TeamAware

TEAM AWARENESS ENHANCED WITH ARTIFICIAL INTELLIGENCE AND AUGMENTED REALITY

Deliverable D3.2

Dataset report V2

Editor(s):	Thales SIX GTS France: Andreina Chietera ; TREELOGIC : Victor Medina, Victor Fernandez Carbajales; EUCENTRE : Ilaria Senaldi, Chiara Casarotti
Responsible Partner:	Thales SIX GTS France
Status-Version:	Final – v1.0
Date:	17/05/2023
Distribution level (CO, PU):	PU

Project Number:	GA 101019808
Project Title:	TeamAware

Title of Deliverable:	Dataset report
Due Date of Delivery to the EC:	01/05/2022

Workpackage responsible for the Deliverable:	WP3 – Visual Scene Analysis System
Editor(s):	THALES SIX GTS FRANCE SAS
Contributor(s):	TREELOGIC, EUCENTRE
Reviewer(s):	HAVELSAN, EUCENTRE
Approved by:	SIMAVI
Recommended/mandatory readers:	WP04, WP09, WP10, WP11, WP12, WP13

Abstract:	The aim of this task is to gather realistic datasets to train and test AI-based algorithms. In order to provide automatic analysis of the features of the visual scene acquired by the E/O systems, AI-based technologies will rely on an offline training process requiring large amounts of labelled representative data. For each target corresponding to the use cases and requirements, there will be a data collection specially identified for each scenarios and contexts.
Keyword List:	Victims detection, images segmentation, smoke & fire detection, structural damages.
Licensing information:	The document itself is delivered as a description for the European Commission about the released software, so it is not public.
Disclaimer	This deliverable reflects only the author's views and the Commission is not responsible for any use that may be made of the information contained therein

Document Description

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	13/02/2022	TOC	THALES
v0.2	15/02/2023	Preliminary content to chapter 2	THALES
V0.3	24/02/2022	Drafted contributions to chapter 3	TREE,
V0.4	28/02/2022	Contributions to chapter 4.	EUCENTRE
V0.5	01/04/2022	Content for the introduction and conclusion	THALES
V0.6	25/04/2022	Feedback for the review points	THALES, TREE
V1.0	17/05/2023	Deliverable submitted	SIMAVI, HAVELSAN

Table of Contents

List of Figures

List of Tables

Terms and abbreviations

AFO	Aerial Dataset of Floating Objects
AWS	Amazon Web Service
AI	Artificial Intelligence
CCTV	Closed Circuit Television
CNN	Convolutional Neural Network
EC	European Commission
DL	Deep Learning
FR	First Responder
HAR	Human Activity Recognition
HMDB	Human Motion Database
IMS	Infrastructure Monitoring System
IR	Infrared
LADI	Low Altitude Disaster Imagery
LEA	Law Enforcement Agency
ML	Machine Learning
MPII	Max-Planck Institute for Informatics
PWR	Passive Wi-Fi Radar
RF	Radio Frequency
RGB	Red Green Blue
RGB-D	Red Green Blue with Depth Sensor
SDR	Software Defined Radio
UAV	Unmanned Aerial Vehicle
USTC	University of Science and Technology of China
UWB	Ultra Wide Band
VSAS	Visual Scene Analysis System
WP	Work Package

Executive Summary

As, Computer vision or more in general Deep learning may require larger amounts of training data to perform well data management issues including how to acquire large datasets and how to improve the quality of large amounts of existing data become more and more relevant [1]. If the data used in artificial intelligence (AI) training is not sufficiently diverse, well balanced, appropriate to the context and unbiased, problems such as artificial "AI bias" may arise.

Accurate data collection techniques in the era of Big data gives motivation to conduct as a first step a comprehensive survey of the data collection literature on different tasks appropriate to the TeamAware project.

The second version of the dataset report has the aim to complete the overview about the open source datasets useful to train model adapted to manmade and natural disaster scenarios. In order to cover the end-user operational requirement 3 main applications are expected to be covered by the Visual Scene Analysis System (VSAS) system covering AI based technologies:

1. Potential victim detection
2. Situational awareness using visual segmentation
3. Damage assessment

According to these 3 main tasks, D3.2 proposes a survey on exploitable dataset in the context of the TeamAware project.

The consortium has made use of synthetic and composite datasets, as well as paired visible-infrared datasets, in order to support the VSAS development. While these datasets may not fully capture the complexity of real-world scenarios, they provide a valuable starting point for developing and testing these algorithms.

The collection of these datasets are presented in this deliverable.

The organisation of the deliverable is as follows:

- Chapter 1: context of this deliverable into the overall project.
- Chapter 2: presents different datasets for victims detections
- Chapter 3: illustrates a collection of datasets for visual segmentations
- Chapter 4: presents a mixed datasets of real data and generated images for damage detections of infrastructure

1 Introduction

This document presents a list of possible public datasets that could be used in the training phase and in the evaluation process of some or all of the computer vision algorithms within the TeamAware project. The list of datasets presented here is not intended as an exhaustive and complete repository, but rather a collection of meaningful data that can be used to train AI based detectors, and serves as a valuable resource in advancing the development of algorithms that can enhance First Responders' capabilities for the demonstration of the scenarios in Figure 1 described for the project in WP2 and later improved in WP13.

The natural disaster event such as an earthquake, involving damaged buildings, human victims, gas leakage, and fire and smoke in an underground station, and in city centre, which cause a chemical explosion and leading to a civil unrest as human-made disaster.

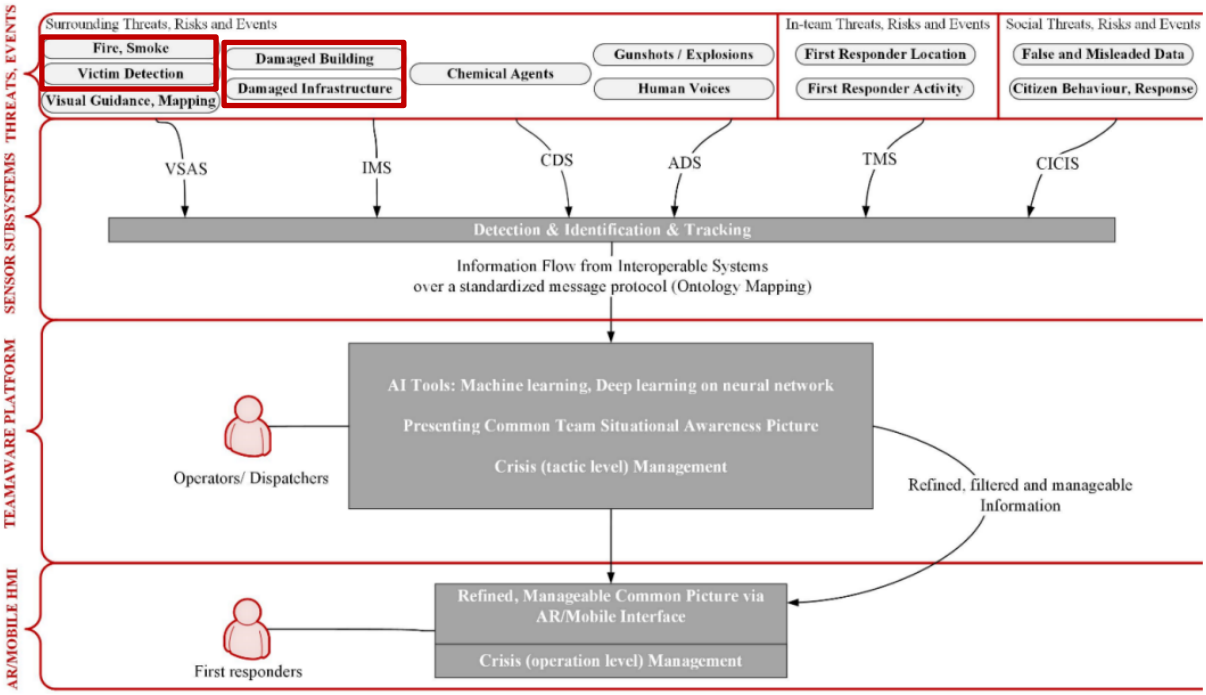

Figure 1. End users' operation in demonstration scenarios. The Red boxes present the main task the dataset of this report has to cover

The datasets presented in the D3.1, in the first phase of the project, were a tool to start evaluating the various algorithms/approximations in the state of the art while waiting for the final datasets.

To have a useful insight on the dataset adapted to the First Responders field, the WP3 team planned to merge appropriate datasets identified in this deliverable survey with synthetic or composite open-source data, rather than relying on real-world data acquisitions.

1.1 About This Deliverable

This first version of the deliverable presents an improved survey of the most suitable datasets for computer vision applications that can be leveraged to solve WP3 tasks, which includes human pose estimation, situational awareness using visual segmentation, and damage assessment. The identified datasets provide a valuable resource for developing deep learning algorithms that can improve First

Responders' ability to detect victims, detect smoke and fire, and assess damage in emergency situations. The deliverable also highlights the challenges faced by the project team in collecting real-world data and the approach taken to overcome these challenges by selecting and using semi-synthetic data.

1.2 Document Structure

The introduction and the executive summary described the needs of data to reach good quality performances for vision base inspection and situational awareness. In particular, these chapters stress the need of visual data in the field of First Responders applications. The introduction also presents the general procedure to follow to obtain new datasets in this domain leveraging, on well suited open-source datasets, pointing out the categories could be representative for visual recognition tasks in TeamAware project.

Chapter 2 presents different collection of public datasets devoted to victim detection. The aim is to use datasets commonly used for video surveillance, smart home assistance applications etc. As the problem of these datasets wants to solve is different from the victim detection, for only specific classes that can be adapted to the application.

Chapter 3 reviews datasets for image segmentation tasks for indoor and outdoor applications. The datasets presented have mainly an aerial point of view, as the TeamAware footages have a drone or a ceiling point of view coming from a CCTV system. The datasets illustrated in this deliverable have the task to map the element of the environment (indoor or outdoor) in an urban environment and detect fire or smoke.

Chapter 4 illustrates a dataset to detect building damages after an earthquake. This dataset merge data from earthquakes in Italy from 2009 to 2017. The aim is also to merge these data with a set of tailor-made data of artificial images, which is currently under development.

Finally the conclusion discusses this survey on datasets for deep learning algorithm training in the context of developing a dataset for First Responders application. The focus is on using existing datasets from different applications, including human pose estimation, visual segmentation, and damage assessment. However, due to the lack of real-world data, the project team had to rely on semi-synthetic data. The deliverable emphasises the importance of leveraging existing datasets to advance the development of algorithms that can improve First Responders' ability to perform their critical work in emergency situations.

1.3 Relation with Other Tasks and deliverables

This report is directly linked to the WP2 deliverables, looking to Figure 2, D3.1 is related to the scenario formulation elaborated in D2.4 and to D2.7 which presents the architecture and the solution selection of the overall TeamAware system. The data collection step of Figure 2 is an essential step to meet the objective of the scenario and their operational and functional requirements forecasted for the VSAS system.

The mapping between the VSAS system and operational requirements, developed in the context of Task 2.2 (“*End-users’ needs, requirements, constraints and scenarios*”), is used as the main entry to support Task T3.1. The identification of the most relevant scenarios for severe disasters together with the end users of the technological partners involved in WP3 and WP4 providing the technological solutions for the visual assessment, as used as the support the reach the goals of this report.

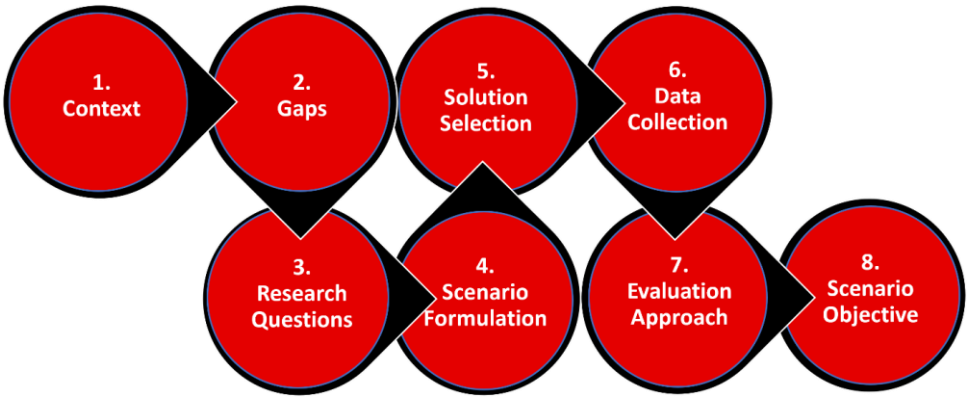

Figure 2 : TeamAware methodology and steps to reach the scenario objectives

2 Survey on Rescue Victim Datasets for Computer Vision applications

Deep learning is a popular method for human detection in the images and videos footprints, and it is applied extensively to the field of pedestrian detection, surveillance or smart home applications. The literature scarce for AI based applications in First Responders (FRs) and Law Enforcement Agency (LEAs) assistance fields. One reason behind this, is the unavailability of public datasets in this domain. As there is not a public dataset adapted to victim detection task, the aim is here to present various datasets, mainly used for human activity detection or pose estimation tasks, that could be filtered using only suitable categories and merge together in order to detect victims (i.e. people and animals) in an injured site.

The aim is to create a dataset for research and rescue applications, such as the tasks proposed in the TeamAware project, that has many defined categories in an attempt to cover all the possible situations and cases that the machine learning approaches (detectors, classifiers, Convolutional Neural Networks (CNN), etc.) may face while classifying the images sent from the injured site.

The categories to detect victims include images of people in various poses: lying down, sitting, and falling etc. from different distances and different angles, where some images would show the whole human body, and others would show parts of the human body. In addition, there would be images captured in a controlled environment and “in the wild” conditions. Finally, the images containing other spaces and rooms with no human would be gathered.

2.1 Relevance to Project Scenarios

There are two distinct scenarios which are explained with details in D2.4 namely, the natural disaster leading to also human-made disaster scenarios. VSAS will take part in both of the scenarios.

The VSAS system will be deployed in two phases:

1. A drone equipped with an RGB-D camera (Depth Sensor - RGB camera) will enter into the disaster zone for a first appreciation of the situation and to look for stranded passengers. The video footprints are sent and will be analysed into the command-and-control station. Thanks to the AI algorithms, previously trained with a victim detection dataset, a first evaluation of the victims will be performed.
2. After this first overview, an LEA wearing a helmet with an infrared (IR) camera and a standard grey camera will enter in the underground tunnel to rescue victims and to inspect in deeper the environment: other victims could be hidden by debris.
3. The helmet and the drone will be deployed and it will use to identify injured people in a demolished abandoned building after a chemical explosion.

2.2 MPII Human Pose

Max-Planck Institute for Informatics (MPII) Human Pose Dataset, describe in [2], is a dataset for human pose estimation “in the wild”, in an indoor or outdoor context. It consists of around 25k images extracted from online videos. Each image contains one or more people, with over 40k people annotated in total. Among the 40k people samples, ∼28k samples are for training and the remainder are for testing. Overall, the dataset covers 410 human activities (Figure 3) and each image is provided with an activity label. Images were extracted from a YouTube video and provided with preceding and following un-annotated frames.

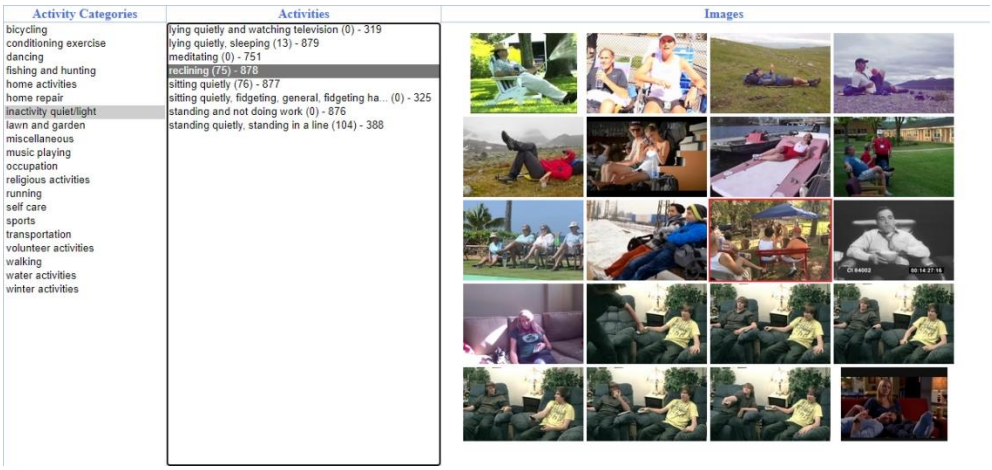

Figure 3 : Example of the inactivity category in the MPII dataset that could be used for victim detections. Different categories are included in the dataset. In the framework of TeamAware project only few categories as laydown, sitting etc. will be considered for victim detection purpose

2.3 UAV-Human

Unmanned Aerial Vehicle (UAV)-Human, described in [3], is a large dataset for human behaviour understanding with UAVs. It contains 67,428 multi-modal video sequences and 119 subjects for action recognition, 22,476 frames for pose estimation, 41,290 frames and 1,144 identities for person re-identification, and 22,263 frames for attribute recognition.

The dataset is interesting in the framework of TeamAware project because it presents data, which was collected by a flying UAV in multiple urban and rural districts in both daytime and night-time over three months, hence covering extensive diversities of subjects, backgrounds, illuminations, weathers, occlusions, camera motions, and UAV flying attitudes.

This dataset can be used for UAV-based human behaviour understanding, including action recognition (155 classes like sit down, stand up, move, walk…), pose estimation, re-identification, and attribute recognition. Furthermore, the drone is equipped with different sensors enabling the dataset to provide rich data modalities (Figure 4) including RGB, depth, IR, fisheye, night-vision, and skeleton sequences. This dataset fits very well to the sensors deployed on the TeamAware project, namely RGB-D, IR and standard grayscale camera in the of the CCTV system.

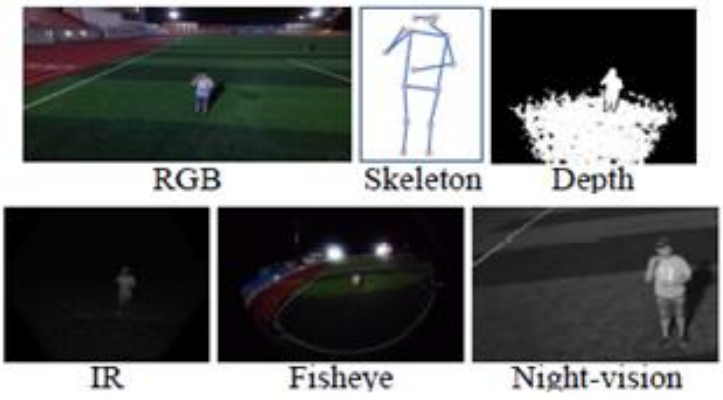

Figure 4 : Example of multisource data (multi modalities) using an UAV recorded in the UAV-Human Dataset

2.4 OPERAnet: A Multimodal Activity Recognition Dataset Acquired from Radio Frequency and Vision-based Sensors

The dataset presented in “Opportunistic Passive Radar for Non-Cooperative Contextual Sensing” OPERAnet [4] is a collection of Indoor Video data obtained with two Kinects and RF measurements. OPERAnet is a comprehensive dataset developed with the aim to evaluate passive Human Activity Recognition (HAR) and localisation techniques with measurements obtained from synchronised Radio-Frequency (RF) devices and vision-based sensors.

The RF data includes Channel State Information (CSI) extracted from a Wi-Fi Network Interface Card (NIC), Passive Wi-Fi Radar (PWR) built upon a Software Defined Radio (SDR) platform, and Ultra-Wideband (UWB) signals acquired via commercial hardware. The vision/Infra-red based data are acquired from Kinect sensors.

Approximately 8 hours of annotated measurements are provided, which are collected across two rooms from 6 participants performing 6 daily activities. This dataset can be exploited to advance Wi-Fi and vision-based HAR, for example, using pattern recognition, skeletal representation, deep learning algorithms or other novel approaches to accurately recognise human activities.

Furthermore, it can potentially be used to passively track a human in an indoor environment. It is suggested to be used for development of new algorithms and methods in the context of smart homes, elderly care, and surveillance applications.

2.4.1 Kinect dataset description

Focusing on the Kinect dataset: the Kinect directory collects files are in “.mat” format and each row in the files corresponds to three-dimensional skeleton information captured from each of the two Kinects at a given point in time:

- exp_no: experiment number which is specified as "exp_002", "exp_003", etc. Note that the Kinect system does not need background scan. Hence, background data for "exp_001" and "exp_019" are omitted for the Kinect data.
- timestamp: UTC+01 00 timestamp in milliseconds when the Kinect skeleton data were recorded.
- activity: ground truth activity labels. The activity is specified as a string of characters with no spacing e.g., "walk", "sit", "stand", "liedown", "standfromlie", and “bodyrotate". These correspond to the activity numbers 1, 2, 3, 4, 5, 6 and 7 in the "Details" respectively.
- person_id: person ID specified as "One", "Two", "Three", etc.
- room_no: room ID specified as "1" (left room) or "2" (right room).

2.5 HMDB: a large human motion database

HMDB (Human Motion DataBase) presented in [5], is a collection of clips from various sources, mostly from movies, and a small proportion from public databases such as the Prelinger archive, YouTube and Google videos. The dataset contains 6849 clips divided into 51 action categories (Figure 5), each containing a minimum of 101 clips. The actions categories can be grouped in five types as general facial actions, facial actions with object manipulation, general body movements including sit down, sit up, etc. body movements with object interaction, body movements for human interaction.

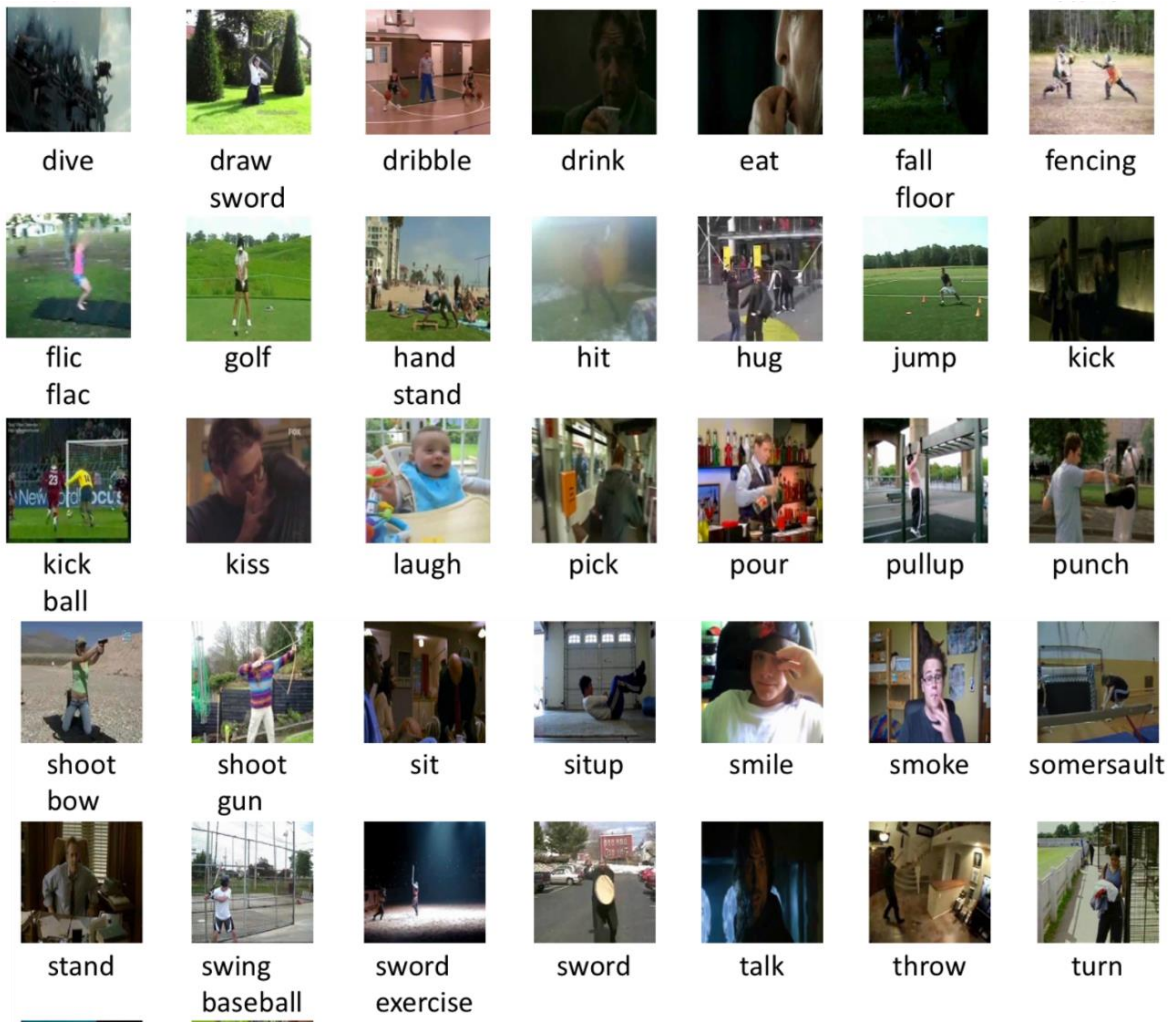

Figure 5 : Sample of data from HMBD: fall floor and sit categories could be used in the TeamAware project

2.6 Hollywood 3D: Recognising Actions in 3D Natural Scenes

The paper presented at IEEE conference in Computer Vision [6], focus on action recognition in unconstrained situations. This is a 3D dataset for action recognition in the wild. The detection and recognition of actions in natural settings is useful in several applications (Figure 6), including automatic video indexing and search, surveillance and assisted living.

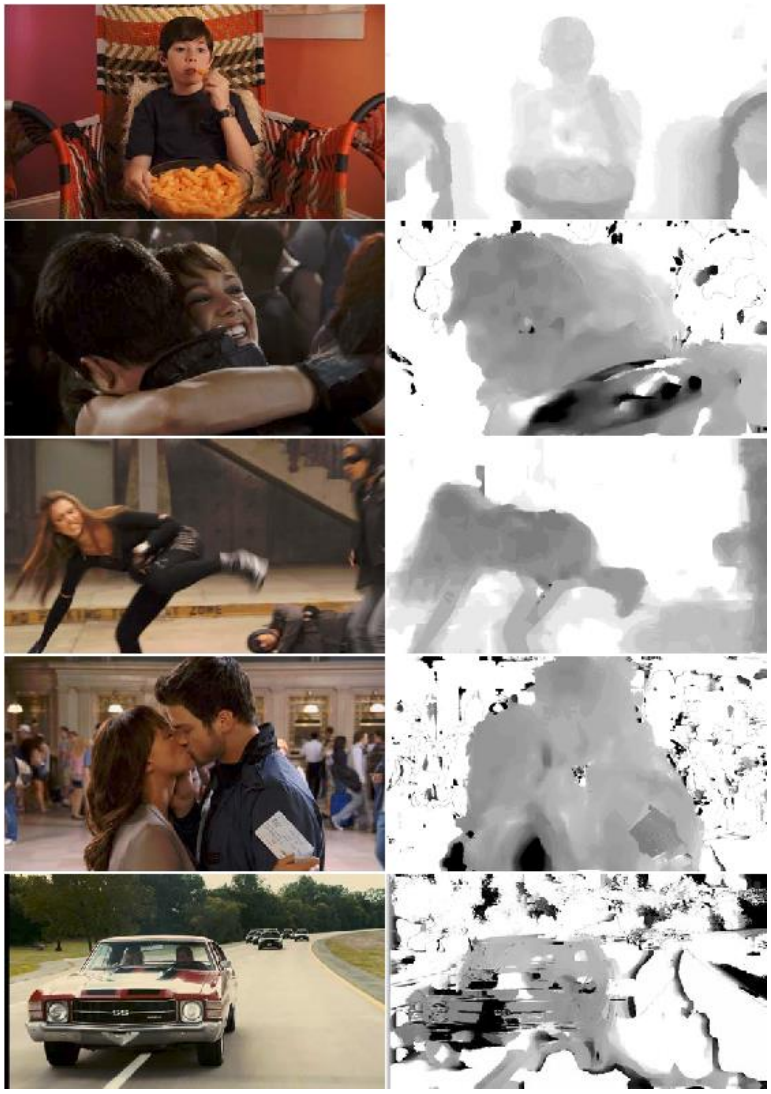

Figure 6 : Sample of data from Hollywood3D with the correspondent depth maps

This dataset presents a new level of complexity to the recognition community, arising from the natural within-class variation of unconstrained data, including unknown camera motion, viewpoint, lighting, background and actors, and variations in action scale, duration, style and number of participants. While this natural variability is one of the strengths of the data, the lack of structure or constraints make classification an extremely challenging task.

The Table 1 below shows the number of training and test sequences available for each action in the dataset, ensuring separate films are used for training and test data.

Table 1: Number of test and train sequences for each category in the Hollywood3D Dataset.

Action	*NoAction*	*Run*	*Punch*	*Kick*	*Shoot*	*Eat*	*Drive*	*UsePhone*	*Kiss*	*Hug*	*StandUp*	*SitDown*	*Swim*	*Dance*	Total
Train	44	38	10	11	47	11	51	21	20	9	22	14	16	45	359
Test	34	39	9	11	50	11	47	20	20	8	21	13	17	7	307

2.7 MSRDailyActivity3D

This dataset was introduced by Jiang Wang et al. in mining action let ensemble for action recognition with depth cameras [7]. DailyActivity3D dataset is a daily activity dataset captured by a Kinect device. There are 16 activity types (Figure 7): drink, eat, read book, call cell phone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down. If possible, each subject performs an activity in two different poses: “sitting on sofa” and “standing”.

The total number of the activity samples is 320. This dataset is designed to cover human’s daily activities in the living room. When the performer stands close to the sofa or sits on the sofa, the 3D joint positions extracted by the skeleton tracker are very noisy. Moreover, most of the activities involve the humans-object interactions.

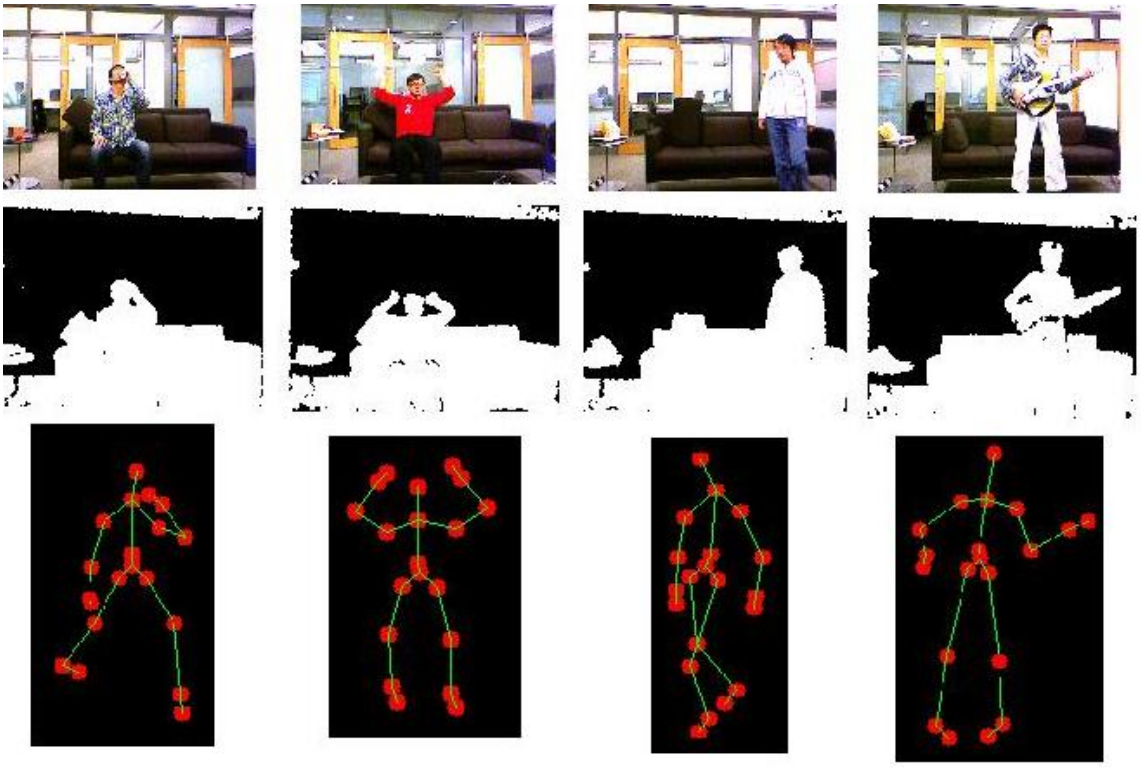

Figure 7 : Data sample with RGB, with related depth maps and skeleton sequences

2.8 VWFP: Virtual World Fallen People Dataset for Visual Fallen People Detection

This dataset [8] contains synthetic images of fallen people with the following classes: standing person, fallen person, and non-human object.

The dataset, as presented in Figure 8, is composed by images extracted from the highly photo-realistic video game Grand Theft Auto V developed by Rockstar North. Each image is annotated and labelled by the game engine providing bounding boxes of people present in the scene. The dataset is composed by 6,071 synthetic images depicting 7,456 fallen and 26,125 non-fallen pedestrian instances in various looks, camera positions, background scenes, lightning, and occlusion conditions.

Figure 8 : Sample of photo-realistic data from VWFP dataset.

2.9 FPDS Dataset

Similar to the previous dataset, this data [9] collection presents set of images of fallen people (see Figure 9) with the following annotated classes: fallen person and non-fallen person is identified with the purpose of this to help train machine learning models to detect potential victim lying on the floor.

“It consists of 6982 images, with a total of 5023 falls and 2275 non falls corresponding to people in conventional situations (standing up, sitting, lying on the sofa or bed, walking, etc). Almost all the images have been captured in indoor environments with very different situations: variation of poses and sizes, occlusions, lighting changes, etc.” [9]

The FPDS dataset is divided into three dataset: training, validation and perform tests.

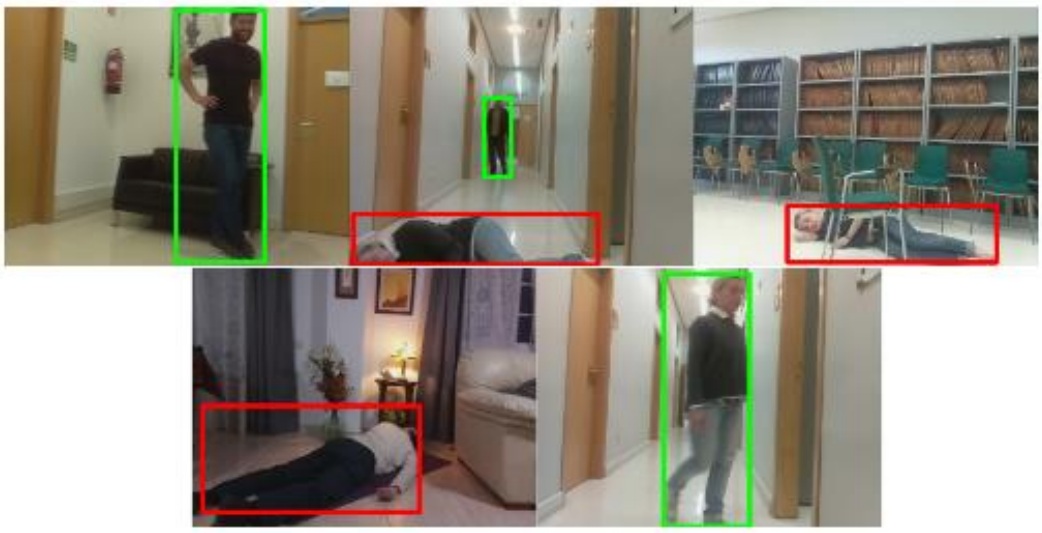

Figure 9 : Sample of standing and fallen people presented in the FPDS datasets.

2.10 VFP290K

This is a large-scale dataset [10], supported by the South Korean government, for the detection of fallen persons composed of fallen person images collected in various real-world scenarios. The Vision-based Fallen Person (VFP290K) dataset, sketched in Figure 10, consists of 294,713 frames of fallen persons extracted from 178 videos, including 131 scenes in 49 locations. The dataset also includes annotations for the location of the fallen person in each image.

This dataset was built to be exploited for different applications, including but not limited to CCTV surveillance, security, health care etc. For this reason the images are collected under diverse environments and in various situations in order to overcome existing datasets limits to only specific environmental conditions and lack diversity.

Figure 10 : Data overview of VFP290K dataset.

2.11 Harmonious Composite Images

This dataset [11] contains composite images of people with occluded body part in the debris.

Focused on the detection of potential victims using deep learning and differently from the other dataset cited in this deliverable that concerns specifically people shown in photos of daily life or sports activities. This set of images show people in the debris after a disaster where only parts of their bodies are exposed. In addition, because of the dust, the colours of their clothes or body parts are similar to those of the surrounding debris.

As images of disaster victims are extremely difficult to obtain, this dataset is obtain from a framework able to generate harmonious composite images for training. The framework allows to paste body parts onto a debris background to generate composite victim images and then it uses a deep harmonisation network to make the composite images look more harmonious.

This dataset, presented in Figure 11, is a result of the EU project INGENIOUS.

Figure 11 : Sample of the composite dataset coming from the project INGENIOUS

2.12 SCUT FIR Pedestrian Dataset

The SCUT FIR Pedestrian Datasets [12] is a large far infrared pedestrian detection in various scenarios and environments. “It consist of about 11 hours-long image sequences (~106 frames) at a rate of 25 Hz by driving through diverse traffic scenarios at a speed less than 80 km/h. The image sequences were collected from 11 road sections under 4 kinds of scenes including downtown, suburbs, expressway and campus in Guangzhou, China. It has 211,011 frames annotated for a total number of 477,907 bounding boxes and around 7,659 unique pedestrians, as shown in Figure 12.” [12]

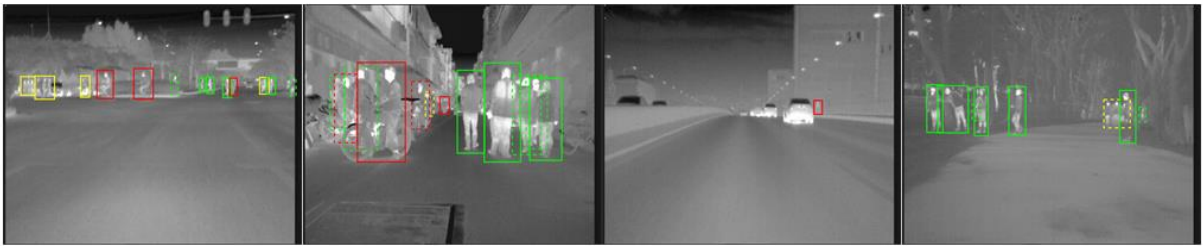

Figure 12 : Sample of the infrared pedestrian dataset called SCUT FIR

2.13 LLVIP: A Visible-infrared Paired Dataset for Low-light Vision

This dataset [13] contains paired visible and infrared images captured in low-light conditions, as presented in Figure 13. The dataset includes annotations for object detection, semantic segmentation, and edge detection. The purpose of this dataset is to help train and evaluate machine learning models for low-light vision applications. The dataset contains multiple classes, including pedestrians, vehicles, and background. This dataset contains 30976 images, or 15488 pairs, most of which were taken at very dark scenes, and all of the images are strictly aligned in time and space.

Figure 13 : Sample of infrared data and their visible version available in the LLVIP dataset.

2.14 FLIR Thermal Dataset for Algorithm Training

This dataset [14] contains thermal infrared images of vehicles and people captured from a moving vehicle. The dataset includes annotations for object detection and classification, as shown in Figure 14. The purpose of this dataset is to help train and evaluate machine learning models for thermal imaging applications. The dataset contains multiple classes, including pedestrians, bicycles, cars, and trucks. The dataset has a total of 26,442 fully annotated frames with 520,000 bounding box annotations across 15 different object categories.

Figure 14 : Sample of a FLIR image including pedestrians and cars.

2.15 Summary of Dataset for Victim Detection

This chapter presents several datasets that will be useful for the TeamAware project. Due to the diversity of sensors (RGB, depth, IR, skeleton information) and the diversity of points of view (frontal, human centric, UAV…), these datasets constitute a first varied and relevant dataset to develop and train models based on the human detection and people pose estimation**.**

After this significant part centred on finding data from databases and publicly available repositories that can help to solve the victim detection problem, it will be necessary to clean the data frames, work with multi-dimensional arrays, and manipulate data frames to aggregate data.

To deploy this data in the TeamAware framework, it will be necessary to building an efficient data architecture, streamlining data processing, and maintaining large-scale data systems.

Beside to that the road map for the deployment of the data imply the definition of the Deep Learning algorithm suitable for the analysis of this data.

The current algorithm aim is to detect people lying on the ground using image analysis, which was not possible to detect with the first approach that has been followed and described in deliverable D3.5.

The algorithm is currently composed of two separated stages: the first one is dedicated to person localisation, the second one classifies fallen persons from standing persons.

The algorithm has been tested on separated data and performs as expected in standard conditions. However, a loss in accuracy in challenging environments is expected such as collapsed buildings. To tackle such issue, detection algorithm is being trained with data representing these situations. The detection task is focused on in this case, rather than the classification task which appears to be less relevant in this type of environment. Due to the difficulty to get annotated data, synthetic images are being used.

3 Survey on Images Segmentation Datasets for VSAS

3.1 Switzerland Aerial Drone Footage Datasets

This dataset can be found at [15]. It contains 4K-quality videos captured with a DJI Mavic Pro drone [9] under different conditions, which are representative of real drone footage in terms of contents and camera angles. The videos have minimal post-processing (colour correction) and do not include any ground truth information, so they would have to be annotated, possibly including multiple object classes in each frame so that they can be used to train different detectors. Although the rich contents in these videos will provide valuable information for general object detection, training data will still be required for scenarios with more specific particularities.

3.1.1 Relevance to Project Scenarios

The dataset contains 13 bird-eye drone videos from natural scenery (snow mountains, forests, sea, roads, urban, etc.) with and without people in it, which can be useful to train reconnaissance and environmental mapping algorithms regarding the natural disaster scenario. Videos contain elements such as trees, mountains, rocks, birds, cars, etc., which can be used to train detectors for different elements (both natural and man-made). Figure 15 shows thumbnails from all the videos included in this dataset, where the relevance of its contents can be seen.

In case of an attack regarding human-made disaster, the first task to be performed is to evaluate the situation. Usually, photographs of the scene are required by analysts to perform such evaluations. In TeamAware, the use of drones, CCTV systems, and head mounted cameras will provide this information, so automatic computer vision algorithms must be able to analyse and map the environment before decisions can be taken.

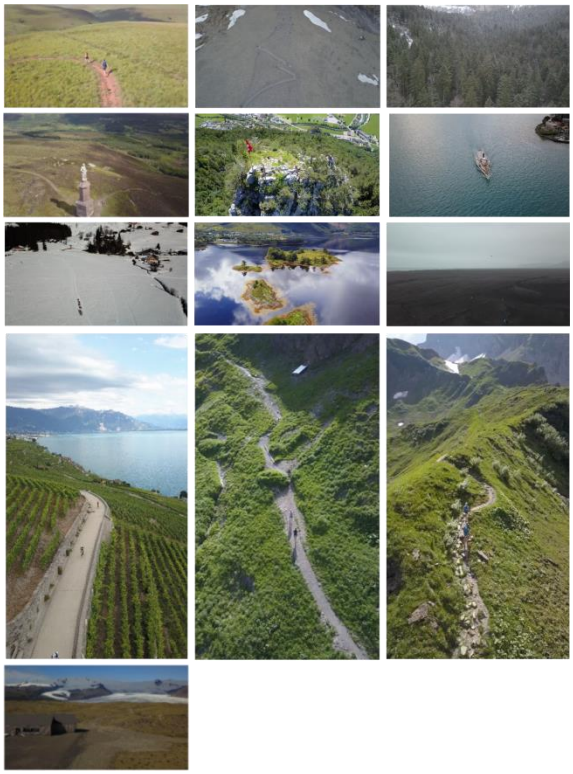

Figure 15: Switzerland Aerial Drone footage dataset thumbnails

3.2 USTC Forest Smoke and Fire Dataset

This dataset has been created by the University of Science and Technology of China (USTC) and it's specific for smoke and fire detection tasks with colour cameras. The full dataset is available at [16] and it requires authentication, but a reduced version has been made publicly available on Kaggle at [17]. The reduced version contains 15.800 images, but no ground truth data is provided, which means that all these images would have to be annotated.

The images are divided into four different sets:

- training sets for smoke detection (Figure 16) and fire detection (Figure 17) tasks
- 2 test sets (one large and one small) containing mixed fire and smoke images.

3.2.1 Relevance to Project Scenarios

The detection of fire and smoke is a task described in DOA and also requested by the end users, and USTC dataset can be a very good starting point to train the detection algorithms on. Figure 18 and Figure 19 illustrate some of the contents in the dataset, merging smoke and fire images for small and big test on the segmentation algorithms.

Figure 16: USTC Forest smoke and fire dataset. Smoke training set preview

Figure 17: USTC Forest smoke and fire dataset. Fire training set preview

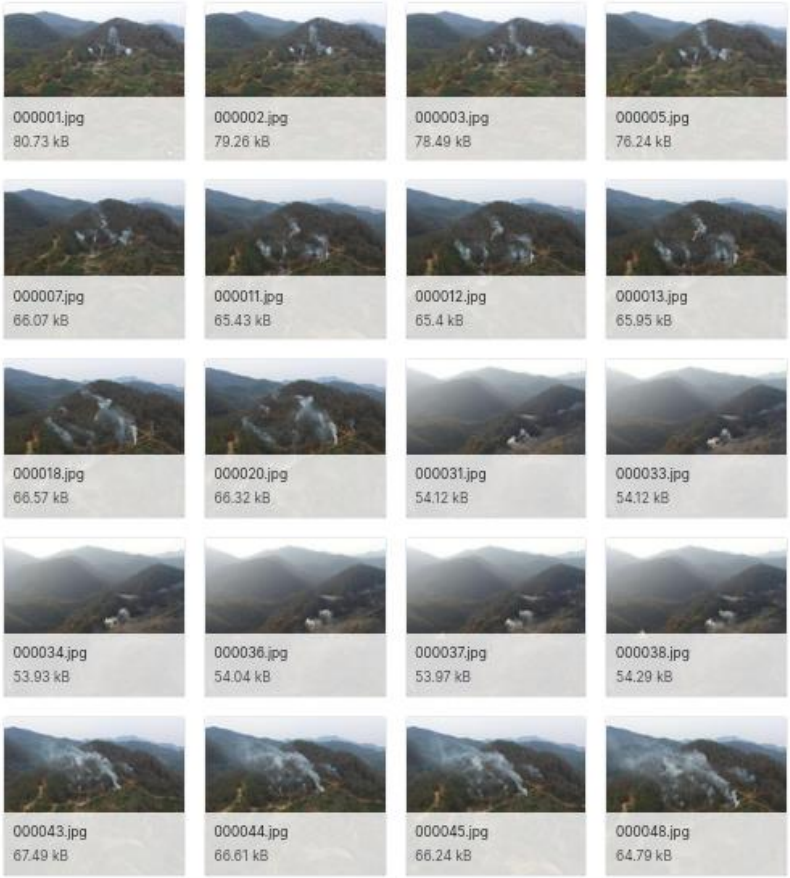

Figure 18: USTC Forest smoke and Fire dataset. "Big" test set preview

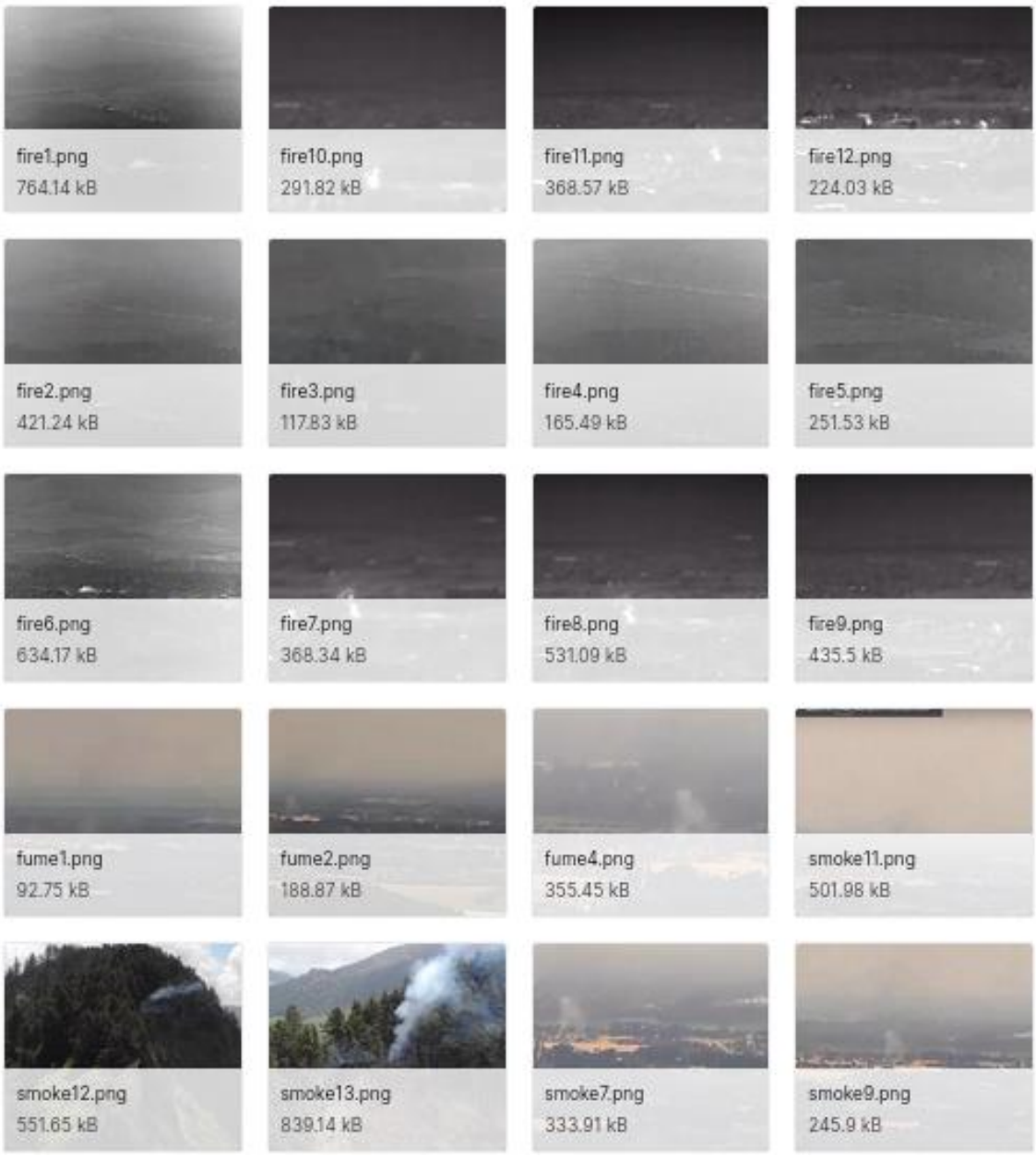

Figure 19: USTC Forest smoke and Fire dataset. "Small" test set preview

3.3 The Stanford Drone Dataset

This dataset has been created by the computational vision and geometry lab at Stanford University. The dataset consists of annotated videos of pedestrians, bikers, skateboarders, cars, buses, and golf carts navigating eight unique scenes on the Stanford University campus. It was created to analyse the behaviour of different types of agents when crossing paths in a real-world outdoor environment, to improve tasks like target tracking or trajectory forecasting. This dataset contains ground truth annotations which simplifies its use in TeamAware.

The original 69GB dataset can be found at [18], although a full compressed and optimised version of only 2GB is available on Kaggle [19]

3.3.1 Relevance to Project Scenarios

Although the original purpose of this dataset (a sample is shown in Figure 20) may not be of much direct use to the TeamAware project, the high number of buildings, streets, and other natural and structural elements found in many public spaces can be of great use for mapping and detection tasks in both of the project scenarios, especially for cases when disasters occur in outdoor public spaces like the ones depicted in this dataset. The downside to this dataset is that, to use it for training on different classes than the ones annotated in it will require additional annotation work.

Figure 20: Stanford drone dataset preview

3.4 Aerial Dataset of Floating Objects (AFO) Dataset

This is a dataset of aerial images for maritime search and rescue applications [20]. With over 40,000 hand-annotated elements in aerial-drone videos, this is the first free dataset of this type for training ML/DL models. It contains images from fifty video clips depicting objects floating on the water surface, captured by different drone-mounted cameras of different resolutions. The dataset has 3.647 images with 39.991 annotated objects, split into three parts: a training set (67.4%), a test set (19.12%), and a validation set (13.48%), where the test set is taken from specific unseen videos to avoid overfitting.

Attention must be paid to the license of this work, as the dataset is published under the Creative Commons Attribution-Non-commercial-Share Alike 3.0 License and so it cannot be used for commercial purposes or without attribution.

3.4.1 Relevance to Project Scenarios

This dataset is a useful resource to train algorithm in the detection of people in different positions like lying face up and face down, sitting, swimming, submerged from the neck down, etc. This could have applications not only in water scenarios but also in other different situations on the ground, both indoors and outdoors. Many of the images show subjects in groups, which can also be used to detect clusters of people in rescue operations, or to detect anomalous behaviour in terrorist scenarios.

Figure 21: Aerial dataset of floating objects. Set 1 preview

Figure 22: Aerial dataset of floating objects. Set 2 preview

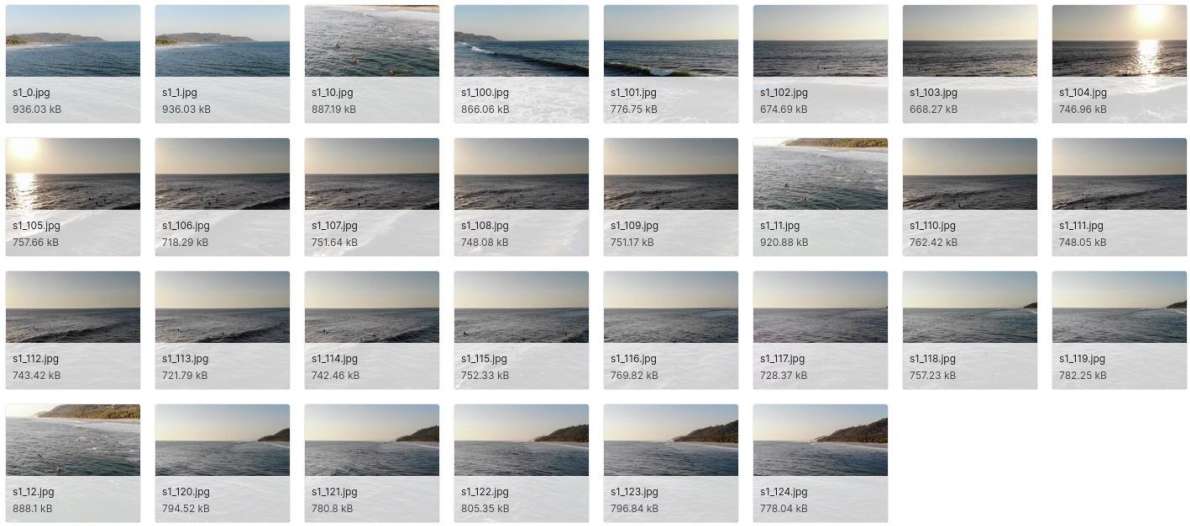

Figure 23: Aerial dataset of floating objects. Set 3 preview

3.5 Low Altitude Disaster Imagery (LADI) Dataset

The LADI dataset provides images that focuses on the Atlantic Hurricane and spring flooding seasons since 2015. Two key distinctions are the low altitude, oblique perspective of the imagery and disaster-related features, which (according to the authors) are rarely featured in computer vision benchmarks and datasets. The dataset uses a hierarchical labelling scheme of a five coarse categorical and then more specific annotations for each category. The five coarse categories are:

- Damage
- Environment
- Infrastructure
- Vehicles
- Water

For each of the coarse categories, there are 4-9 more specific annotations (**Error! Reference source not found.**).

Table 2: Hierarchical labels use in the annotation of LADI dataset.

Damage	Environment	Infrastructure	Vehicles	Water
damage (misc.)	dirt	bridge	aircraft	Flooding
flooding / water damage	grass	building	boat	lake / pond
Landslide	lava	dam / levee	car	Ocean
road washout	rocks	pipes	truck	Puddle
rubble / debris	sand	utility or power lines / electric towers		river / stream
smoke / fire	shrubs	railway		
	snow / ice	road		
	trees	water tower		
		wireless / radio communication towers		

Information and documentation on the dataset can be found on GitHub [21] as well as the dataset's page on AWS's open data registry [22]. The full dataset (327,2GB) can be downloaded from the project's AWS S3 bucket at [23].

Figure 24: Sample images from the LADI dataset

3.5.1 Relevance to project scenarios

The applications of this vast dataset to the TeamAware projects are numerous. From the detection of several landmarks (lakes and other bodies of water, green areas, forests, etc.), houses, mountains, etc., to locating landing sites for responders, roads, tracks, access points and many other areas that can be of interest during search and rescue operations.

3.6 Indoor Semantic Dataset (2D-3D segmentation)

The dataset [24] includes semantic and geometric data in 2D, 2.5D, and 3D domains, as well as their instance-level annotations. The dataset consists of about 70,000 RGB images, along with the corresponding depths, surface normal, semantic annotations, global XYZ images, as well as camera information. Apart from these images and information, there are also point clouds and raw 3D meshes registered and semantically annotated. In itself, this dataset allows the development of joint and intermodal learning models and potentially unsupervised approaches that use the regularities present in large-scale interior spaces.

3.6.1 Relevance to project scenarios

This dataset collects on 6 large-scale indoor areas originating from 3 different buildings of primarily office and educational use. These data are clearly useful in any of the scenarios proposed by the project (natural or man-made disaster), since in both cases, it is necessary to recognise the interior spaces of buildings to plan necessary rescues.

The contained data is displayed in two modes: 3D and 2D. In 3D mode, the data contains coloured point clouds and textured meshes for each scanned area. 3D semantic annotations for objects and scenes are provided for both modalities, with corresponding point-level and face-level labels. Annotations were initially done on the point cloud and projected onto the surfaces of the models, as well as indicating the bounding boxes of the objects in the scene. In the 2D modality, the data included is the RGB and depth raw images in full high definition with a resolution of 1080x1080, as well as their annotations of the objects present in them. Examples of these annotations can be seen in Figure 25 (2D data) and Figure 26 (3D data).

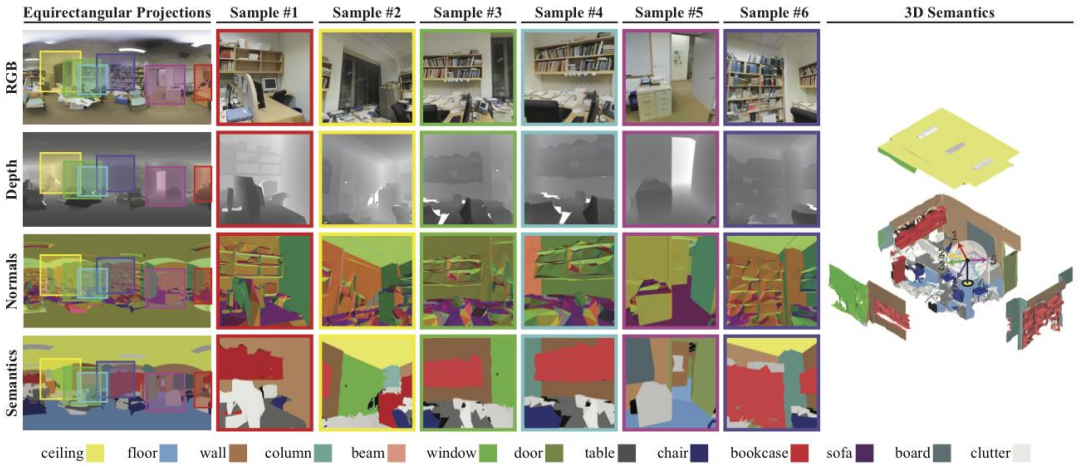

Figure 25: Sample images from the 2D modalities catalogue of the dataset.

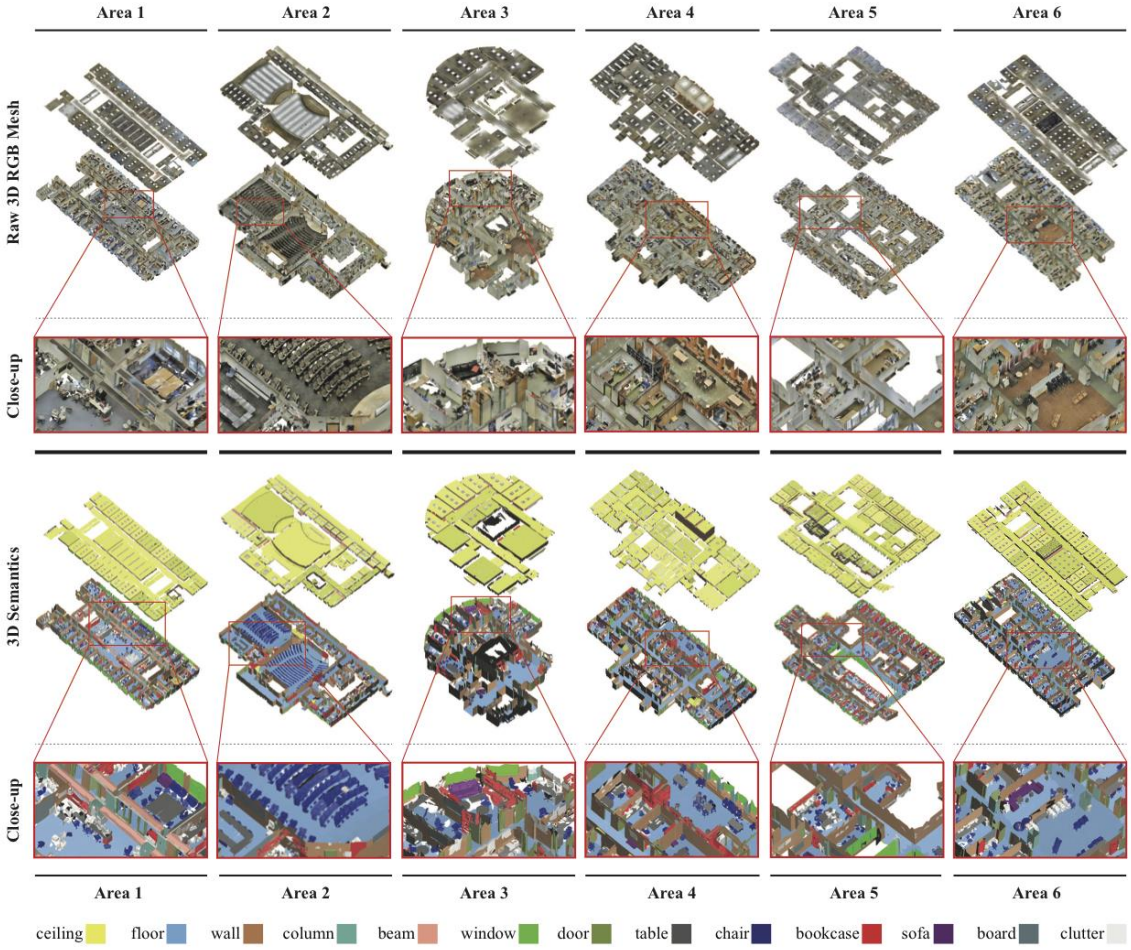

Figure 26: Dataset in 3D modalities includes: the textured 3D mesh models, semantic annotation and their point clouds

3.7 Amazon Web Service Open Data Registry

The Amazon Web Service (AWS) Open data registry hosts around 320 searchable datasets for specific tasks, 45 out of which are specific to disaster recovery alone. Most of these datasets provide landmark satellite imagery of different types, mostly used for agriculture purposes, which could be useful for certain detection and mapping tasks in TeamAware.

3.8 Flood image DataBase System (FloodDBS)

This dataset (FloodDBS) [18] consists around 9000 images of flood, mainly, in outdoor scenes. These images are collected through several sources like Twitter, YouTube, and the South Carolina Department of Transportation (SCDOT). The data are focused on the classification of images that present floods, showing urban environments (mainly) along with vehicles, signs, buildings, etc. Images are high resolution and have been standardised to 1080p resolution.

3.8.1 Relevance to project scenarios

Floods, along with fire and smoke, are one of the main events to be detected within the scope of TeamAware. In addition, the interaction of these elements together with other classifications/detections/segmentations will be key to infer by the TeamAware system of hazards, for example the detection of people in flooded environments. This data set must be re-annotated since it is focused on classification and not detection and/or segmentation. Examples of images that compose FloodDBS are show in Figure 27.

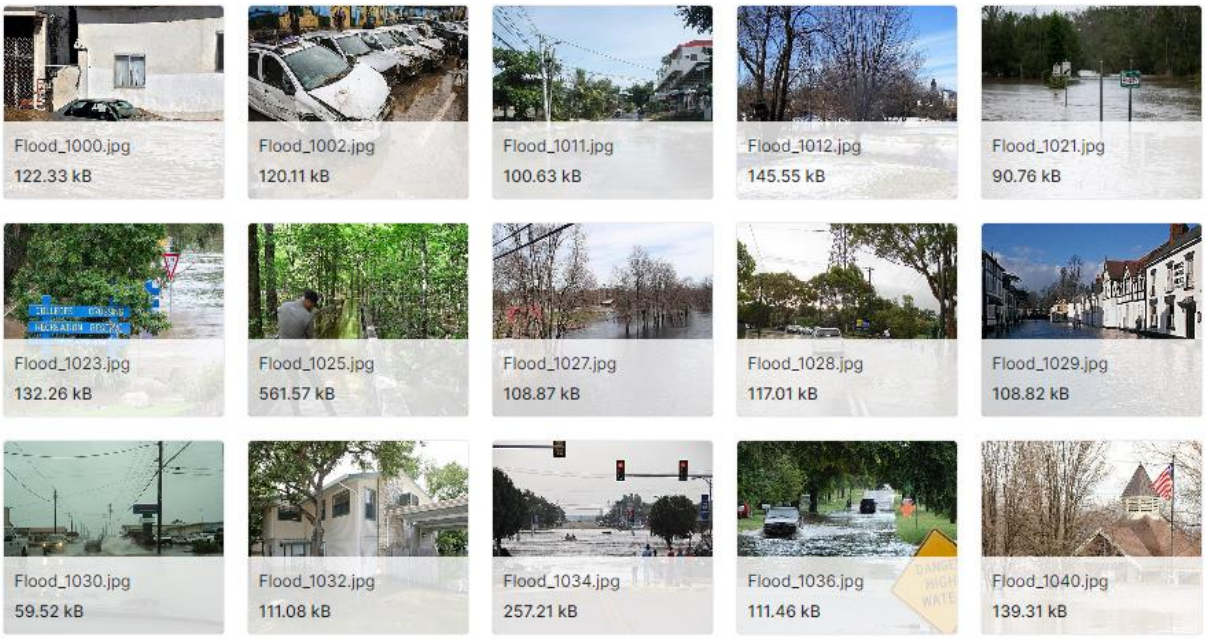

Figure 27: Examples of outdoor flood images of FloodDBS

3.9 Normal Use Object Datasets

Objects for daily use are a large set of objects, but considering the tasks/scenarios proposed by TeamAware, the objects that imply the presence of people in the vicinity of transport systems are being focused on, as well as items that can contain dangerous elements.

A single data set has not been found to cover these cases; four different sets have been brought together to formalise it:

- Suitcase/Luggage Dataset Indoor Object Image [26]: This dataset is an extremely challenging set of over 7000+ original Suitcase/Luggage images captured and crowdsourced from over 800+ urban and rural areas. 99% of images have a HD resolution and above.
- ImageCV Suitcase dataset [27]: This dataset contain 432 images labelled of suitcase. Close a far view of suitcases with different backgrounds and typologies. All images have a HD resolution.
- Backpack Image Dataset [28]: With a total of 500 images (400/50/50 – train/valid/test) of backpack is one of most complete datasets about this type of object. All images have at least HD resolution and contain high variability. Additionally, present more than 10 different codifications of the annotation as well as the necessary data augmentation.

- Mobile Phone Dataset [29]: This dataset has a set of over 3000 original mobile phone images captures in many environments and with various damage states. All the images are of high resolution, and it is the only one with a great variety of damages and defects.

3.9.1 Relevance to project scenarios

As previously introduced, these everyday items, especially focused on people who use transportation, are of special interest to first responders. Especially, this group of data sets allows to cover the training need of deep learning algorithms for the detection and segmentation of these elements. Additionally, the union of these several data subsets implies a normalization of annotations between them, as well as the transcoding of the current ones to a common format (COCO). The Figure 28 shows examples of the images included in mobile phone and suitcase dataset of Kaggle[19].

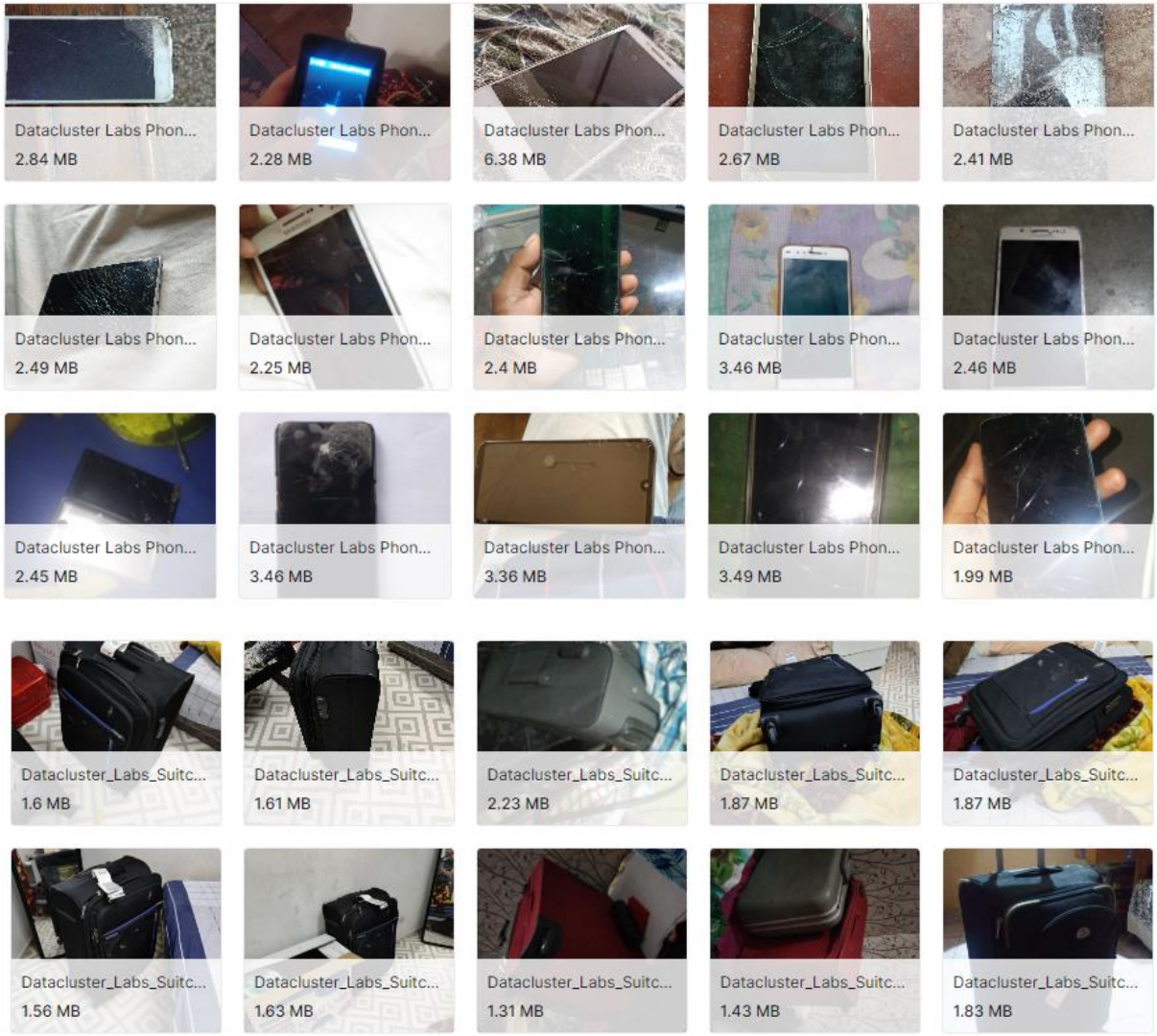

Figure 28: Examples daily object (mobile phones) and travel objects (luggage, suitcase, and backpack)

3.10 Oxford Pet Dataset

The Oxford Pets dataset [30] is a collection of images and annotations labelling various breeds. There are approximately 100 examples of each of the 37 breeds. This dataset contains the object detection portion of the original dataset with bounding boxes around the animals' heads. 80% of the samples in the data set are from cats and dogs, since they are the most common pets in households. The advantages of this data set are that it has high-resolution images and it is annotated for detection and segmentation.

3.10.1 Relevance to project scenarios

Animals, along with people, are part of the objectives indicated in the DOA, so this data set focused on common animals in indoor spaces is essential to train the detection and segmentation of these objects within the project. Examples of the images and annotations of the data set can be seen in the following figure.

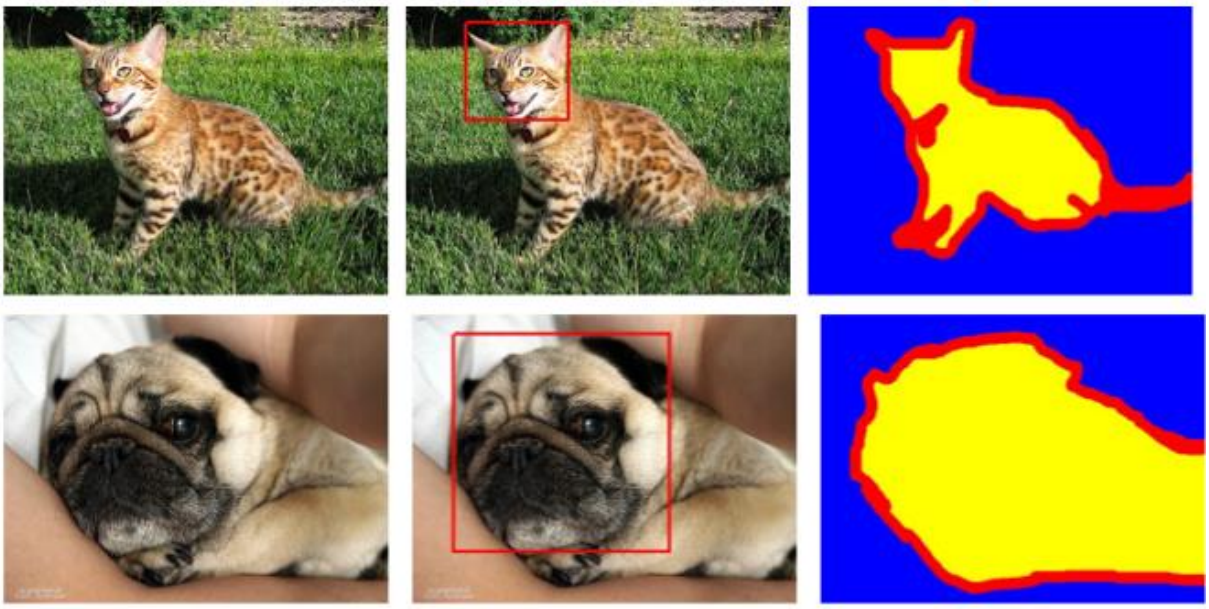

Figure 29: Domestic animals/Pet dataset of Oxford University

3.11 People (Adult & Children) Dataset

This dataset [31] is focused on the differentiation of people into adults and children. The detection of people is a task surpassed in the state of the art, but the robust classification between age groups is not. This set contains 240 high-quality adult and child images, with an associated rating annotation. Images are high resolution and normalised.

3.11.1 Relevance to project scenarios

The detection and categorization of people is key within the project. One of the important categorizations for first responders is knowing the age range of victims or people present in order to prioritize or direct certain tasks. This dataset allows us to train a classifier of the people detected and therefore include it in the list of project objects.

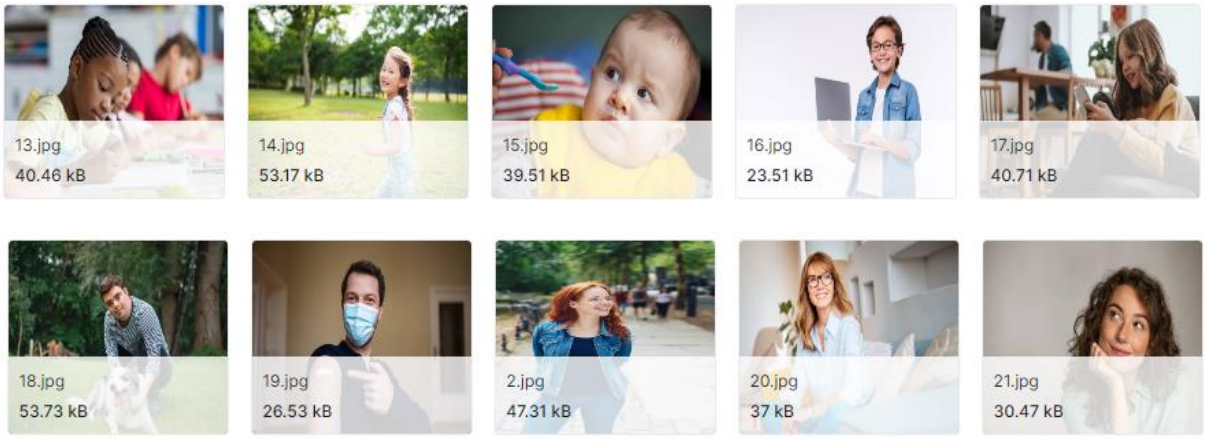

Figure 30: Examples of adult/child images

3.12 ADE20K Dataset

The ADE20k [32] dataset contains about 25,000 high-resolution images annotated with various semantic concepts, with about 2,700 different types of objects (including 500 subparts thereof). These images incorporate both indoor and outdoor scenes.

3.12.1 Relevance to project scenarios

The classes included within this dataset are very diverse, but it does have some subsets that are highly relevant to end users. From structural elements of the house (doors, windows, chimneys, staircases etc.), objects of common use indoors (clothes, garbage cans, etc.), clothes, vehicles, even electrical appliances. Many of these classes are relevant to end users and therefore cover many needs for training neural networks for detection and segmentation within WP3.

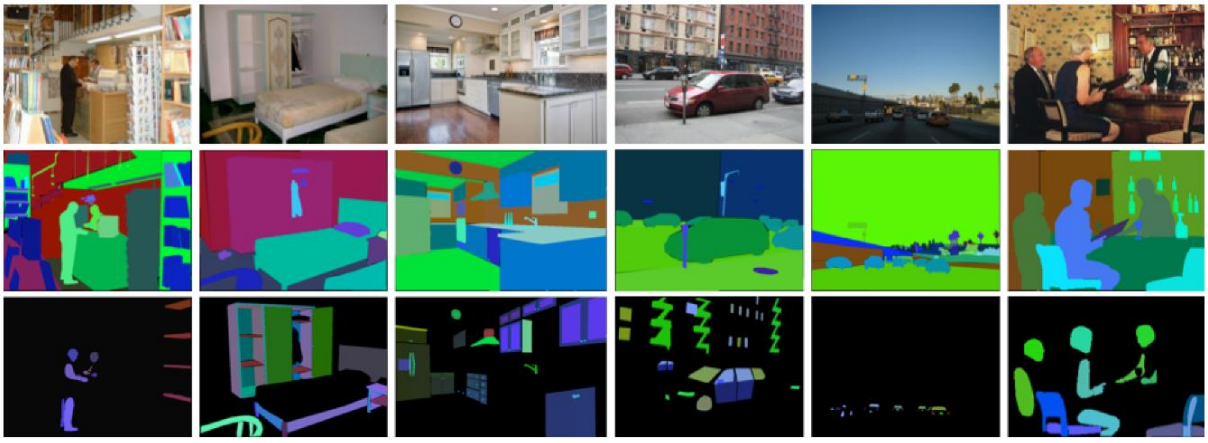

Figure 31: Examples of image and annotation of ADE20K dataset

3.13 Vehicles Detection Datasets

Within this category, has been included a selection of various datasets focused on vehicle detection and its subclasses. Within this section we will describe the following datasets:

- Motor Bike Dataset (MB1000) [33]: This dataset contains 10,000 annotated images taken with a Phantom 4® drone, with an HD camera under windy conditions, which affected the image

stabilizer capabilities. Images were resized to 640 x 364 pixels, although it is also available in 1920x1080, containing 56,975 ROI annotated objects, with a minimal height size set to 25 pixels. 60% of the annotated data corresponds to occluded motorcycles. Objects partially occluded with height less than 25 pixels were not annotated.

- Stanford Card Dataset [34]: The Stanford Car Dataset is a vehicle dataset taken by non-monitoring cameras with a bright vehicle appearance. This dataset includes 19,618 categories of vehicles covering the brands, models, and production years of the vehicles.
- TRANCOS Dataset [35]: The TRaffic ANd COngestionS (TRANCOS) dataset, a novel benchmark for (extremely overlapping) vehicle counting in traffic congestion situations. It consists of 1244 images, with a total of 46796 vehicles annotated. All the images have been captured using the publicly available video surveillance cameras of the Dirección General de Tráfico of Spain.
- Laroca & Boslooper Dataset [36]: This dataset collects 14,935 images from approximately 1,000 different wagons. These captured images present different types of wagons and under different conditions. The images were acquired with four different cameras in Full HD resolution.
- ImagesCV Bicycles Dataset [37]: The data set is composed of 705 images of bicycles of various models. Some images incorporate the presence of people on them, as well as several backgrounds.

This union of different data sets gives a broad view of the various types of vehicles that are currently used for detection and use in vision systems.

3.13.1 Relevance to project scenarios

The detection of vehicles and determining some of their subclasses has been highlighted by end users as being of interest, as well as being included in the list of objects to be detected by the VSAS (see

ANNEX I – List of Objects of VSAS and IMS). End users indicate that these elements are of interest for the following purposes:

- Can indicate the location of victims.
- Are flammable items.
- Are obstacles in the intervention paths of first responders.

These data sets allow the detection of these vehicles outdoors, in a wide range of distances, as well as their sub-classification. Figure 32, Figure 33,Figure 34and Figure 35 illustrate some of the contents of these datasets.

Figure 32: Stanford Car Dataset examples

Figure 33: TRANCOS Dataset to vehicle detection

Figure 34: Examples of trains/wagons of Laroca&Boslooper Dataset

Figure 35: Bicycle Dataset examples

3.14 Summary of Image Segmentation Datasets for VSAS

This chapter presents a series of datasets that can be useful to train some of the TeamAware project's detection algorithms, for the sample scenarios proposed in the project as well as others that may come up later on. The datasets provide valuable resources to develop and train deep learning algorithms for detection and recognition tasks from drones and wearable cameras, which are the main source of video footage that will be used in the project. Although some of these datasets require additional annotation work (which could potentially extend the duration of some of the tasks), the benefits of having so much data (especially in cases where training data is scarce) could justify this additional work.

4 Survey on Building Defect Datasets

4.1 Relevance to Project Scenarios

Photos of real damages after catastrophic events and synthetic images of damaged structures constitute the datasets on building/infrastructure damage owned and implemented by EUCENTRE. Such datasets will be deployed in TeamAware for training and internal validation processes for the implementation of the AI algorithm for damage detection, which is at the base of the Infrastructure Monitoring System (IMS), as described in deliverables D2.7 and D4.2. These datasets, in particular the two composed by photos of reals cases, are constituted by annotated images, in which several structural elements and structural damages are identified. For the datasets, the event of reference is mostly a natural disaster, in particular earthquakes but also ground deformations (such as landslides, settlements or subsidence) that may jeopardise the structural safety. The use of the datasets may also be extended to those manmade disasters that could cause damages to structural components or to those conditions in which a suitable maintenance level for infrastructure operability is not guaranteed.

4.2 Real-cases images Datasets of Post-event Damages and Structural Inspections

Two datasets of real-cases images of annotated damages have been selected for the purpose of training and testing the IMS damage detection algorithm.

The first one, the EUCENTRE's dataset, collects photos from surveys performed on different occasions:

- Post-event structural safety assessments after the last three major seismic sequences in Italy: the L'Aquila earthquake (2009), the Emilia earthquake (2012) and the Central Italy earthquake (2016-2017). A sample is shown in Figure 36;
- Structural inspections on infrastructures to monitor their maintenance level and structural stability.

The dataset, constantly increased and updated, is at the time of writing constituted by approximately 40000 images of which the greatest part regards post event seismic assessments. Each image collected undergoes a process of annotation, based first on the identification of the structural element and subsequently of the damage (whose classification depends on the structural element typology). There are two particular cases in which the damage is not associated with a single element: the first may be a phenomenon globally affecting the structure (e.g., total collapse), the second could be a single type of damage associated with non-structural elements. Besides information of structural characteristics and damage level, images are annotated taking into consideration the geographical location, the use of the structure, the period in which the survey was performed.

Most part of the photos were taken with typical RGB cameras, with camera sensors of different image resolutions (e.g., Canon EOS 400D).

Some of the inspections of infrastructures (bridges) for maintenance monitoring were performed with optical RGB drones:

- DJI Spark: 1/2.3" CMOS, Effective pixels: 12 MP, Lens35 mm Format Equivalent, aperture: f/2.8, image size 3968×2976;
- DJI Mavic 2 PRO: 1'' CMOS, Effective Pixels: 20 MP, Lens: 35 mm Format Equivalent, aperture: f/2.8-11, image size 5472×3648;
- Anafi Parrot: 1/2.4'' CMOS, Effective Pixels: 21 MP, Lens35 mm Format Equivalent, image size 5344x4016.

Figure 36: Example of an annotated image belonging to EUCENTRE'S dataset (L'Aquila 2009 event, shear damage on unreinforced masonry piers and spandrels)

The environment in which the photos are shot is typically outdoor, with good visibility, usually in daylight with sun or light clouds, rarely rainy or severely foggy. A limited sample of images is shot indoor, again in good visibility conditions (the number of those images is limited to few hundred samples at most).

The point of view in general is at ground level for the post event surveys after earthquakes, although few hundred samples of photos of bridges are characterised by an aerial point of view, since they were taken with drones.

The dataset belongs to the EUCENTRE Foundation, with all rights reserved. Access to the dataset may be allowed upon request and for research purposes only with appropriate citation.

The second dataset is the so-called COncrete DEfect Bridge IMage dataset (CODEBRIM) [38] [39] [40] , created for the classification of commonly appearing concrete defects in concrete infrastructure, especially bridges. CODEBRIM is conceived as a labelled multi-target dataset with overlapping defect categories for use in machine learning and the design of deep neural networks.

The dataset includes five mutually non-exclusive classes of damage (with an additional class for non-defective background): crack, spalling, efflorescence, exposed bars, corrosion (stains). The images were acquired at high-resolution from 30 unique bridges, partially using drones to gain close-range access, and feature varying scale and context. Images were also taken under changing weather conditions to include wet/stained surfaces with multiple cameras at varying scales. The dataset is

constituted by approximately 1600 high-resolution images, with defects annotated with bounding boxes (largely with overlapping defects). Few examples of defects are given in Figure 37.

Figure 37 : CODEBRIM dataset examples of images with cracking, spalling, exposed rebars, corrosion.

4.3 Semi-synthetic images dataset for Post-event Damages and Structural Inspections

Deep Neural Networks are the state of the art of most of object detection algorithms. Such method, however, needs a significant number of examples of each object class to be detected (in the order of several thousands) in order to be sufficiently robust for industrial applications. This constraint is particularly difficult to overcome when the target object classes are attributable to anomalies of some kind, which, in the case of structural/infrastructural damages, may be rare hence determining datasets of limited volume.

Within the context of WP4, a methodology for the generation has been specifically developed, implemented and validated for the purpose of increasing the available number of images, in support to the operations of training and validation of object detection algorithms (Please refer to deliverable D4.2 for a detailed description of the methodology and to deliverable D4.6 for its first validation). The motivation for which the EUCENTRE Foundation is creating a tailor-made dataset of semi-synthetic images is because the artificial generation provides images of sufficiently comparable characteristics to real ones.

The dataset, at the date of editing of the present deliverable, is currently constituted by a first set of semi-synthetic images and it is currently being expanded with further sets of images.

The images that have been generated and included in the dataset represent the following:

- A subset of the most common building/infrastructure typologies (e.g., residential buildings, reinforced concrete girder bridges, unreinforced masonry churches, etc.).
- A subset of the most common structural damages aimed to represent the effect of both natural catastrophes (in particular earthquakes, landslides in which the ground deformation affects the structural stability) or lack of maintenance and weathering (i.e. cracks, spalling, leaching and corrosion are included).

The dataset is created starting from 3D point cloud or 3D mesh models obtained from aerial surveys with drones of real structures. The images of damages to train the Object Detection algorithm at the base of TeamAware's Infrastructure Monitoring System (IMS) have produced and extracted thanks to the use of the software Blender, considering as a pre-condition that the simulated survey is conducted only outdoors in daylight and with good visibility, by varying the following:

- point of view from which the image is shot (either reproducing a survey performed with a typical photo camera at ground level or an aerial survey with a drone at different heights and distances from the structure);
- light and exposure conditions, by varying luminosity (for sun, clouds, time of the day simulation), presence and intensity of shades;
- texture and colour of structural elements;
- typology, level of severity and location of the selected typologies of damages.

Examples of the semi-synthetic images included in the dataset are given in Figure 38.

Figure 38: Some examples of semi-synthetic images with bounding boxes annotation of damages (different colours mark different type of damages)

4.4 Summary of Building Defect Datasets

This chapter presents the three datasets that will be useful for the training and validation of the IMS in TeamAware's platform. The datasets, two constituted by real-word images, the other by artificially generated ones, are representative of several classes of damages, deemed relevant for structural safety assessment.

5 Conclusions

In conclusion, the deliverable presents a survey on datasets for deep learning algorithm training for First Responders application, with a focus on highlighting datasets useful for the TeamAware project. The partners involved in this deliverable have selected and currently use semi-synthetic data and data drawn from existing datasets commonly used for other applications, such as home assistant, video surveillance, autonomous navigation, etc. Since the datasets that have been used are the generic ones shared with public, the background and the detected poses may not be similar. Therefore, accuracy would increase if the dataset is collected from the field, in case there is possibility in future.

The deliverable identifies three main applications from which to draw relevant data: human pose estimation and human activities for victim detection tasks, situational awareness using visual segmentation for smoke/fire detection, and damage assessment datasets. The selected datasets are being reused by selecting appropriate categories to comply with the end-users' requirements in the framework of the TeamAware project.

The WP3 team has made significant progress and has identified several datasets that can be used to train deep learning algorithms for First Responders applications.

Overall, the deliverable highlights the importance of leveraging existing datasets to train deep learning algorithms for First Responders applications, particularly in the absence of real-world data. The use of semi-synthetic and composite data can be an effective solution to address this challenge and can help to advance the development of algorithms that can improve First Responders' ability to perform their critical work in emergency situations

6 References

[1] S. H. Bach, B. He, A. Ratner and C. Re, "'Learning the structure' of generative models without labeled data," ICML, pp. 273–282, 2017.

[2] M. Andriluka, L. Pishchulin and P. B. Gehler, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2014.

[3] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang and Z. Li, "UAV-Human: A Large Benchmark for Human Behavior Understanding with unmanned aerial vehicles," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 16266--16275, 2021.

[4] M. J. Bocus, W. Li, S. Vishwakarma, R. Kou, C. Tang, K. a. C. I. Woodbridge, R. McConville, R. Santos-Rodriguez, K. Chetty and others, "OPERAnet: A Multimodal Activity Recognition Dataset Acquired from Radio Frequency and Vision-based Sensors," 2021.

[5] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: a large video database for human motion recognition," *IEEE 2011 International conference on computer vision,* pp. 2556--2563, 2011.

[6] S. Hadfield and R. Bowden, "Hollywood 3D: Recognizing Actions in 3D Natural Scenes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3398--3405, 2013.

[7] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1290--1297, 2012.

[8] "VWFP: Virtual World Fallen People Dataset for Visual Fallen People Detection," https://zenodo.org/record/6394684.

[9] "FPDS Dataset," https://gram.web.uah.es/data/datasets/fpds/index.html .

[10] "VFP290K: A Large-Scale Benchmark Dataset for Vision-based Fallen Person Detection," https://github.com/DASH-Lab/VFP290K.

[11] "Harmonious Composite Images," https://github.com/noahzn/VictimDet.

[12] "SCUT FIR Pedestrian Dataset," https://github.com/SCUT-CV/SCUT_FIR_Pedestrian_Dataset.

[13] "LLVIP: A Visible-infrared Paired Dataset for Low-light Vision," https://bupt-ai-cz.github.io/LLVIP/.

[14] "FLIR Thermal Dataset for Algorithm Training," https://www.flir.com/oem/adas/adas-dataset-form/.

[15] "Switzerland Aerial Drone footage dataset," https://www.kaggle.com/kmader/drone-videos.

[16] “USTC Forest smoke and fire dataset,” http://smoke.ustc.edu.cn/datasets.htm.

[17] “Smoke & Fire,” https://www.kaggle.com/kutaykutlu/forest-fire.

[18] S. d. d. set, https://cvgl.stanford.edu/projects/uav_data/.

[19] “ Stanford dronedaset compressed,” https://www.kaggle.com/aryashah2k/stanford-drone-dataset .

[20] “Aerial data set,” https://www.kaggle.com/jangsienicajzkowy/afo-aerial-dataset-of-floating-objects.

[21] “Git LADI data set,” https://github.com/ladi-dataset.

[22] “LADI Registry,” https://registry.opendata.aws/ladi/ .

[23] “LADI data,” https://ladi.s3.amazonaws.com/index.html.

[24] “Indoor data,” http://buildingparser.stanford.edu/dataset.html.

[25] “FloodDBS,” [Online]. Available: https://www.kaggle.com/datasets/hhrclemson/flooding-image-dataset?select=Flood+Images.

[26] “Suitcase/Luggage Kaggle Dataset,” [Online]. Available: https://www.kaggle.com/datasets/dataclusterlabs/suitcaseluggage-dataset.

[27] “ImageCV Suitcase Dataset,” [Online]. Available: https://images.cv/dataset/suitcase-image-classification-dataset.

[28] “Roboflow Backpack Dataset,” [Online]. Available: https://universe.roboflow.com/small-scale-experiments/backpack-00fut/dataset/2.

[29] “Mobile Phone Kaggle Dataset,” [Online]. Available: https://www.kaggle.com/datasets/dataclusterlabs/mobile-phone-image-dataset.

[30] “Osford Pet Dataset,” [Online]. Available: https://public.roboflow.com/object-detection/oxford-pets.

[31] “Children vs Adults Classification,” [Online]. Available: https://www.kaggle.com/datasets/die9origephit/children-vs-adults-images.

[32] “ADE20K,” [Online]. Available: https://groups.csail.mit.edu/vision/datasets/ADE20K/.

[33] “MB1000,” [Online]. Available: http://velastin.dynu.com/videodatasets/UrbanMotorbike/mb7500.htm.

[34] “Stanford Car Dataset,” [Online]. Available: http://ai.stanford.edu/~jkrause/cars/car_dataset.html.

[35] "TRANCOS Dataset," [Online]. Available: https://gram.web.uah.es/data/datasets/trancos/index.html.

[36] R. Laroca, A. C. Boslooper and D. Menotti, "Automatic Counting and Identification of Train Wagons," *Computer Vision and Pattern Recognition,* 2020.

[37] "ImageCV Bicycle Dataset," [Online]. Available: https://images.cv/dataset/bicycle-image-classification-dataset.

[38] "CODEBRIM," https://github.com/MrtnMndt/meta-learning-CODEBRIM .

[39] "CODEBRIM 2," https://zenodo.org/record/2620293#.ZAIBIHbMJD_.

[40] S. M. S. M. P. P. V. R. Martin Mundt, "Meta-learning Convolutional Neural Architectures for Multi-target Concrete Defect Classification with the COncrete DEfect BRidge IMage Dataset," IEEE CVPR, 2019.

[41] D. C. Luvizon, D. Picard and H. Tabia, "2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning," CVPR, 2018.

[42] W. L. e. a. M. J. Bocus, "OPERAnet: A Multimodal Activity Recognition Dataset Acquired from Radio Frequency and Vision-based Sensors," 2021.

[43] "Pedrestian Dataset," [Online]. Available: https://www.kaggle.com/datasets/karthika95/pedestrian-detection.

[44] "Harmonious Composite Images," https://github.com/noahzn/VictimDet.

ANNEX I – List of Objects of VSAS and IMS

This annex collects all the objects proposed by end users and included in scope of WP3/WP4. Each of these objects includes:

- **Purpose:** First responders' target to include this type of objects.
- **Relevance:** Relevance of objects and why they should or should not be included.
- **Indoor/Outdoor:** Type of scenes where object detection is applied.
- **Sensor:** Indicates in which hardware these elements can be detected.
- **End-user:** Indicate the end user of TeamAware with interest in the detection of this type of object.
- **Accuracy:** Expected accuracy for the detection of these elements.
- **System:** Indicates the subsystem of TeamAware that produce the detection/segmentation/classification of these objects.
- **Indoor/Outdoor:** Where they are and can detect the indicated element.

Object	Purpose	Relevance	Indoor / Outdoor	Sensor	End-user	Expected accuracy (ACC)	System
Motorcycles / Bikes	These elements are important because they present potential hazards (e.g., explosions during fires), indications of the presence of people in places or areas of interest or areas of victim location.	Medium	Outdoor	Outdoor EO Drone	JOAFG	70%	VSAS
Car			Outdoor	Outdoor EO Drone	JOAFG	80%	VSAS
Bus			Outdoor	Outdoor EO Drone	JOAFG	80%	VSAS
Rails			Outdoor/Indoor	Outdoor EO Drone / Indoor Drone / Helmet	JOAFG	80%	VSAS

Window	These access structural elements are relevant in the guidance and mapping of zones.	Medium	Both	Outdoor EO Drone / Indoor Drone / Helmet	JOAFG	90%	VSAS
Door			Booth	Outdoor EO Drone / Indoor Drone / Helmet	JOAFG	90%	VSAS
Chimney	These access structural elements are relevant in the guidance and mapping of zones.	Medium	Indoor	Outdoor EO Drone / Indoor Drone / Helmet	JOAFG	90%	VSAS
Staircase	These transit structural elements are relevant in the guidance and mapping of zones.	Medium	Both	Outdoor EO Drone / Indoor Drone / Helmet	JOAFG	90%	VSAS
Cracks	Damage of structures is highly relevant since it represents potential dangers in the intervention of	High	Outdoor	Outdoor EO Drone	JOAFG	90%	IMS
Spalling / Loss of material			Outdoor	Outdoor EO Drone	JOAFG	90%	IMS
Exposed rebars / Corrosion			Outdoor	Outdoor EO Drone	JOAFG	90%	IMS

	the first responders.						
People	The detection of civilians (regardless of age) in disaster situations is a priority.	High	Both	Outdoor EO Drone / Indoor Drone / Helmet	HSEPC AAHD AHBVP JOAFG	90%	VSAS
Adult			Both	Outdoor EO Drone / Indoor Drone / Helmet	HSEPC AAHD AHBVP JOAFG	70%	VSAS
Child			Both	Outdoor EO Drone / Indoor Drone / Helmet	HSEPC AAHD AHBVP JOAFG	70%	VSAS
Victim	The detection of victims in any of the situations or states	High	Indoor	Indoor Drone / Helmet	HSEPC AAHD AHBVP	90%	VSAS

	indicated is a priority				JOAFG		
Trapped person in ruins			Indoor	Indoor Drone / Helmet	HSEPC AAHD AHBVP JOAFG	90%	VSAS
Fainted persons			Indoor	Indoor Drone / Helmet	HSEPC AAHD AHBVP JOAFG	90%	VSAS
Person in damaged train			Indoor	Indoor Drone / Helmet	HSEPC AAHD AHBVP JOAFG	90%	VSAS
Fire/Explosion	Events involving fire, floods, etc. they are extremely relevant since they are the main effects of disasters.	High	Both	Outdoor EO Drone / Indoor Drone / Helmet	AAHD AHBVP JOAFG	90%	VSAS
Smoke / Dust / Fog			Both	Outdoor EO Drone /	AAHD	90%	VSAS

				Indoor Drone / Helmet	AHBVP JOAFG		
Flood			Both (Mainly in outdoor)	Outdoor EO Drone / Indoor Drone / Helmet	AAHD AHBVP JOAFG	90%	VSAS
Common domestic animal	Domestic animals are very frequent in the defined scenarios and are an important focus for the rescue and detection of victims.	Low	Indoor	Indoor Drone / Helmet	None	80%	VSAS
Backpack	These elements are important sources of danger as they may contain explosive elements, be flammable, etc.	Medium	Both	Outdoor EO Drone / Indoor Drone / Helmet	AAHD AHBVP JOAFG	80%	VSAS
Suitcase			Both	Outdoor EO Drone / Indoor Drone / Helmet	AAHD AHBVP JOAFG	85%	VSAS

Luggage			Both	Outdoor EO Drone / Indoor Drone / Helmet	AAHD AHBVP JOAFG	80%	VSAS
Mobile Phone			Both	Outdoor EO Drone / Indoor Drone / Helmet	AAHD AHBVP	80%	VSAS