MMFlood: A Multimodal Dataset for Flood Delineation from Satellite Imagery

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 August 2022, accepted 2 September 2022, date of publication 8 September 2022, date of current version 19 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3205419

APPLIED RESEARCH

MMFlood: A Multimodal Dataset for Flood Delineation From Satellite Imagery

FABIO MONTELLO[1], **EDOARDO ARNAUDO**[1,2], **AND CLAUDIO ROSSI**[1]
[1]AI Data and Space (ADS), LINKS Foundation, 10138 Turin, Italy
[2]Dipartimento di Automatica e Informatica (DAUIN), Politecnico di Torino, 10138 Turin, Italy

Corresponding author: Edoardo Arnaudo (edoardo.arnaudo@polito.it)

This work was supported in part by the Horizon 2020 Projects Sustainable Historic Environments hoListic reconstruction through Technological Enhancement and community based Resilience (SHELTER) under Agreement 821282, and in part by the Structured Approaches for Forest fire Emergencies in Resilient Societies (SAFERS) under Agreement 869353.

ABSTRACT Accurate flood delineation is crucial in many disaster management tasks, such as risk map production and update, impact estimation, claim verification, or planning of countermeasures for disaster risk reduction. Open remote sensing resources such as the data provided by the Copernicus ecosystem enable to carry out this activity, which benefits from frequent revisit times on a global scale. In the last decades, satellite imagery has been successfully applied to flood delineation problems, especially considering Synthetic Aperture Radar (SAR) signals. However, current remote mapping services rely on time-consuming manual or semi-automated approaches, requiring the intervention of domain experts. The implementation of accurate and scalable automated pipelines is hindered by the scarcity of large-scale annotated datasets. To address these issues, we propose MMFlood, a multimodal remote sensing dataset purposely designed for flood delineation. The dataset contains 1,748 Sentinel-1 acquisitions, comprising 95 flood events distributed across 42 countries. Along with satellite imagery, the dataset includes the Digital Elevation Model (DEM), hydrography maps, and flood delineation maps provided by Copernicus EMS, which is considered as ground truth. To provide baseline performances on the MMFlood test set, we conduct a number of experiments of the flood delineation task using state-of-art deep learning models, and we evaluate the performance gains of entropy-based sampling and multi-encoder architectures, which are respectively used to tackle two of the main challenges posed by MMFlood, namely the class unbalance and the multimodal setting. Lastly, we provide a future outlook on how to further improve the performance of the flood delineation task. Dataset and code can be found at `https://github.com/edornd/mmflood`.

INDEX TERMS Computer vision, deep learning, image processing, machine learning, semantic segmentation, remote sensing, natural disaster dataset.

I. INTRODUCTION

One of the most impacting consequences of climate change is the intensification of the water cycle, a phenomenon that in some regions is causing more intense rainfall, increasing the risk of severe flood events. The most recent reports estimate that, by the end of this century, intense precipitation events that would typically occur two times per century would occur twice as often [1]. At the same time, coastal areas are expected to see a constant sea-level rise throughout the century, contributing to more frequent and severe coastal flooding in low-lying areas [1]. Floods are the most significant disaster type in terms of affected people, producing a wide range of health impacts with short, medium, and long terms effects [2], [3]. Moreover, floods cause a significant amount of economic losses. For example, in Italy, it is estimated that the damages caused by floods events will amount to 654 million euros per year in the interval between 2014 and 2100 [4]. For countries in Central Europe, a 1-in-20-year flood (probability of 5%) could lead to losses within the range of 2.2-10.7% of the government revenues [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

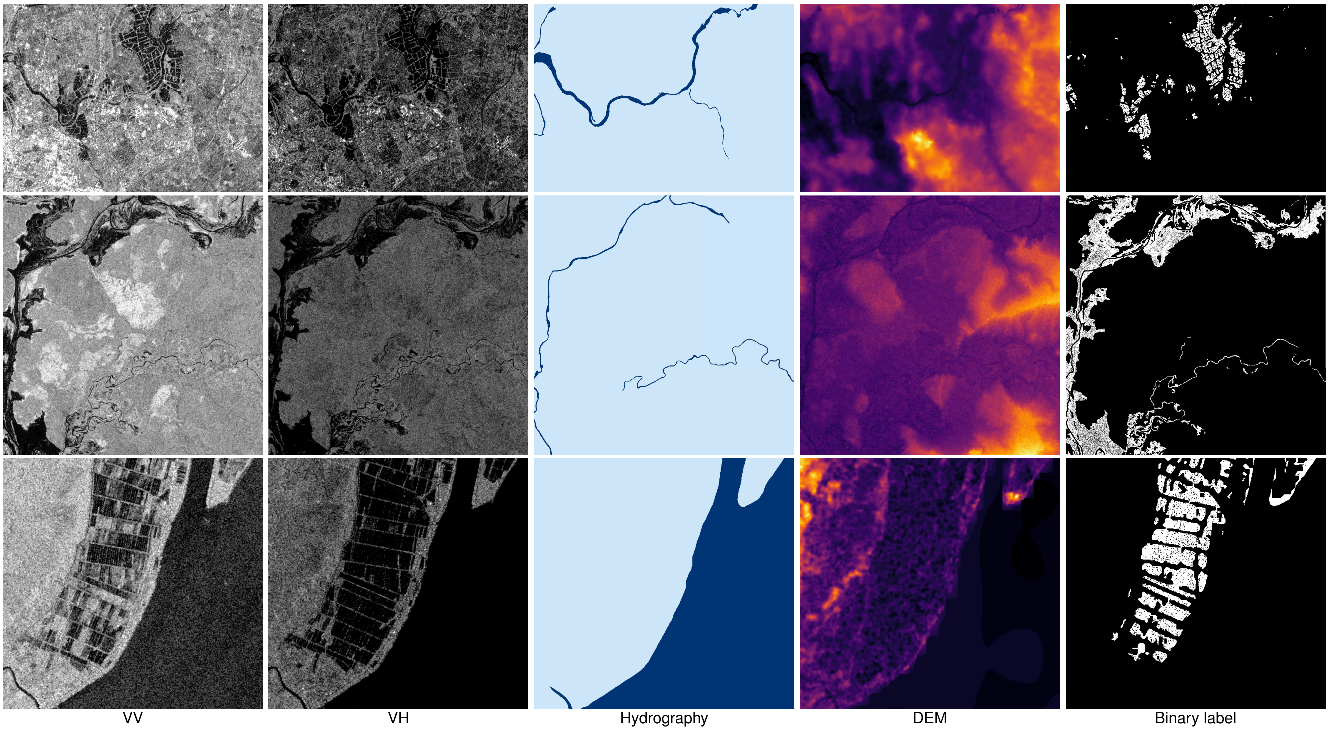


FIGURE 1. Samples from the MMFlood dataset. From left to right, SAR bands (VV and VH polarizations), hydrography, elevation model and binary label.

Information becomes a critical resource in case of flood events: providing geographical extent, severity, and socioeconomic impacts of a natural hazard using near real-time Earth Observations (EO) can improve the assessment of the affected areas more quickly and efficiently [6]. EO-based imagery, either derived from aerial or satellite acquisitions, is becoming crucial for the detection and the monitoring of natural hazards, as well as to support activities related to the restoration and adaptation phase. In recent years, the increased accessibility of remote sensing services has allowed for new and improved applications, from urban development [7] to agricultural [8], [9], and emergency scenarios [10], [11]. Considering floods, aerial or remote sensing images have been extensively and successfully employed in the last decade, with a focus on SAR signals. Despite the typically lower image quality compared to optical or multispectral variants, SAR technology has several advantages. First, it is not sensitive to light because the acquisition is based on the backscatter coefficient, thus allowing the provision of actionable data also at night. Second, the SAR signal is less affected by atmosphere or clouds, making it possible to map the underlying soil in different weather conditions. Last, its inherent sensitivity to the physical and dielectric properties of the traversed materials make SAR imagery extremely valuable for water and soil moisture analysis [12].

However, the vast majority of remote sensing applications aimed at mapping flood events are typically carried out through manual or semi-automated approaches [13], [14]. The challenge of this task mainly resides in the complex and time-consuming activity carried out by domain experts that manually inspect and annotate the available images along different emergency management phases and often in several iterations [13]. In recent years, this process has been progressively improved. First, using standard image processing, with techniques such as spectral indices [14] or thresholding [15]. Then, with fully automated supervised machine learning approaches, where the most effective solutions are provided by deep convolutional neural networks aimed at binary segmentation, i.e., the pixel-wise classification of the image in flooded or not flooded areas [10]. Nonetheless, a major setback of supervised machine (and deep) learning solutions remains the strong reliance on large amounts of annotated data. While this issue is being gradually mitigated in the context of optical data [8], [9], a low number of datasets are including SAR signals, even less targeting flood delineation. Moreover, the scope of such data is often limited to specific geographical regions [16], or they include large portions of weakly annotated data [17] that makes it harder to realise highly accurate supervised models.

In this paper, we propose *MMFlood*, a novel remote sensing dataset derived from Sentinel-1, with the aim of providing a complete and well-rounded set of data specifically designed for the delineation of flooded areas. While building *MMFlood*, we considered three key aspects: worldwide distribution, the availability of manual annotations and multiple data sources. To obtain the labelled data, we exploit the Copernicus Emergency Management System (EMS) service, [18], cross-referencing hundreds of manually produced flood delineation maps with the raw SAR data obtained from Sentinel-1. In the time span ranging from 2014 to 2021, we obtain 1,748 SAR acquisitions and the corresponding pixel-level annotations from 95 flood events distributed in 42 different countries around the world. These acquisitions translate into a total of 8,522 images using a tiling dimension of 512 × 512 pixels, or 30,855 at 256 × 256, which is comparable or superior to similar existing datasets [16], [17] (see Sec. V-A). Given the challenging task of delineating the flooded areas from a single data source, we add the Digital Elevation Model (DEM) sourced from MapZen [19]. In fact, DEM has been shown to improve the accuracy of models in discriminating the area of interest from permanent water bodies [20], [21]. We also include the hydrography map, derived from OpenStreetMap (OSM) [22], which can provide additional information about the location of large rivers and water basins in the area. Both the DEM and the hydrography feature the same spatial resolution and the same coverage of the corresponding Sentinel-1 data. A visual sample of the dataset is presented in Fig. 1. In summary, the contributions of this paper can be broken down as follows:

- We present *MMFlood*, a multimodal dataset created specifically for the task of flood delineation, describing the construction process and its features in comparison to the most known remote sensing datasets;
- We provide a benchmark evaluation on *MMFlood* using state-of-the-art deep learning segmentation algorithms to provide a solid performance baseline for future works;
- We explore the performance improvements of entropy-based sampling and multi-encoder architectures to address the class unbalance problem, which is inherent to this task because floods typically cover a limited portion of land, and the multimodal setting enabled by the availability of multiple data sources, respectively.

The remainder of this document is organised as follows. Section II introduces related works, focusing on existing remote sensing datasets related to SAR data and natural disasters, as well as on existing techniques in computer vision and machine learning for binary segmentation tasks. Section III introduces the *MMFlood* dataset, describing its construction and validation process, while Section IV delineates the experimental setup, together with the preprocessing steps and the methodology employed. Section V details and discusses the results obtained from our experiments. Finally, Section VI draws the final conclusions, highlighting possible directions in view of future works.

II. RELATED WORK

A. DATASETS

Aerial and remote sensing datasets are still relatively limited with respect to the number of available resources and their extension. If we consider the largest and most popular datasets in the EO field, we find that they are typically dedicated to land cover classification, or its derivative tasks. Among them, it is worth mentioning by popularity the Vaihingen and the Potsdam datasets [23]. However, they are both limited to acquisitions over a single city, counting few Very-High Resolution (VHR) imagery in multiple bands, as detailed in Table 1. Larger datasets are often aimed towards general-purpose tasks like BigEarthNet [24] or DeepGlobe [26], dedicated to land cover, or derived from aerial acquisitions such as Agriculture-Vision [9]. The DeepGlobe Land Cover challenge encompasses different geographical areas and contains thousands of images, paired with pixel-level semantic annotations for 7 categories, from urban to agricultural areas. However, the images only cover the visible spectrum and they do not include location data, hence limiting its use in other contexts. BigEarthNet consists of 590,326 Sentinel-2 image patches annotated with multiple land-cover classes, considering 10 different countries. Given the classification task, this dataset only contains coarse labels, indicative of possible elements present in the image, from vegetation to water, from agriculture to urban areas. Lastly, Agriculture-Vision represents one of the first, publicly available, large-scale aerial dataset for agricultural land cover, providing RGB and Near-Infrared (NIR) bands and with different agricultural patterns manually annotated by experts. It contains 94,986 images at resolutions varying from 10 to 20 m/pixel, sampled from American farmlands.

Focusing on disaster management and flood delineation, the amount of open datasets available is further reduced, typically only including images in the visible spectrum, thus providing coarser annotations. This is, for instance, the case of xBD [27], a large-scale dataset providing 9,168 annotated images, comprising 5 disaster types, including floods, wildfires and earthquakes. A similar example specific for flood segmentation is FloodNet [28], an aerial dataset comprising more than two thousands VHR drone images acquired after hurricane Harvey and annotated with semantic categories including flooded areas. On top of the limitation represented by the sole availability of optical imagery, we also have the lack of geographical diversity. Among the possible satellite instruments available to overcome this limitation, the most common for the flood mapping task is in fact the SAR. The advantage in this case is not being disrupted by cloud cover (extremely common during flood events) or lack of illumination. A major obstacle for machine learning analysis for flood mapping is the scarce availability of SAR datasets, together with the lack of annotations and robust ground truths. To this day, among the data source available for supervised training methods, it is worth mentioning SEN12-FLOOD [29], a co-registered optical and SAR images time series for the detection of flood events, built from 336 time series containing Sentinel-1 and Sentinel-2 images of areas hit by major flooding event during 2019, including east Africa, south-west Africa, Middle-East, and Australia. An improvement on top of this dataset is provided by Sen1floods11 [17], a set of images based on Sentinel-1 and Sentinel-2 of flooded areas which are composed of both manual and automated annotations, including in different proportions permanent water masks, flood water masks, and raw Sentinel imagery. Sen1floods11 consists of 4,831 tiles with

TABLE 1. Brief summary and comparison between MMFlood and other aerial datasets. S1 and S2 refer to Sentinel-1 and Sentinel-2 respectively. Modalities include visible spectrum (RGB), Digital Surface Maps (DSM), Digital Elevation Maps (DEM), all the 12 bands of Sentinel-2, SAR imagery with VV and VH polarizations (VV-VH), and hydrography (Hd). Tasks refer instead to *classification* (C), *segmentation* (S), *object detection* (OD), and *visual question answering* (VQA).

Dataset	Source	Modalities	Geolocated	Task	Images	Img. size	Resolution
Vaihingen [23]	Aerial	RG-IR, DSM	✓	S	33	<2,500×2,000	10cm
Potsdam [23]	Aerial	RGB-IR, DSM	✓	S	38	6,000×6,000	10cm
BigEarthNet [24]	S2	12 bands	✓	C	590,326	120×120	10-60m
LandCoverNet [24]	S2	12 bands	✓	C	9,000	256×256	10-60m
Agriculture-Vision [9]	Aerial	RGB-NIR	×	S	94,986	512×512	10-20cm
HRSID [25]	S1, TerraSAR-X	HH,VV,HV	✓	OD	5,604	800×600	1-5m
DeepGlobe [26]	Maxar	RGB	×	S	1,146	2,448×2,448	50cm
xBD [27]	Maxar	RGB	×	OD	9,168	1,024×1,024	80cm
FloodNet [28]	Aerial	RGB	×	S, C, VQA	2,343	4,500×3,000	1.5cm
SEN12FLOOD [29]	S1, S2	VV-VH (S1), 12 bands (S2)	✓	C	336*	512×512	10-60m
sen1floods11 [17]	S1, S2, JRC	VV-VH (S1), 12 bands (S2), Hd.	✓	S	4,831	512×512	10-60m
ETCI-2021 [16]	S1, NASA	VV-VH, Hd.	×	S	33,405	256×256	20m
MM-Flood	S1, MapZen, OSM	VV-VH, DEM, Hd.	✓	S	1,748	<2,000x2,000	20m

worldwide coverage, comprising 11 different floods events. Among the total amount of images, 446 are hand-labeled images of surface water from flood events, 814 are images of publicly available permanent water data labels from Landsat and the remaining 4,385 images are of surface water classified from Sentinel-2 images from flood events. Given the delineation carried out through Sentinel-2 imagery, many annotations contain several missing parts due to high cloud coverage which, especially during floods events, is often an issue for multi-spectral imagery. Another comparable proposal is the ETCI 2021 dataset [16], containing more than 30, 000 raw Sentinel-1 SAR tiles from five different geographical regions. Although the available datasets are extremely valuable, we argue that most of them lack in certain aspects: first, they often contain a limited number of images that represent an actual flood event, where the vast majority represents water basins. Second, they lack a DEM, which has been proven to be useful in assisting the models to discriminate between flooded areas and permanent water bodies, resulting in a more precise prediction [20], [21], [30]. Last, only a few proposals include hydrography map or include it as part of the segmentation mask, which does not help the model to learn how to differentiate only the flood event.

We try to address these issues through our *MMFlood* dataset. In the first place, we exploit Copernicus EMS as ground truth, generating masks from high quality activations that provide three main features: they have been produced using SAR imagery, they have been manually validated by experts, and most importantly they only include flooded areas. Then, we include both DEM and hydrography, when available, of the same areas of interest to exploit additional modalities at training time.

B. ALGORITHMS

Thanks to the continuously increasing availability of freely available remote sensing data, many studies concerning the delineation of flooded areas from satellite acquisitions have been proposed over the years. The approaches proposed in these works cover a wide range of techniques, both in terms of algorithms and data sources used. Among the possible satellite instruments available, most of the flood mapping literature revolves around the use of SAR data from different satellite networks as TerraSAR-X [31], RADARSAT [32], COSMO-SkyMed [33] and Sentinel-1 [34]. The latter is one of the most convenient options available to this day, given the worldwide coverage at medium-high spatial resolution, short revisit times and the open data availability. Preliminary approaches in this field were mainly using masking and thresholding, combined with careful data preprocessing [15], [35], [36], [37] or by using Fuzzy Logic approach [38], [39], [40]. With the increasing applications of Artificial Intelligence and Deep Learning techniques in the Computer Vision field, many supervised machine learning classifiers have been devised and exploited for these tasks. In particular, previous works have proposed Support Vector Machines [41], [42], Fully Convolutional Neural Networks [43] Bayesian Networks [44], Deep Belief Networks [45], or Random Forests [10]. To this day, deep learning solutions have mostly focused on flood delineation on ground level [46] or through drone and aerial imagery [28]. However, several works have been carried out on remote sensing SAR data, from image despeckling [47], [48], to the detection of large objects such as ships [25], [49] and land cover classification [8], [24]. SAR imagery is often enriched with information derived from additional sources, including optical data such as Sentinel-2 [24], or even completely different modalities such as Automatic Identification Systems (AIS), which is often exploited in vessel detection as both supplementary knowledge through domain adaptation [50] or ad-hoc data fusion [51].

While classical machine learning techniques provide robust results across different tasks, Convolutional Neural Networks (CNN) remain the dominant architecture, with encoder-decoder modules such as U-Net [10], or multi-scale extraction such as DeepLab [52].

III. THE DATASET

This section describes the dataset in all its components, and the post-processing operations applied to the raw data, with

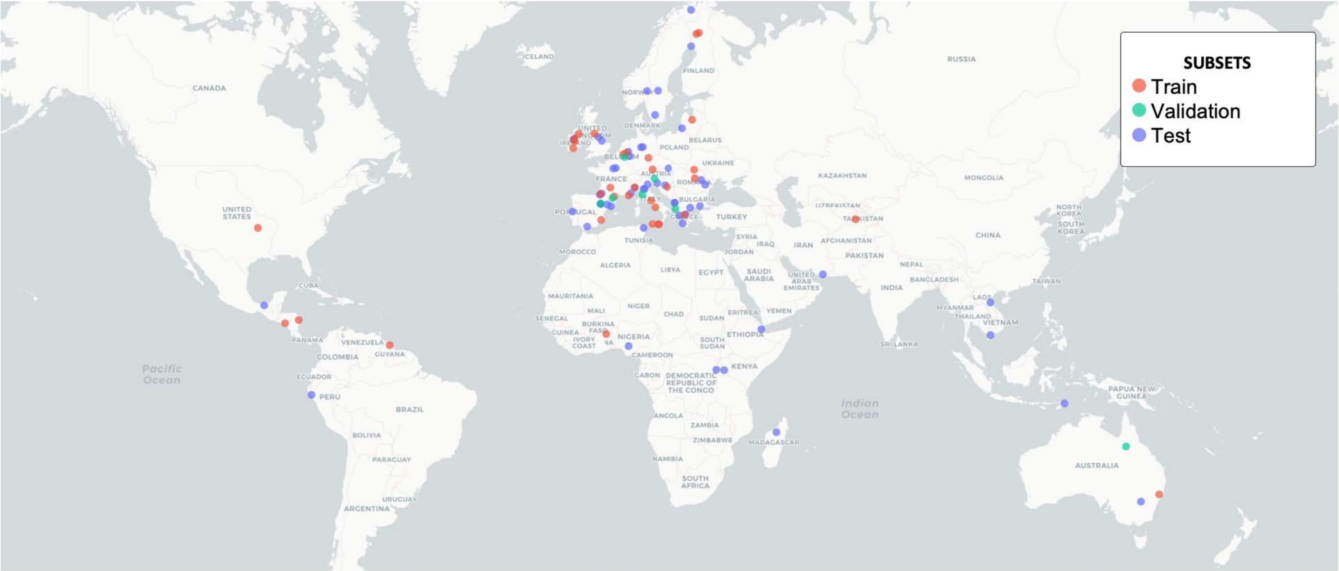


FIGURE 2. Locations of the major floods derived from Copernicus EMS and exploited for the dataset generation. Different colours correspond to different splits.

the aim of refining and optimising the available information for training purposes. The proposed dataset is composed of a set of 1,748 tuples, in turn comprising four separate pieces of information: (i) SAR satellite acquisitions, gathered from the Copernicus Sentinel-1 mission, (ii) pixel-wise DEM maps, (iii) hydrography information, and (iv) binary annotations, obtained through the Copernicus EMS service and delineating flooded or not flooded areas.

TABLE 2. Summary of activation counts, grouped by country. Major events from a total of 42 countries are present in the final dataset.

Italy	11	Togo	1
France	10	Djibouti	1
Spain	7	Slovenia	1
Germany	6	Portugal	1
Greece	6	Croatia	1
Ireland	5	Moldova	1
United Kingdom	4	Lithuania	1
Australia	3	Timor-Leste	1
Albania	3	Guyana	1
Finland	3	Peru	1
Romania	3	Tajikistan	1
Sweden	3	Honduras	1
Nicaragua	2	Mexico	1
Netherlands	2	Iran	1
Norway	2	Nigeria	1
Vietnam	2	Madagascar	1
Uganda	2	Belgium	1
Ukraine	2	Austria	1
Latvia	1	Slovakia	1
Tunisia	1	Bosnia and Hze...	1
Belgium	1	United States	1

A. FLOOD EVENTS

The whole construction process of this dataset revolves around Copernicus EMS [18], which represents the main source of information for the identification and delineation of flooded areas. Copernicus is an EU program that has the purpose of developing European information services based on satellite EO and in-situ data. Its main objective is to monitor and forecast the state of the environment on land, sea and in the atmosphere, in order to support climate change mitigation and adaptation strategies.

The information tools provided by Copernicus are available to authorities and first responders on a fully open and free-of-charge basis. Among these, the EMS service is exploited to retrieve the mapping of flooded areas, as shown in Fig. 3A. The latter provides information belonging to past emergency responses in relation to different types of disasters, including meteorological hazards, geophysical hazards, deliberate and accidental man-made disasters, and other humanitarian disasters, as well as prevention, preparedness, response and recovery activities. Specifically, the EMS Rapid Mapping products provide a large data collection, a provision of geospatial information within hours or days from the disaster, including flood delineations. To carry out the data gathering process, we consider the vector packages with flood delineation products, comprising sets of geographical files with the purpose of assessing the extent of the flood events. The EMS damage maps are thus used as ground truth for training machine learning models.

The retrieval of the vector packages is executed programmatically, extracting the vector file containing the delineation of the flooded area, and considering all the activations prior to July 2021 (Fig. 3B). Due to differences in package structure, naming and inconsistencies in the geometry, a small portion of the hundreds of available activations was discarded. For this reason, after retrieval, we performed a thorough manual inspection over the activations to ensure a certain degree of correctness and uniformity among the obtained packages and their contents. The final activation list considers a total of 95 individual flood events, including a large variety of data spanning seven years, starting from 2014 (i.e., dating the initial Sentinel-1 acquisitions), and encompassing 42 different countries (see Table 2). Given the inherent focus of the Copernicus products on the European soil, Italy, France, and Spain gather the largest number of activations, while an equally large portion of data is scattered around the world. The least represented continent[1] in terms of absolute numbers is South America, with 2 unique flood events. A complete list of the EMS activations selected for the dataset is listed in Appendix.

For each activation, we collect all the necessary information to provide both precise annotations and additional contextual details, specifically: the event date, intended as estimate of the flood starting time, a general location of the area of interest, the bounding box of the event, and the polygons corresponding to the actual flood delineation. Each activation usually groups together several disasters, happened around the same region and caused by the same agent; for this reason the former may contain one or multiple vector packages, defining different flooded sub-regions within the same event. In order to maximise the recall for the dataset construction, we consider each one of them separately for the subsequent image acquisition and rasterization phase (Fig. 3C). Flood polygons have been manually validated by both the service provider that generated such delineations, and the Joint Research Centre (JRC) services to ensure a certain degree of accuracy [13]. From these polygons, we automatically derive the minimum bounding box fully containing the flooded areas, necessary for the image retrieval phase. We also pad the areas in each direction by a variable amount corresponding to 10% of its maximum extent, to account for possible inaccuracies in the delineations and to include contextual information.

B. DATA ACQUISITION

Given the list of geolocated EMS activations with vector delineations, the main objective is to retrieve a set of image tuples, including remote sensing acquisitions, additional modalities and binary labels, with pixel-level alignment.

For the creation of this dataset we focus on SAR images given their inherent properties particularly suitable for the task of flood delineation, especially considering the visibility in cloudy environments. Sentinel-1, the first mission in the Copernicus Programme conducted by the European Space Agency (ESA) in 2014, represents one of the best options in this field, given the worldwide coverage, the relatively short revisit time, and its open and free availability. Its constellation comprises two twin satellites with polar orbit, Sentinel-1A and Sentinel-1B, which share the same orbital plane. The radar acquisition is carried independently through a C-band SAR instrument, thus providing a variable revisit time that differs according to the latitude, estimated as 3 days at the equator, while even less than 1 day at the poles. Given its purpose, we focus the acquisition on Level-1 Interferometric Wide (IW) Ground Range Detected (GRD) products, which provide only the signal amplitude without its phase, but can maintain an approximately square spatial resolution and square pixel spacing without further resampling, and display reduced speckle due to the multi-look processing. In order to facilitate the retrieval of the images only in desired areas and periods, we exploit the Sentinel-Hub platform,[2] a third-party service providing complex filtering and compositing functionality with direct access to several open remote sensing sources, such as Sentinel or Landsat. For each bounding box collected in the previous step, we retrieve the corresponding raw SAR signal with two channel variants, namely VV (Vertical transmit and Vertical receive) and VH (Vertical transmit and Horizontal receive), obtaining every image in a window with a maximum of 4 days from the day of the event to minimise the discrepancies between images and flood delineations (Fig. 3D). Since each area may not be fully covered by a single Sentinel-1 tile, we exploit the mosaicking features of Sentinel-Hub to automatically merge neighbouring acquisitions. In the rare event where such adjacent tiles cannot be found, the pixels missing in the SAR acquisition will also be masked out in the corresponding ground truth.

All the images are provided as orthorectified, georeferenced TIFF (GeoTIFF) files, with a spatial resolution of 20m/pixel and 32-bit floating point precision.

In addition to the SAR imagery, we collect the DEM of the same areas of interest (Fig. 3E) based on Mapzen's terrain tiles, a static collection mainly based on the Shuttle Radar Topography Mission (SRTM30) [53]. Its main purpose is to provide additional context to the flood prediction process, by suggesting which areas are more likely to be flooded based on elevation. The provided data is inherently static, therefore it does not take into account small variations of terrain due to side natural events (e.g. landslides), which are common phenomena during heavy rains. On the other hand, static data allows for a much wider coverage and a rarer presence of noise in the retrieved images. While the original resolution of the available DEM varies from 5 to 30m/pixel, we provide a resampled version at the same 20m/pixel resolution of the SAR image, as a separate single-band GeoTIFF with 32-bit floating point precision.

With a similar process, we further expand the available information by including the hydrography map of the areas of interest. Given the worldwide scope of the retrieved EMS activations, we leverage on OSM for this task. Exploiting the provided Overpass API, we extract polygons for every water layer available in every region inside the image bounds, when available. With the aim of maintaining a pixel-wise alignment among tuples, we rasterise each set of hydrography vectors at 20m/pixel to maintain same resolution of previous modalities. While not every area provided high-quality polygons of the water basins, more than half of the available SAR acquisitions is accompanied by an hydrography raster.

[1] https://en.wikipedia.org/wiki/Continent

[2] https://www.sentinel-hub.com/

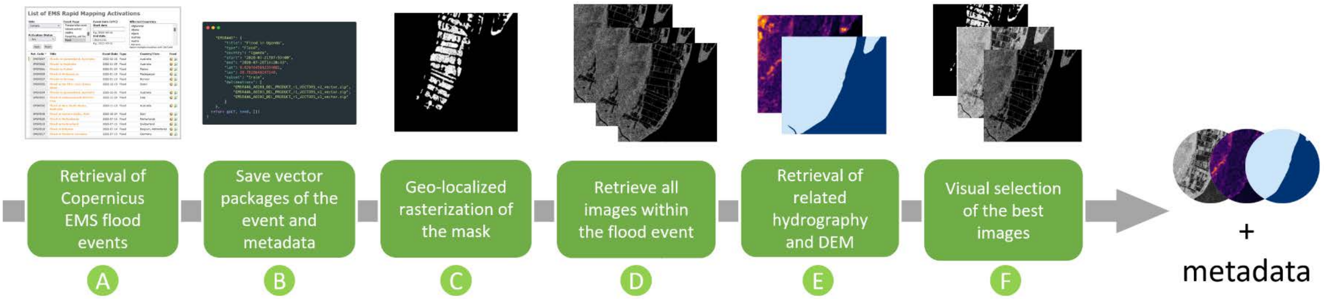

Retrieval of Copernicus EMS flood events
A
Save vector packages of the event and metadata
B
Geo-localized rasterization of the mask
C
Retrieve all images within the flood event
D
Retrieval of related hydrography and DEM
E
Visual selection of the best images
F
+
metadata

FIGURE 3. Processing steps taken to generate the dataset from Copernicus EMS activations: retrieval and conversion to raster images (A-C), acquisition of images and additional modalities (D-E), and manual inspection and filtering (F).

After the image retrieval phases, a thorough manual filtering is carried out to further improve the quality of the final product (Fig. 3F): among every Sentinel-1 image retrieved within the 4-day period, starting from the event date of each activation, only the one that was visibly matching the flood mask is maintained, discarding all the others. Furthermore, given the possible inconsistencies among different geographical regions, DEM rasters are also manually verified, filtering out all those tuples with invalid elevation map. Concerning the hydrography, the availability is much sparser, either because of missing data or absence of water basins in the requested area. We still opted to keep every sample without corresponding hydrography, both for future studies and as optional post-processing step to further subtract permanent water areas from the final prediction of the models, if required.

TABLE 3. MMFlood data specification, including format and bands.

Modality	Format	Bands
Sentinel-1 IW GRD	GeoTIFF, Float32	0: VV, 1: VH
DEM	GeoTIFF, Float32	0: elevation
Hydrogr.	GeoTIFF, Uint8	0: hydrography
Flood Mask	GeoTIFF, Uint8	0: ground truth

The released dataset is specified in Table 3: it is grouped in 1,748 elements containing SAR, DEM, and mask images, out of which 1,012 also include the correspondent hydrography images (i.e., 57.9%). Each group contains additional metadata describing the activation code, the date of the event, and the country in which the event has happened. Moreover, SAR data are also annotated with the acquisition date of the satellite signal itself, which may differ from the actual event date. On average, the acquisition happens 1 day and 21 hours later than the reported event date. Further technical details about the metadata are available in the main repository.[3] Focusing on the semantic segmentation task, we keep the original image size for each area of interest, while making sure that the minimum input dimensions of 512×512 are kept across the whole set. The smallest image size is 531×524 pixels, while the largest is $1,944 \times 1,944$ pixels.

For training and benchmark purposes, the set has also been split into images for training, validation, and testing. To ensure a robust subdivision, data has been randomly split based on the EMS activation with geographical awareness, meaning that each set has a proportionally equal number of examples for different locations around the globe, thus preventing biases due to specific land types or geographical areas. Specifically, the three splits, whose location is visible in Fig. 2, comprise 54 activations for the training set, 34 for the test set, and 7 for the validation set.

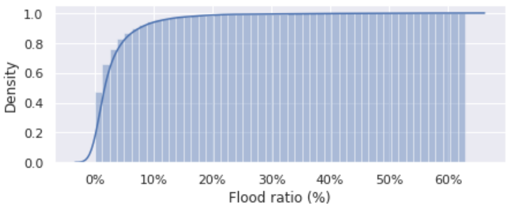


FIGURE 4. Cumulative distribution function of the flooded areas ratios over the full MMFlood dataset.

C. DATASET CHALLENGES

Despite the relatively simple downstream task of binary segmentation, the dataset provides several challenges that need to be addressed for an effective training. First, areas affected by floods usually represent a small portion of the segmentation mask, because of the inherent unbalance of the actual flooded area within the image extremely skewed towards background pixels, especially when the ordinary water basins are excluded in the final mask. As shown by the cumulative distribution function in Fig. 4, 50% of the images have less than 1.4% of pixels marked as flood, and 95% of the dataset consists of images that contain less than 11% of flooded area. To tackle this recurring issue in remote sensing scenarios, we test the effectiveness of different solutions: we first adopt a simple, albeit effective, filtering and down-sampling approach based on flood-background ratio with different thresholds, comparing it with a dynamic, entropy-based weighted sampling, to cope with lack of information in labels with mostly background or flood pixels.

[3] https://github.com/edornd/mmflood

A second challenge is given by the SAR data itself, as radar signals are inherently noisy, therefore a proper pre-processing pipeline should typically involve radiometric correction based on elevation maps and speckle noise reduction, to obtain an optimal image for further downstream tasks [13]. We purposely avoid these manuals steps for two main reasons: first, we aim to provide a dataset with clean and coherent data, while at the same time maximising its reusability in other contexts, and refraining from biasing the final product towards a specific set of applications. Second, we expect deep learning algorithms to extract similar features on their own during training, when provided with the necessary information. As detailed in the following sections, we cope with the noisy images by means of random augmentations, while testing the effectiveness of the multimodal setting exploiting the additional DEM data.

Last, another challenging aspect of the proposed dataset is the provision of multiple modalities. Using SAR data through VV and VH polarisations, in combination with the elevation raster, introduces the problem of having input sources of different nature. In this paper, we test several combinations with and without DEM in order to assess the advantage brought by the additional modality, by simply expanding the input weights [9], or further refining the model structure to support early and late fusion [54], to better merge modality-specific features. The following sections provide further details regarding the methodologies and training configurations adopted, together with the experimental results on the selected baselines and more complex scenarios.

IV. METHODOLOGY

We focus the evaluation of the MMFlood dataset on deep learning models, considering only SAR and DEM as training modalities, as the hydrography can be exploited as post-processing step. The task of flood delineation can be described as a binary classification problem, where the objective is to classify each pixel as *flood* or *background*. Formally, we can define a set of samples X, containing pairs of images x_s and x_d, respectively representing SAR and DEM inputs, with matching and constant dimensions $H \times W$, that are associated with a set of labels $y \in Y$ with the same dimensions. For each pixel i, the image label provides an annotation $y_i = \{0, 1\}$, where 0 indicates background cells, while 1 refers to flooded areas. A binary segmentation training can be formalised as training a model f_θ with parameters θ to map from the image space to the label space, namely $f_\theta : X \to \mathbb{R}^{|H \times W|}$. In the multimodal context, we provide a model with two separate encoders, g_s and g_d, then fuse the extracted features through a shared decoder h, to obtain a final model $f = h(g_s(x_s) \oplus g_d(x_d))$. In the following sections we describe the methodology adopted to address the highlighted challenges, focusing on class unbalance and multiple modalities.

A. BASELINE MODELS

With the aim of providing benchmark results as baseline, we first select and test different combinations of encoders and decoders, to assess the effectiveness of different models in this context. Among the encoders, we adopt ResNet50 [55], the *de-facto* standard choice in many segmentation frameworks, TResNet [56] a ResNet variant that aims to boost accuracy and efficiency in both training and inference, EfficientNet-V1 [57], a more lightweight approach that scales all dimensions using a compound coefficient, and DenseNet [58], where all layers are connected with matching feature-map sizes to provide more semantically rich layers. Among state-of-the-art decoders, we include U-Net [59], a standard solution in aerial imagery initially proposed for datasets in the medical field [60], which reverses the computation back to the pixel level using a mirrored upscaling pipeline. In terms of decoders based around mult-scale feature extraction, we include DeepLabV3+ [52], based around atrous convolutions and spatial pyramid pooling, and PSPNet [61], which utilises a pyramid parsing module to exploit global and local clues.

B. CLASS UNBALANCE

The unbalance between flood and background pixels represents the first and major challenge of this dataset. We test two complementary approaches to address this, respectively based on downsampling and upsampling: in the first case, we simply impose a threshold on the percentage of flood pixels in each tile, therefore avoiding the selection of images almost completely covered with background and focusing on those tiles rich of information. More formally, from the final list of preprocessed tiles, as described in Sec. V, we filter out images having a flood pixel ratio under a threshold τ, computed as:

$$\tau = \frac{\sum_i [y_i = 1]}{\sum_i [y_i]} \tag{1}$$

where y_i represents the label associated with a generic pixel i. Simply put, the ratio of pixels marked as *flood* with respect to the total amount of pixels in the label.

In the second case, we better exploit the sparsity of the dataset by considering the entropy of the available labels. In information theory, the latter can be interpreted as *amount of information*, which can also translate into how much the given data sample is effective in terms of content. The larger the entropy, the higher information content is contained [62]. This property can be exploited to provide an estimation of the importance of each label with respect to the training process, meaning that those annotations with higher entropy are more likely to be informative for the given task. Similar to [62], we exploit this property to perform a standard weighted sampling, where however each sample is weighted by its content information computed as:

$$w_j = \lambda \left(- \sum_y p(y) \log_2 p(y) \right), \quad \forall j \in |X| \tag{2}$$

where $p(y)$ represents the distribution of pixels among the two available classes in the label y, λ represents a modulating

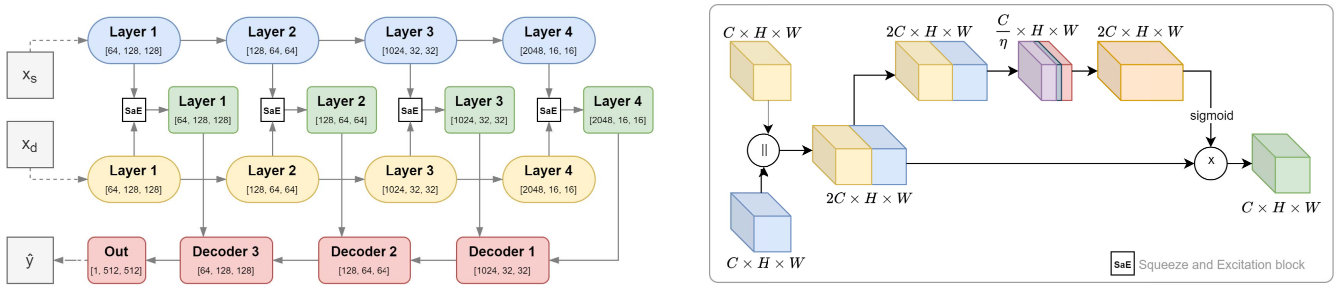


FIGURE 5. Multi-encoder architecture: two twin, separate encoders (blue, yellow) are merged layer-wise through Squeeze-and-Excitation blocks to produce an equivalent multimodal layer (green). The decoder (red) is agnostic w.r.t. the multi-encoder setup.

factor $0 \leq \lambda \leq 1$, and j uniquely identifies each item of the set of samples X.

C. MULTIPLE MODALITIES

A second challenge of this dataset is the inclusion of the DEM for training purposes. Given the matching spatial extents of SAR and DEM modalities, we can consider this additional information as extra channel, to combine with the satellite acquisitions. As for the unbalance problem, we test two increasingly complex approaches, namely input channel expansion and multi-encoder networks, to assess the benefits of including this additional data as input. As baseline, we simply provide every SAR band and the extra DEM image as a 3-channel image, training the network from scratch and letting it extract features from each modalities independently.

For a more effective approach, we construct a custom residual U-Net network [63], including two separate, lightweight twin encoders, one for each modality. Similar to [54], we provide both early and late fusion by means of a Squeeze-and-Excitation block (SaE) [64]. At each layer i, the encoders provide a comparable tensor with shape $H_i \times W_i \times C_i$. As shown in Fig. 5 (right), the SaE module merges these outputs of the encoders by first concatenating them channel-wise to produce an output with size $H_i \times W_i \times 2\ C_i$, which are then calibrated through a weight map, generated by the second branch of the module. The weights are obtained by first *squeezing* the input into a channel descriptor with $\frac{C_i}{\eta}$ channels, and subsequently expanding it back to $2C_i$. This tensor is eventually fed to a Sigmoid activation, that scales each value in the range [0, 1], and multiplied by the original inputs. Last, the $2C_i$ channels are reduced to the original shape $H_i \times W_i \times C_i$, so that the final output becomes virtually indistinguishable from the single-encoder ones, and remains perfectly compatible with the decoder, regardless of the encoding phase. An example schema of multi-encoder architecture is displayed in Fig. 5 (left).

V. EXPERIMENTS

In this section, we provide an overview of the experimental tests carried out on the MMFlood dataset and their results. We first detail the initial preprocessing steps, in order to transform the available images into an actual model input. Then we describe the preliminary benchmark tests, with the purpose of finding the best performing encoder-decoder combination for subsequent tests. Last, we discuss the inclusion of the additional components described in Sec. IV, namely Entropy Weighted Sampling (EWS), the inclusion of DEM as extra channel, and the multi-encoder setup, highlighting strengths and weaknesses of the proposed approaches.

A. IMPLEMENTATION DETAILS

Given the purposely generic dataset, we first process the available images to obtain a set of data more suitable for training. Considering SAR imagery, we compute a base-10 logarithm over the acquisitions to obtain decibel values and reduce the effects of high backscatter peaks. Concerning the DEM image, since the values are expressing the raw altitude in meters, we clip the available values within reasonable values in the range $[-100, 6,000]$ to reduce noise. In every experiment, we further normalise each input by first computing mean and standard deviation for each channel.

Because of large images with various scales, we resort to offline tiling to both speed up the training process and maintain a standard setup. From each available image and modality, we extract the smallest possible number of tiles covering the whole surface of the original image, considering a window size of 512 × 512, which is a common standard in most semantic segmentation applications [9], [52], as it provides a good compromise between pixel-wise class balance and contextual information. Since the image extent is often incompatible with the tile size, we dynamically overlap by the minimal amount of pixels required to cover the whole image without overflow, thus avoiding padding and maximising the visual content of each window. Excluding the test partition, the training dataset contains a total of 6,182 tiles, together with 560 tiles for validation.

With the aim of maintaining the correct pixel count, we only apply this procedure to the training and validation sets, while keeping the full test images: during the test phase, we apply an online tiling with fixed overlap, combining the predictions into a single output with the same extents of the inputs.

We also apply several online data augmentations to avoid overfitting at training time. In case of multiple modalities, every input is first processed by a common set of transformations, namely random rotation, cropping, horizontal and vertical flipping, grid and elastic transforms, each with probability $p = 0.5$. Then, we apply additional pixel-level transformations to SAR images, dealing with its inherent speckle noise by introducing random Gaussian blur and random multiplicative noise, alternately.

In every experiment, we train for 100 epochs and early stopping criterion with patience set to 30 epochs. Given the peculiar inputs, we do not use any pretrained weights. We use a batch size of 16, except for the combinations of UNet with ResNet50 and DenseNet121, where the batch was reduced to 12 images for memory constraints. Moreover, given the stride incompatibility of the outputs provided by DenseNet with the DeepLabV3+ decoder, we did not include such variant in the baseline result. We adopt an AdamW optimiser with a base learning rate $\lambda = 10^{-3}$ and a weight decay coefficient of 0.01. We also employ a polynomial LR scheduler with $\gamma = 3$ and ending learning rate $\lambda = 10^{-4}$.

Considering the loss function, we perform every experiment using a focal Tversky loss [65], a generalisation of the soft Dice score with the addition of a focal component, particularly suited for imbalanced problems. The loss was configured with hyperparameters $\alpha = 0.6$, $\beta = 0.4$ and $\gamma = 2$. To provide a complete overview, four metrics are considered: we report the mean value over all test images for Precision, Recall, IoU, and F1 Score (with equal weighting of all chips).

All the procedures and experiments described in this paper were performed on a workstation equipped with an Intel Core Intel(R) Xeon(R) Silver 4216 CPU and 4 Nvidia GTX 2080Ti GPUs. The framework is based on Python, exploiting the *PyTorch* library for training and testing.

B. RESULTS

1) BASELINES

We first test the effectiveness of different combinations of encoders and decoders, to test the best combination for subsequent experiments. To assess the feasibility of the task without including additional information, we train on SAR images only, using a flood threshold of 2% to maximise the performances (see Section V-C). Results for this set are shown in Table 4. We observe similar performances across most decoders, with the most relevant differences dictated by the encoder choice: ResNet variants (RN50, TResNet) display the best performance with 0.63 and 0.59 mIoU respectively. On the other hand, EfficientNet and DenseNet variants did not achieve optimal results for this specific task, with a maximum score of 0.56 mIoU for the latter, despite the added complexity. This may be due to several factors, such as the relatively smaller amount of data when compared to natural images domain, where these models typically outperform the standard ResNet. While this performance gap may be closed exploiting more hardware and data resources to effectively train more advanced architectures, we leverage on ResNet50 for subsequent experiments for simplicity, as it remains a competitive solution to this day [66].

Considering decoders, performance differences are generally negligible, however, both UNet and DeepLabV3+ reach comparable results, with 0.63 mIoU on average, while PSPNet remains a solid choice, despite the 1% performance loss, providing the most consistent results across the experiments. While we note that the performances with other decoders are comparable, for simplicity we maintain the combination of ResNet50 and DeepLabV3+ as model architecture for future tests given the obtained results, except for the multi-modal setting where we employ a TResNet-m variant for performance reasons. Overall, deep learning approaches appear to provide much more robust results when compared to the Otsu baseline with a +0.3 increment in most cases, as also confirmed by the qualitative results in Fig. 6, where the latter is noticeably noisier than the counterparts. This issue could be mitigated with careful processing, however we note that neural networks were able to cope with the same noisy inputs without additional steps.

TABLE 4. Baseline results on the test set, considering different combinations of encoders and decoders, without including additional modalities.

Model	Precision	Recall	IoU	F1
Otsu	0.2895	0.4627	0.1963	0.2895
ResNet50 + UNet	0.6910	0.8710	0.6269	0.7706
ResNet50 + DeepLabV3+	0.6733	0.9031	**0.6279**	**0.7714**
ResNet50 + PSPNet	0.6659	0.8858	0.6132	0.7603
TResNet + UNet	0.6151	**0.9178**	0.5830	0.7366
TResNet + PSPNet	**0.7376**	0.7462	0.5897	0.7419
TResNet + DeepLabV3+	0.6331	0.6299	0.4614	0.6315
EfficientNet + UNet	0.6856	0.6242	0.4853	0.6534
EfficientNet + DeepLabV3+	0.4491	0.2050	0.1638	0.2815
EfficientNet + PSPNet	0.7143	0.6211	0.4976	0.6645
DenseNet121 + PSPNet	0.6050	0.8967	0.5656	0.7225
DenseNet121 + UNet	0.5954	0.9054	0.5605	0.7184

2) DEM AND CLASS BALANCE

Experiments considering different combinations of modalities and class unbalance countermeasures are shown in Table 5. We observe that the sole inclusion of the DEM as extra input, without further processing or considerations, is enough to improve the performance of the baseline model by 1.5%, with 0.63 mIoU, maintaining the flood threshold $\tau = 0.02$. This shows the importance of including terrain information in this context, where elevation and the nominal extents of water basins is crucial for flood segmentation. These results are confirmed and further improved by including EWS, where the performance improvement is more substantial, reaching 0.65 mIoU. This demonstrates the importance and the effectiveness of the guided sampling of the most informative labels during training. While recall is maintained at 0.9 across most experiments, results are more remarkable when considering precision, going from 0.67 to 0.71 after the inclusion of DEM and EWS.

TABLE 5. Results obtained from the incremental addition of our proposed solutions, starting from the single encoder (SE), single encoder with DEM as extra channel, to the entropy-weighted sampling (EWS), and the multi-encoder model (ME) with all the additions.

Model	Thresh.	DEM	EWS	Precision	Recall	IoU	F1
SE	2%			0.6733	**0.9031**	0.6279	0.7714
SE	2%	✓		0.6814	0.8979	0.6324	0.7748
SE	×		✓	0.6976	0.8955	0.6451	0.7843
SE	×	✓	✓	0.7173	0.8893	0.6585	0.7941
ME	×	✓	✓	**0.7319**	0.8794	**0.6652**	**0.7989**

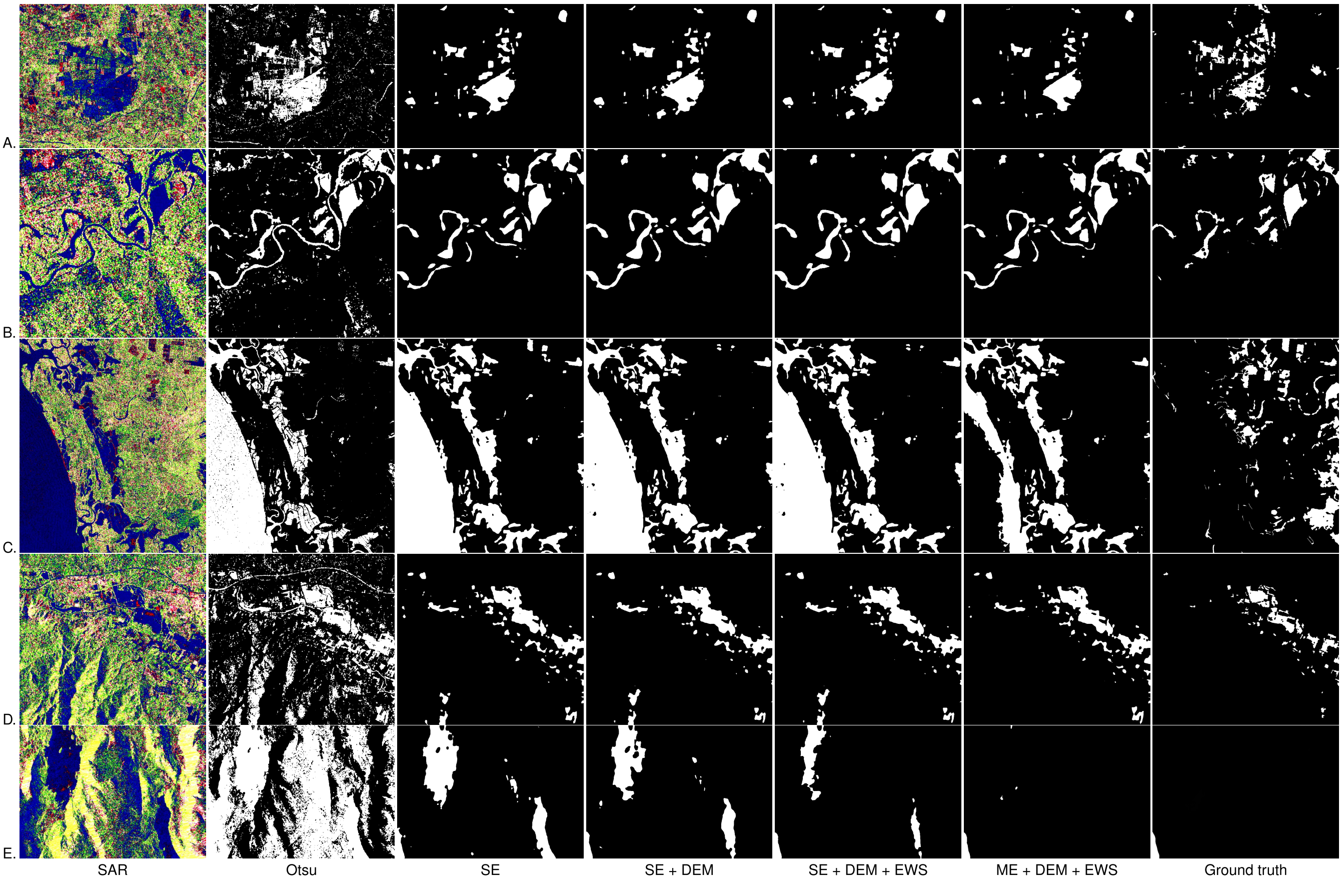


FIGURE 6. Qualitative results obtained on the test set. From left to right, Otsu baseline, single encoder on SAR data only (SE), single encoder with DEM (SE+DEM), single encoder with DEM and EWS (SE + DEM + EWS), the multi-encoder setup (ME + DEM + EWS), and the ground truth.

TABLE 6. Study on the influence of the downsampling threshold and the effectiveness of the entropy-based sampling (EWS).

Thresh.	EWS	Precision	Recall	IoU	F1
0%	×	0.4118	0.2274	0.1717	0.2930
2%	×	0.6733	0.9031	0.6279	0.7714
5%	×	0.6301	**0.9097**	0.5930	0.7445
0%	✓	**0.6976**	0.8955	**0.6451**	**0.7843**
2%	✓	0.6700	0.9024	0.6247	0.7690
5%	✓	0.6486	0.9096	0.6094	0.7573

Last, we also test the multi-encoder architecture, reaching more than 0.66 mIoU and surpassing every other variant, notwithstanding the encoder change. As shown by the qualitative results in Fig. 6, despite the small performance differences the multi-encoder setup appears more resilient to terrain changes, where shadows on mountain ranges or large permanent water basins could be erroneously classified as floods (Fig. 6-D, 6-E).

C. SAMPLING VARIATION STUDY

We perform a variation study to assess the effectiveness of different combinations of sampling procedures over the MMFlood dataset, without including the elevation model because of the focus on sampling. We test four different downsampling thresholds, namely 0% meaning no downsampling, 2% and 5% to test the effectiveness of the reduced training size on the test set. We further experiment with each variant with and without EWS, effectively performing a mix of oversampling over informative inputs, and downsampling tiles with mostly background beforehand. For every experiment listed in Table 6, we maintain the same configurations and hyperparameters as described in Sec. V-A. Considering the variants without sampling, results confirm that the optimal threshold for this context remains $\tau = 0.02$: lower values undermine the performance of the model, while higher values slowly introduce a bias towards flooded areas with less precise predictions, decreasing from 0.67 to 0.63 in the 5% setup. Including the entropy-based sampling introduces a strong performance boost without thresholds, improving over the baseline by 4.5% with 0.64 mIoU, while leaving the remaining variants almost unchanged. This highlights the effectiveness of the sampling solution, which is most likely constrained by the lower number of images in the threshold-based approaches, and works best when applied to the whole set.

VI. CONCLUSION

We introduced MMFlood, an ad-hoc dataset for flood delineation tasks, generated from 1,748 Sentinel-1 acquisitions based exclusively on flooded areas delineated from 95 Copernicus EMS flood events around the globe. MMFlood builds on top of other comparable datasets available in literature in terms of images and global distribution, the multiple modalities of the dataset, including also elevation model and hydrography, and the use of masks which are specific for the task of flood delineation, without including irrelevant water basins. Considering future developments, MMFlood could be expanded in several ways. First, new images from future flood events mapped by Copernicus EMS could be incrementally added. Second, by adding other Copernicus satellites such as Sentinel-2. Third, including all the available acquisitions within the temporal span of the flood events. Lastly, by considering co-registered VHR optical and SAR satellite networks, such as TerraSAR-X [35], as additional modalities. We provided an extensive benchmark applying state-of-the-art solutions, to evaluate the effectiveness of MMFlood in the context of binary segmentation, considering both classical thresholding and recent deep learning approaches. Furthermore, we proposed various solutions to common aerial segmentation problems such as class unbalance and the exploitation of the additional modalities by means of a multi-encoder architecture. The inclusion of the elevation model provided a consistent boost in performance in every setup, especially in the multi-encoder setup. In future works, the model performances could be further improved investigating different multimodal architectures, as proposed by recent works on RGB-D [67], [68], which have a similar problem setting. Additionally, future experiments could also investigate the use of weights pretrained on natural images, as bootstrap strategy for training. While aimed at completely different domains, an effective knowledge transfer from recent large-scale models such as Vision Transformers [69] may provide better results.

APPENDIX
COPERNICUS EMS ACTIVATIONS

See Table 7.

TABLE 7. Complete list of Copernicus EMS activations included and selected for the construction of the MMFlood dataset.

EMSR107	EMSR260	EMSR342	EMSR450
EMSR117	EMSR261	EMSR358	EMSR456
EMSR118	EMSR265	EMSR388	EMSR465
EMSR120	EMSR267	EMSR397	EMSR467
EMSR122	EMSR268	EMSR399	EMSR468
EMSR141	EMSR271	EMSR407	EMSR470
EMSR147	EMSR273	EMSR410	EMSR471
EMSR149	EMSR275	EMSR411	EMSR479
EMSR150	EMSR277	EMSR414	EMSR487
EMSR151	EMSR279	EMSR416	EMSR492
EMSR154	EMSR280	EMSR417	EMSR496
EMSR156	EMSR283	EMSR419	EMSR497
EMSR162	EMSR284	EMSR421	EMSR498
EMSR165	EMSR287	EMSR422	EMSR501
EMSR166	EMSR293	EMSR424	EMSR502
EMSR167	EMSR314	EMSR427	EMSR504
EMSR184	EMSR319	EMSR429	EMSR507
EMSR187	EMSR321	EMSR437	EMSR511
EMSR192	EMSR324	EMSR438	EMSR514
EMSR199	EMSR326	EMSR441	EMSR517
EMSR203	EMSR330	EMSR442	EMSR518
EMSR215	EMSR332	EMSR444	EMSR520
EMSR238	EMSR333	EMSR445	EMSR548
EMSR258	EMSR337	EMSR446	

REFERENCES

[1] V. Masson-Delmotte *et al.*, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, U.K.: Cambridge Univ. Press, 2021.

[2] W. Du, G. J. FitzGerald, M. Clark, and X.-Y. Hou, "Health impacts of floods," *Prehospital Disaster Med.*, vol. 25, pp. 265–272, Jun. 2010.

[3] S. N. Jonkman, "Global perspectives on loss of human life caused by floods," *Natural Hazards*, vol. 34, pp. 151–175, Feb. 2005.

[4] L. Carrera, G. Standardi, E. Koks, L. Feyen, J. Mysiak, J. Aerts, and F. Bosello, "Economics of flood risk in Italy under current and future climate, CMCC Found., Italy, Tech. Rep. 272, 2015.

[5] J. Pollner, *Financial and Fiscal Instruments for Catastrophe Risk Management: Addressing the Losses From Flood Hazards in Central Europe*, Washington, DC, USA: World Bank, 2012.

[6] P. C. Oddo and J. D. Bolten, "The value of near real-time Earth observations for improved flood disaster response," *Frontiers Environ. Sci.*, vol. 7, p. 127, Sep. 2019, doi: 10.3389/FENVS.2019.00127.

[7] B. Neupane, T. Horanont, and J. Aryal, "Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis," *Remote Sens.*, vol. 13, no. 4, p. 808, Feb. 2021.

[8] G. Tseng, I. Zvonkov, C. L. Nakalembe, and H. Kerner, "CropHarvest: A global dataset for crop-type classification," in *Proc. 35th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 1–14.

[9] M. T. Chiu, X. Xu, Y. Wei, Z. Huang, A. G. Schwing, R. Brunner, H. Khachatrian, H. Karapetyan, I. Dozier, G. Rose, D. Wilson, A. Tudor, N. Hovakimyan, T. S. Huang, and H. Shi, "Agriculture-vision: A large aerial image database for agricultural pattern analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2825–2835.

[10] G. Palomba, A. Farasin, and C. Rossi, "Sentinel-1 flood delineation with supervised machine learning," in *Proc. 17th Int. Conf. Inf. Syst. Crisis Response Manage. (ISCRAM)*, 2020, pp. 1072–1083.

[11] S. Monaco, S. Greco, A. Farasin, L. Colomba, D. Apiletti, P. Garza, T. Cerquitelli, and E. Baralis, "Attention to fires: Multi-channel deep learning models for wildfire severity prediction," *Appl. Sci.*, vol. 11, no. 22, p. 11060, 2021.

[12] J. C. Curlander and R. N. McDonough, *Synthetic Aperture Radar*, vol. 11. New York, NY, USA: Wiley, 1991.

[13] A. Ajmar, "Response to flood events: The role of satellite-based emergency mapping and the experience of the Copernicus emergency management service," *Flood Damage Surv. Assess. Insights Res. Pract.*, vol. 228, pp. 213–228, Jun. 2017.

[14] H. A. Ganaie, H. Hashaia, and D. Kalota, "Delineation of flood prone area using normalized difference water index (NDWI) and transect method: A case study of Kashmir valley," *Int. J. Remote Sens. Appl.*, vol. 3, no. 2, pp. 53–58, 2013.

[15] S. Martinis, A. Twele, and S. Voigt, "Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution TerraSAR-X data," *Natural Hazards Earth Syst. Sci.*, vol. 9, no. 2, pp. 303–314, 2009.

[16] *Interagency Implementation, NASA Advanced ConceptsTyler, and IEEE GRSS Earth Science Informatics Technical Committee. ETCI 2021 Competition on Flood Detection*, NASA, Washington, DC, USA, 2021.

[17] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1–11.

[18] *Copernicus Emergency Management Service*, European Union, Brussels, 2022.

[19] MapZen. (2022). *Terrain Tiles*. [Online]. Available: https://registry.opendata.aws/terrain-tiles/

[20] V. Klemas, "Remote sensing of floods and flood-prone areas: An overview," *J. Coastal Res.*, vol. 314, pp. 1005–1013, Jul. 2015.

[21] I. Ogashawara, M. P. Curtarelli, and C. M. Ferreira, "The use of optical remote sensing for mapping flooded areas," *J. Eng. Res. Appl.*, vol. 3, pp. 1956–1960, 2013.

[22] OpenStreetMap Contributors. (2017). *Hydrography Obtained From*. [Online]. Available: https://planet.osm.org and https://www.openstreetmap.org

[23] *International Society for Photogrammetry and Remote Sensing*. Datasets, Potsdam, Germany, 2013.

[24] G. Sumbul, A. D. Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl, "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, Sep. 2021.

[25] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[26] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[27] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, "Creating xBD: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 10–17.

[28] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[29] C. Rambour. (Sep. 2020). *Sen12-Flood : A SAR and Multispectral Dataset for Flood Detection*. [Online]. Available: https://ieee-dataport.org

[30] N. Chaouch, M. Temimi, S. Hagen, J. Weishampel, S. Medeiros, and R. Khanbilvardi, "A synergetic use of satellite imagery from SAR and optical sensors to improve coastal flood mapping in the Gulf of Mexico," *Hydrol. Processes*, vol. 26, pp. 1617–1628, Sep. 2011.

[31] J. Herbert Kramer, *Observation of the Earth and Its Environment*. Berlin, Germany: Springer, 2002.

[32] R. K. Raney, A. P. Luscomber, E. J. Langham, and S. Ahmed, "RADARSAT (SAR imaging)," *Proc. IEEE*, vol. 79, no. 6, pp. 839–849, Jun. 1991.

[33] F. Caltagirone, G. Angino, A. Coletta, F. Impagnatiello, and A. Gallon, "COSMO-SkyMed program: Status and perspectives," in *Proc. 3rd Int. Workshop Satell. Constellations Formation Flying*, 2003, pp. 11–16.

[34] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, and M. Brown, "GMES Sentinel-1 mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, May 2012.

[35] K. Voormansik, J. Praks, O. Antropov, J. Jagomagi, and K. Zalite, "Flood mapping with TerraSAR-X in forested regions in Estonia," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 562–577, Feb. 2014.

[36] J. Lu, L. Giustarini, B. Xiong, L. Zhao, Y. Jiang, and G. Kuang, "Automated flood detection with improved robustness and efficiency using multi-temporal SAR data," *Remote Sens. Lett.*, vol. 5, pp. 240–248, Mar. 2014.

[37] I. Ali, V. Freeman, S. Cao, and W. Wagner, "Sentinel-1 based near-real time flood mapping service," in *Proc. ISCRAM*, 2018.

[38] A. Twele, W. Cao, S. Plank, and S. Martinis, "Sentinel-1-based flood mapping: A fully automated processing chain," *Int. J. Remote Sens.*, vol. 37, no. 13, pp. 2990–3004, 2016.

[39] L. Pulvirenti, N. Pierdicca, M. Chini, and L. Guerriero, "An algorithm for operational flood mapping from synthetic aperture radar (SAR) data using fuzzy logic," *Natural Hazards Earth Syst. Sci.*, vol. 11, pp. 529–540, Feb. 2011.

[40] S. Martinis, J. Kersten, and A. Twele, "A fully automated TerraSAR-X based flood service," *ISPRS J. Photogramm. Remote Sens.*, vol. 104, pp. 203–212, Jun. 2015.

[41] G. Ireland, M. Volpi, and P. G. Petropoulos, "Examining the capability of supervised machine learning classifiers in extracting flooded areas from landsat TM imagery: A case study from a Mediterranean flood," *Remote Sens.*, vol. 7, pp. 3372–3399, Mar. 2015.

[42] A. Benoudjit and R. Guida, "A novel fully automated mapping of the flood extent on SAR images using a supervised classifier," *Remote Sens.*, vol. 11, p. 779, Apr. 2019.

[43] W. Kang, Y. Xiang, F. Wang, L. Wan, and H. You, "Flood detection in Gaofen-3 SAR images via fully convolutional networks," *Sensors*, vol. 18, p. 2915, Sep. 2018.

[44] N. Kussul, A. Shelestov, and S. Skakun, "Flood monitoring from SAR data," in *Use of Satellite and In-Situ Data to Improve Sustainability*, F. Kogan, A. Powell, and O. Fedorov, Eds. Dordrecht, The Netherlands: Springer, 2011, pp. 19–29.

[45] C. Bayik, S. Abdikan, G. Ozbulak, T. Alasag, S. Aydemir, and F. B. Sanli, "Exploiting multi-temporal Sentinel-1 SAR data for flood extend mapping," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, vol. 42, no. 3, p. W4, 2018.

[46] L. Lopez-Fuentes, C. Rossi, and H. Skinnemoen, "River segmentation for flood monitoring," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3746–3749.

[47] F. Lattari, B. G. Leon, F. Asaro, A. Rucci, C. Prati, and M. Matteucci, "Deep learning for SAR image despeckling," *Remote Sens.*, vol. 11, no. 13, p. 1532, Jun. 2019.

[48] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Speckle2Void: Deep self-supervised SAR despeckling with blind-spot convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.

[49] Y.-L. Chang, A. Anagaw, L. Chang, Y. Wang, C.-Y. Hsiao, and W.-H. Lee, "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, vol. 11, no. 7, p. 786, Apr. 2019.

[50] H. Lang, S. Wu, and Y. Xu, "Ship classification in SAR images improved by AIS knowledge transfer," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 439–443, May 2018.

[51] A. Galdelli, A. Mancini, C. Ferrá, and A. N. Tassetti, "A synergic integration of AIS data and SAR imagery to monitor fisheries and detect suspicious activities," *Sensors*, vol. 21, no. 8, p. 2756, 2021.

[52] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," 2016, *arXiv:1606.00915*.

[53] T. G. Farr, P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, and L. Roth, "The shuttle radar topography mission," *Rev. Geophys.*, vol. 45, no. 2, 2007.

[54] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *Int. J. Comput. Vis.*, vol. 128, pp. 1239–1285, Jul. 2020.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[56] T. Ridnik, H. Lawen, A. Noy, E. B. Baruch, G. Sharir, and I. Friedman, "TresNet: High performance GPU-dedicated architecture," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1400–1409.

[57] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[58] G. Huang, Z. Liu, and Q. K. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[60] S. A. Sriram, A. Paul, Y. Zhu, V. Sandfort, J. P. Pickhardt, and M. R. Summers, "Multilevel UNet for pancreas segmentation from non-contrast CT scans through domain adaptation," *Proc. SPIE*, vol. 11314, Mar. 2020, Art. no. 113140K.

[61] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2016, *arXiv:1612.01105*.

[62] L. Li, H. He, and J. Li, "Entropy-based sampling approaches for multi-class imbalanced problems," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 11, pp. 2159–2170, Nov. 2020.

[63] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet—A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[64] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[65] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.

[66] R. Wightman, H. Touvron, and H. Jégou, "ResNet strikes back: An improved training procedure in timm," 2021, *arXiv:2110.00476*.

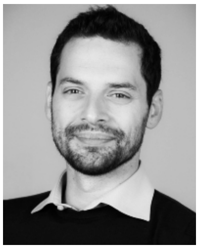
[67] R. Girdhar, M. Singh, N. Ravi, L. V. D. Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16102–16112.

[68] H. Liu, J. Zhang, K. Yang, X. Hu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," 2022, *arXiv:2203.04838*.

[69] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Austria, May 2021. [Online]. Available: https://openreview.net/group?id=ICLR.cc/2021/Conference

FABIO MONTELLO received the M.S. degree *(summa cum laude)* in data science from La Sapienza, University of Rome, in 2020. He has an experience in web development while working for a start-up based in Auckland, New Zealand. Since 2020, he has been a Applied Researcher with the AI, Data and Space Department, LINKS Foundation, Turin, Italy. His research interests include computer vision, remote sensing, and security applications.

EDOARDO ARNAUDO received the M.S. degree *(summa cum laude)* in computer science with specialisation in machine learning and artificial intelligence from the Università degli Studi di Torino (UniTo), in 2019. He is currently pursuing the Ph.D. degree in computer vision. In 2019, he started as Applied Researcher with AI, Data & Space, LINKS Foundation, Turin, Italy, in several multidisciplinary projects, with a focus on aerial and remote sensing applications. His research interest includes semantic segmentation applied to aerial and satellite imagery.



CLAUDIO ROSSI received the M.S. degree in electrical and computer engineering from the University of Illinois at Chicago (UIC), in 2005, and the graduate degree *(summa cum laude)* in electronics Politecnico di Torino, where he received the Ph.D. degree, in 2014. He was a Software Analyst at Consorzio per il Sistema Informativo (CSI), Piedmont, Italy, and a Project Manager at Fiat Group Automobiles (FGA). He was a Research Intern at Telefonica I+D, Politecnico di Torino. Currently, he is a Program Manager with LINKS Foundation, Turin, where he leads the research program AI for Industry and Security, coordinating several research and development projects from the Horizon 2020 Programme.