghimi, M.; Azagra, P.; Montesano, L.; Murillo, A.C.; Belongie, S. Experiments on an rgb-d wearable vision system for egocentric activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 597–603.
61. Wan, J.; Zhao, Y.; Zhou, S.; Guyon, I.; Escalera, S.; Li, S.Z. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 56–64.
62. Xu, N.; Liu, A.; Nie, W.; Wong, Y.; Li, F.; Su, Y. Multi-modal & multi-view & interactive benchmark dataset for human action recognition. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1195–1198.
63. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3d joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.
64. Sung, J.; Ponce, C.; Selman, B.; Saxena, A. Human activity detection from RGBD images. In Proceedings of the Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011.
65. Sung, J.; Ponce, C.; Selman, B.; Saxena, A. Unstructured human activity detection from rgbd images. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, St Paul, MI, USA, 14–18 May 2012; pp. 842–849.
66. Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T.L.; Samaras, D. Two-person interaction detection using body-pose features and multiple instance learning. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 28–35.
67. Liu, J.; Song, S.; Liu, C.; Li, Y.; Hu, Y. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM Trans. Multimed. Comput. Commun. Appl. (Tomm)* **2020**, *16*, 1–24. [CrossRef]
68. Masoumian, A.; Rashwan, H.A.; Cristiano, J.; Asif, M.S.; Puig, D. Monocular Depth Estimation Using Deep Learning: A Review. *Sensors* **2022**, *22*, 5353. [CrossRef] [PubMed]
69. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3828–3838.
70. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.

71. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
72. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
73. Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; Gupta, A. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 510–526.
74. Gu, C.; Sun, C.; Ross, D.A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6047–6056.
75. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
76. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In Proceedings of the ICML, Online, 18–24 July 2021; Volume 2, p. 4.
77. Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; Zisserman, A. A short note about kinetics-600. *arXiv* **2018**, arXiv:1808.01340.
78. Contributors, M. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmaction2 (accessed on 7 May 2024).
79. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
80. Robbins, H.; Monro, S. A stochastic approximation method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [CrossRef]
81. Amel, O.; Stassin, S. Multimodal Approach for Harmonized System Code Prediction. In Proceedings of the 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 4–6 October 2023; pp. 181–186. [CrossRef]
82. Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; González, F.A. Gated multimodal units for information fusion. *arXiv* **2017**, arXiv:1702.01992.

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