Land Cover Mapping from Multiple Complementary Experts under Heavy Class Imbalance

Valerie Zermatten, *Student Member, IEEE*, Xiaolong Lu, Javiera Castillo-Navarro, *Member, IEEE*, Tobias Kellenberger, Devis Tuia, *Senior Member, IEEE*,

***Abstract*—Deep learning has emerged as a promising avenue for automatic mapping, demonstrating high efficacy in land cover categorization through various semantic segmentation models. Nonetheless, the practical deployment of these models encounters important challenges from the imbalanced distribution of samples between the classes, a problem inherent to real-world datasets. This results in models biased towards frequent classes that perform poorly on rare classes. While existing approaches to fight class imbalance mainly focus on image classification, here we propose to address this issue for semantic segmentation with a multiple complementary experts (MCE) structure. Taking inspiration from ensemble models, each expert in our MCE specializes in certain classes and works with other experts in a complementary manner to generate robust predictions for rare classes. We compare our approach to other existing methods and also explore different logit aggregation methods, to identify the performance upper bounds and improvement directions. Our model is evaluated on a large-scale and challenging alpine land cover dataset that we make openly available. Additionally, we evaluated our model on an imbalanced land cover mapping dataset, FLAIR, to highlight its adaptability. Overall, our MCE model yields notable improvement in performances on the medium and rare classes compared to baseline methods, while only slightly compromising on the overall accuracy. Despite its simplicity, the MCE approach stands as a practical solution for more operational semantic segmentation models, not trading off performances on rare but important classes.**

***Index Terms*—Class imbalance, land cover mapping, remote sensing, multi-expert model.**

I. Introduction

LAND cover (LC) mapping provides information about the characteristics and spatial distribution of the Earth's surface. It plays a crucial role in many scientific and operational applications, including environmental monitoring, natural resources management, planning, disaster management and climate change research [1]. The ever-increasing amount of generated data created a shift in LC map production from labour-intensive human annotations towards machine-driven products. In the past decades, the development of LC maps with machine learning algorithms has drawn considerable attention from the scientific community. Computer vision methods and more specifically deep learning (DL) based models exhibited outstanding performance [2] when confronted with visual recognition tasks such as LC mapping [3], classification [4] or object detection on remote sensing images [5].

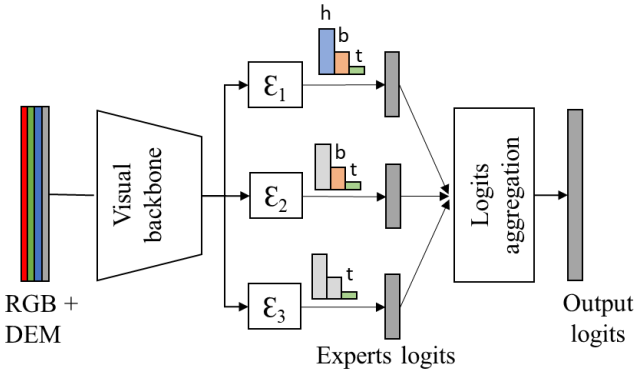


Fig. 1: Multiple complementary expert (MCE) semantic segmentation model. First, a common backbone extracts visual features from the input images; then each expert's model predicts land cover classes, with one expert focusing on all classes, while other experts focus on body and tail or tail-only classes. The extracted features from each expert are combined via different aggregation methods to produce the final network predictions.

While proficient on carefully balanced datasets with numerous samples, the effectiveness of DL methods diminishes when confronted with unequal distributions of instances over different classes [6], [7], [8]. This issue originates from the classifier's objective of minimizing overall error rates during training: Head classes dominate the network parameters update, which tends to greatly diminish the recognition of tail classes by the model [6], [9]. This problem called class imbalance is an intrinsic problem for semantic segmentation, and especially for land cover mapping: LC classes naturally have very different surface coverage; some appear very frequently and cover large areas, i.e. forested areas, whereas other LC like wetlands have a limited frequency of occurrence and/or only cover small proportions of the territory. This leads to an asymmetry in the number of pixels representing each LC type, with the number of pixels of rare or 'tail' classes being lower by several orders of magnitude than pixels belonging to frequent or so-called 'head' classes. Nevertheless, rare classes hold particular significance for LC mapping since they often represent specialized and unique regions of the territory that may be of high interest, for instance, biodiversity hot spots [10], or indicators of new patterns appearing in the

V. Zermatten, J. Castillo-Navarro and D. Tuia are with the Laboratory for Environmental Computational Science and Earth Observation, EPFL.

X. Lu is with the Dept. of Civil, Environmental and Geomatic Engineering, ETH Zurich and the Laboratory for Environmental Computational Science and Earth Observation, EPFL.

T. Kellenberger is with the Federal Office of Topography swisstopo.

landscape [4].

The class imbalance problem is a well-known problem in computer vision and numerous methods have been proposed to tackle it. The most intuitive approaches address the problem at the data level by deploying data re-sampling [7] or re-weighting methods [11], [12], [13] that give more visibility to samples from rare classes. In practice, these methods have been shown to lead to a rapid over-fitting of the rare classes [6], [14] and improve the tail performances at the cost of performances on the frequent classes [7], [9]. Other methods also focus on augmenting the information via data augmentation [15], on experimenting with different network architectures i.e. by learning an ensemble [6] or by improving the learning process by decoupling the training [16]. Recently, multi-expert/multi-branch models have shown promising results [17], [18], [19]. These networks strategically associate multiple diverse experts/branches to obtain the final predictions. While a majority of the works tackle the imbalance problem for image classification, only a few studies address this problem for semantic segmentation [20], [21], [22]. The transfer of long-tail classification methods to semantic segmentation is not straightforward due to the inherent differences between the two tasks: Unlike classification, where each instance is treated individually, pixels occurring within the same frame are spatially dependent. Methods that work for classification might not capture the spatial correlations needed for segmentation. Similarly, due to class co-occurrence in each image, classes cannot be isolated easily for sampling or data augmentation purposes [6].

Inspired by the success of multi-expert models in image classification [14], [23], [18], we design a multiple complementary experts network (MCE) for semantic segmentation. MCE combines a shared backbone with a set of learnable modules trained on several overlapping subsets of classes. We adopt some training techniques developed from classification problems and observe their effectiveness for semantic segmentation tasks. We explore different training strategies and aggregation methods to identify the performance's upper bounds and some improvement directions. We evaluate our approach on two large-scale datasets; a large-scale alpine land cover dataset that we developed and made publicly available, and the FLAIR dataset, a heavily imbalanced benchmarking dataset for land cover mapping. Compared to other state-of-the-art methods, our MCE network manages a significant improvement in rare classes accuracy while minimally decreasing the performances in most frequent classes. Therefore biases of classification towards majority classes are reduced, closing the gap with national agencies' production requirements.

II. Related work

A. Semantic segmentation for land cover mapping

LC mapping can be designed for image classification, i.e. to provide a single label for an image, but also as a semantic segmentation task, where each pixel from the input image receives a label. Deep learning-based methods fit well with the characteristics of remote sensing images thanks to the large volume of data available. Rapid advances were made during the past decade driven primarily by the development of powerful architectures such as deep convolutional networks (CNNs). The fundamentals for DL semantic segmentation were established by Long et al. in 2014 with the fully convolutional networks (FCNs) [24], a CNN that can learn dense predictions by preserving the spatial structure of the input image in an efficient and end-to-end way. Ronneberger et al. designed U-Net [25], an encoder-decoder network for semantic segmentation that consists of a contracting path and an expansive path, which allows it to learn both local and global features of an image. Later, the DeepLab series was introduced by Chen et al. [26] where they proposed atrous convolution and atrous spatial pyramid pooling to enlarge the receptive field of the network and capture multi-scale contextual information. These remain de facto standard approaches for LC mapping that are being constantly improved upon, for example considering rotation invariance [3], interpretability [27] or including more modalities [28].

B. Class imbalance in classification problems

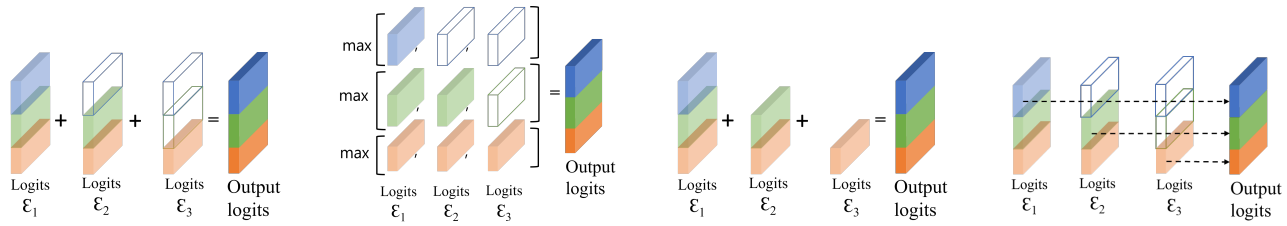
The problem of imbalanced distribution is well-studied and diverse methods have been developed in recent years. Here we review briefly two types of methods, class re-balancing and ensemble learning, we refer to [6] for a more complete review.

Class re-balancing via re-sampling and re-weighting. Re-balancing methods aim at fighting the imbalance by introducing prior information about the class distribution and giving more importance to unfrequent samples. An intuitive solution is data re-sampling which tries to achieve a balanced distribution across classes through over-sampling tail classes or under-sampling head classes [29], [7]. Practically, over-sampling methods lead to rapid overfitting of the rare classes, whereas under-sampling discards part of the data that may contain important information.

Instead of acting at the sample level, the re-weighting methods operate at the loss level and introduce a balancing factor to adjust the loss value for different classes. The vanilla solution is the weighted softmax that uses the inverse of the class frequency as a factor. Some approaches also use the prediction difficulty [11] or the effective number of samples [12] to re-balance the loss, while others directly modify the gradients [30] or the logits based on the training labels frequencies [31]. These methods based on loss function weighting improve rare classes' performance but can often be detrimental to the recognition of frequent classes.

Ensemble learning. Multi-expert or multi-branch models combine multiple network modules in parallel that aim to extract different representations from the data. Such ensemble learning method is seen as state-of-the-art for imbalanced visual recognition [23], [6] and their benefits are typically attributed to the expert's diversity: making diverse mistakes [32] or exploring different local minima in the loss landscape [33].

The diversity among various experts can be introduced by learning on different groups of classes or different class distributions. Various grouping of classes have been experimented with: BNN [19] uses a two-branches architecture, one



(a) Average logits (b) Maximum logits (c) Zero non-target logits (d) Oracle model

Fig. 2: Illustration of the aggregation methods for a MCE model with 3 experts. The blue, green and orange colours correspond to the logits of head, body and tail classes respectively.

focusing on original distribution, the other on a re-balanced version of it. BAGS [34] or LFME [17] groups classes with similar numbers of samples. ACE [14] and ResLT [35] form overlapping groups of classes so that experts have specific and complementary recognition skills.

C. Multi-expert model for imbalanced semantic segmentation

Semantic segmentation can be formulated as a per-point/pixel classification, thus several of the methods presented above have been extended to segmentation tasks. Inspired by re-weighting methods, Zhou et al. [36] have proposed a dynamic sample weighting algorithm. Wang et al. [30] introduced the Seesaw loss that re-balances the expert's gradients based on the label frequency and the number of misclassified samples for each class. Recently, the imbalance problem has been observed from the representation learning perspective and focused on learning more balanced features in the latent space [20], [37]. Closer to our work, [38] uses a region-re-balance branch to better learn rare class features during training. Other multi-expert models for semantic segmentation have been developed recently [22], [18] but without specifically addressing the class imbalance problem.

In this work, we take inspiration from the success of multi-expert models developed for classification problems and we adapt them to semantic segmentation tasks. We adopt an approach with an overlapping grouping of classes, that allows the ensemble to learn specific, but complementary skills.

III. Method

A. Model overview

The overall architecture of our multiple complementary experts (MCE) network for alpine land cover mapping is shown in Figure 1. The architecture comprises a shared visual backbone, followed by several separated trainable layers forming the experts, and finally, an aggregation module that combines the experts' predictions and produces the final model output.

B. Experts design

The i expert modules $E = \{\mathcal{E}_1, ..., \mathcal{E}_i\}$ are branched out from the same visual backbone. They all share the same architecture but have different parameters to reflect the specificity of their respective inputs. Each expert is composed of one convolutional layer with kernel 3x3 followed by a ReLU [39] activation, a batch normalization layer [40] and a final convolutional layer with a kernel size of 1x1. Li et al. [41] observed that learning a more uniform distribution with fewer samples is sometimes easier than learning a long-tailed distribution with more samples. Moreover, following the spirit of re-balancing, tail classes should be more exposed to the model. Therefore, we divide the $\mathcal{C}$ classes into more balanced but overlapping splits, similarly to ACE [14]. We assign the first expert $\mathcal{E}_1$ to all the classes, and the consecutive experts with progressively rarer and rarer classes (i.e. $\mathcal{E}_2$ focuses on middle and rare classes, $\mathcal{E}_3$ only sees rare classes). Practically, we feed each expert with all the pixels while attributing a loss weight of 0 for samples whose class does not belong to the target classes of the experts. Consequently, the expert $\mathcal{E}_i$ only focuses on its target classes $\mathcal{C}_i$, and the losses from pixel belonging to non-target classes $\tilde{\mathcal{C}}_i$ are not back-propagated.

Following the Linear Scaling Rule[42] for multi-expert models [14],*'when the batch size is increased by a factor k, the learning rate should be multiplied by a factor k'*, we adapt each expert learning rate and assign smaller values to those trained with less data to avoid overfitting. The base learning rate η_0 is used to train the backbone and the most general expert $\mathcal{E}_1$. The i-th expert $\mathcal{E}_i$ is trained with the **adapted learning rate** η_i :

$$\eta_i = \eta_0 \frac{\sum_{c \in C_i} n_c}{\sum_{j \in C} n_j} \tag{1}$$

where n_c is the number of pixel in class c belonging to the target classes $\mathcal{C}_i$ seen by expert $\mathcal{E}_i$ and $\mathcal{C}_i \subset \mathcal{C}$.

C. Loss function

The model is trained with a combination of classification losses and complementary losses. We compute a classification loss on each expert output separately, to learn expert-specific features. Given the label y_c and the logits $z_i \in \mathbb{R}^{1 \times C}$ from expert $\mathcal{E}_i$, the classification loss for expert $\mathcal{E}_i$ is a cross-entropy loss over its target classes C_i :

$$\mathcal{L}_{cls,i} = -\sum_{c \in C_i} y_c \log(\sigma(z_i)) \tag{2}$$

with $\sigma(.)$ representing the SoftMax operation.

Similarly to [14], we use a complementary loss that penalizes the experts for predicting any of the non-target classes $\tilde{C}_i$. Since the classification loss is not computed on the pixels belonging to these classes, no gradient updates the parameters,

therefore their output should be close to zero. It is defined as a L_2-penalty term :

$$\mathcal{L}_{com,i} = -\sum_{c \in \tilde{C}_i} \|z_{i,c}\|^2 \tag{3}$$

where $z_{i,c} \in \mathbb{R}$ is the logit of $\mathcal{E}_i$ for the class c belonging to the non-target classes $\tilde{C}_i$. Thus the network loss for K experts for a given pixel can be written as :

$$\mathcal{L} = \sum_{i=1}^{K} \mathcal{L}_{cls,i} + \sum_{i=1}^{K} \mathcal{L}_{com,i} \tag{4}$$

It is interesting to notice that there is no loss on the final model output; we avoid updating together the experts' weights to ensure diversity between their predictions.

D. Aggregation

The aggregation module combines the logits from all the experts into the final model output via algebraic operations as shown in Figure 2.

- The output logits o_c of class c can be computed as the **mean of the logits** from all experts:

$$o_{c,mean} = \frac{1}{K} \sum_{i=1}^{K} z_{i,c} \tag{5}$$

  where $z_{i,c}$ are the logits for class c from expert i. The logits coming from **non-target classes** (i.e. logits from a head class in the rare class expert) can be ignored by setting them to zero at the expert level, i.e. $z_{i,c} = 0$ if $c \in \tilde{C}_i$.
- As an alternative, we consider the **group maximum logits** for each class, where:

$$o_{c,max} = \max_{1 \le i \le K} \{z_{i,c}\} \tag{6}$$

- We conduct an **oracle** case study to establish an upper bound on the performance of the MCE model. As each expert concentrates more than the others on some subset of the classes, i.e. the tail expert predicts best the tail classes, the model would achieve optimal performance when each pixel gets predictions only from the expert who specialises in its category. For this assumption, we introduce prior knowledge about the best expert for every pixel in the inference phase and obtain the oracle results. Note that this approach is used only for benchmarking, as it requires the ground truth to be applied at inference.

IV. Experiments

In this section, we present the data, the experimental details and the evaluation metrics.

A. TLM Dataset

1) The TLM raw image data: consists of very high-resolution aerial images and a digital elevation model (DEM) that covers approx. $2,300\text{km}^2$ of land above $2,000$m altitude in the southwestern part of Switzerland as shown in Figure 4. The raw data are made openly available by the Federal Office

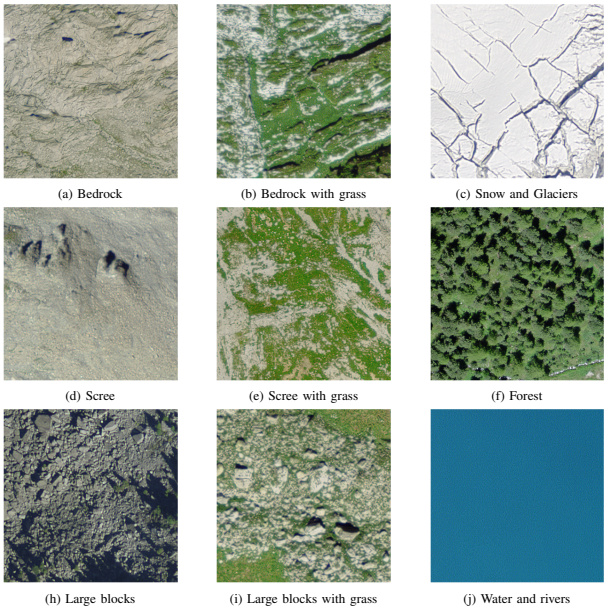


Fig. 3: Examples of alpine land cover classes from our dataset, where the entire surface is occupied by the specified target land cover category.

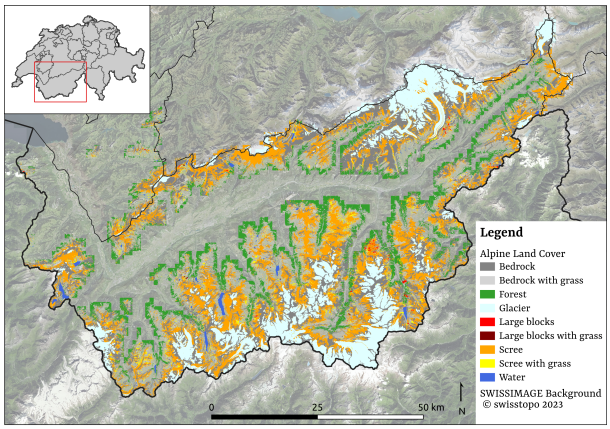


Fig. 4: Map of the study area located in southwest Switzerland with altitude above 2000m.

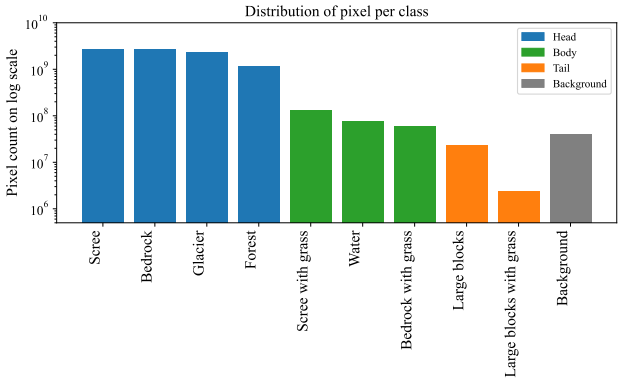


Fig. 5: Pixel distribution among the alpine land cover classes in the TLM dataset on a logarithmic scale

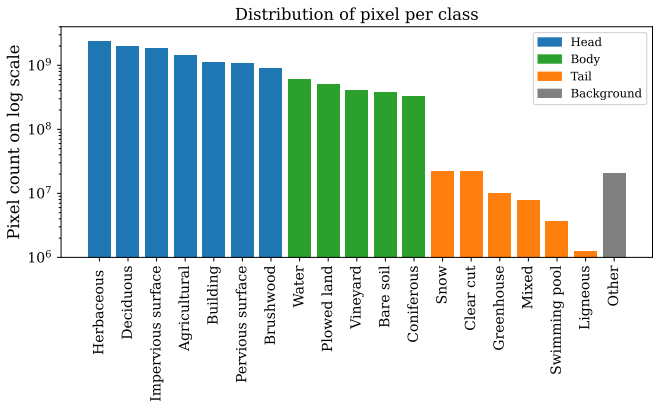


Fig. 6: Pixel distribution in the FLAIR dataset on a logarithmic scale

of Topography swisstopo[1] under the Open Government Data policy. The aerial images with RGB channels have been taken from the swissIMAGE product from the year 2020 and have RGB bands and a spatial resolution of 25cm. The DEM is derived from the swissALTI3D product from the year 2019 with a ground resolution of 0.5m and accuracy of 1m to 3m. RGB data have been upsampled to 50cm to have a consistent resolution with the DEM.

2) The TLM land cover labels: used in this study focuses on alpine land cover (hence the choice of limiting the dataset to areas higher than $2,000$m). The labels are taken from the Swiss Topographic Landscape Model model (swissTLM3D) layer generated through visual photo-interpretation by experts from swisstopo based on the aerial images from 2014-2017. The nine land cover types in our study area include *bedrock, bedrock with grass, large blocks, large blocks with grass, scree, scree with grass, water area, forest and glacier* and are shown in Figure 3. Multiple classes within rocky areas present high visual similarity, adding to the challenge of imbalance. Areas without labels are regarded as a background class and images with background (unlabeled) pixels over 10% of the total number of pixels are removed. The final dataset contains $229,538$ tiles with a size of 200×200 pixels (1ha) and is available for download [2].

3) Dataset construction and splitting: Figure 5 shows the distribution of pixels among classes and presents a typical case of a long-tailed distribution with an imbalance factor, defined as the ratio of the most frequent to the rarest class, close to 1000. The classes are grouped by frequency into 'head', 'body', and 'tail' groups. We divide the images into training, validation and test sets with a ratio of $[0.6 : 0.2 : 0.2]$. To achieve a balanced distribution in each split, we perform stratified sampling at the tile level based on the most frequent label in each tile.

B. FLAIR dataset

The FLAIR [43] dataset was developed by the French National Institute of Geographic and Forest Information (IGN) and is a comprehensive benchmarking dataset that combines aerial imagery with land cover annotations. This dataset contains an extensive collection of more than $77,412$ high-resolution patches, each measuring 512×512 pixels, with a spatial resolution of 0.2m and with 19 semantic classes. For our work, we use the RGB bands, the elevation channel and the land cover labels. We discard other spectral bands. The pixel distribution across different land cover classes exhibits a significant long-tailed distribution with an imbalance factor of approximately $2,000$, as shown in Figure 6.

C. Experimental setting

We selected two semantic segmentation visual backbones for our experiments: the DeepLabv3+ [26] architecture and the U-Net [25]. Both backbones use a ResNet-50 [44] image encoder that is initialized using the pre-trained weights from ImageNet-1K [45] for the RGB channels and the DEM channel weights are copied from the first (red) channel. Other convolution layers are initialized with the He initialization method [46] and normalization layers, with zero mean and unit norm. We use the Adam [47] optimizer with a weight decay value of 0.01, a base learning rate of $1e^{-4}$, and a batch size of 64 for the DeepLabv3+ backbone, respectively $1e^{-5}$ and 32 for the U-Net backbone. The hyper-parameters were chosen through a grid search on the validation set. We train all models for 50 epochs, with a learning rate decay factor of 0.1 if no improvement occurs over the last 10 epochs.

We use basic data augmentation (random horizontal and vertical flips, rotations, colour jittering and normalization) for training all models, but only normalization for testing and validation. All the experiments are implemented with Pytorch and run with one GPU *NVIDIA GeForce RTX 3090*, delivering 120 samples/s for inference, and 68 samples/s in training.

D. Evaluation metrics

The performance of different classes is usually considered to be equally important in long-tail recognition. We thus report results with overall accuracy (OA), mean Intersection-over-Union (mIoU) as well as macro-average accuracy (mAcc) on head, body and tail groups and accuracy per class. The results are computed over a separate test set with a distribution similar to the training and validation set.

V. Results

A. Comparison of different methods

Table I presents the results for our MCE network with 2 (MCE-2) or 3 experts (MCE-3) with both DeepLabV3+ and U-Net backbones on the TLM dataset. We compare them with similar visual backbones trained with a cross-entropy loss (CEL), with inverse frequency weights (WCEL), with class balanced weights (CBL) [12] with $\beta = 0.9999$, and with the seesaw loss (SL) [30]. We also add the results for the performance of the Oracle model based on the MCE-3 model. We observe an improvement in performance for mIoU, mAcc and tail classes accuracy for both U-Net and DeepLabv3+ backbones with our MCE-2 and MCE-3 models,

[1]https://www.swisstopo.admin.ch/en/geodata/images
[2]https://dx.doi.org/10.21227/n61c-k282

TABLE I: Comparison of our approach with two different backbones, DeepLabv3+ and U-Net. We present the results for the MCE model with 2 (MCE-2) or 3 experts (MCE-3) with mean aggregation, with a cross-entropy loss model (CEL), a weighted cross entropy loss (WCEL), a class balanced loss (CBL), a seesaw loss (SL). Results on head, body and tail classes are average accuracy per group. The Oracle model is based on the MCE-3 model. The best results are in **bold**, second best are underlined.

Backbone	Deeplabv3+						U-Net					
Methods	**mIoU**	**mAcc**	**OA**	**head**	**body**	**tail**	**mIoU**	**mAcc**	**OA**	**head**	**body**	**tail**
CEL	53.2	59.6	**89.2**	**91.7**	60.8	23.6	45.2	51.1	**86.8**	**89.6**	50.9	0.0
WCEL	27.1	41.5	68.6	73.7	40.0	0.0	22.4	35.7	63.0	68.9	27.2	0.0
CBL	40.4	52.6	77.0	80.4	68.0	0.0	33.2	44.1	79.6	83.2	36.2	0.0
SL	50.5	64.6	86.9	89.4	71.3	37.6	48.0	65.0	86.3	88.6	**74.4**	36.4
MCE-2	**53.9**	68.8	87.6	89.7	**73.7**	54.2	**52.0**	66.0	86.1	88.4	73.0	44.0
MCE-3	53.6	**70.0**	87.2	89.3	72.8	**62.0**	49.0	**69.1**	85.0	87.5	71.2	**63.4**
Oracle	69.5	82.3	89.3	90.6	92.7	91.1	69.2	81.2	88.0	89.6	91.7	89.6

TABLE II: Comparison of our MCE network with the DeepLabv3+ backbone on the FLAIR dataset with 2 (MCE-2) or 3 experts (MCE-3). We use mean aggregation, with a cross-entropy loss model (CEL), a weighted cross entropy loss (WCEL), a class balanced loss (CBL), and a seesaw loss (SL). Results on head, body and tail classes are average accuracy per group. The Oracle model is based on the MCE-3 model. The best results are in **bold**, second best are underlined.

	FLAIR					
Methods	**mIoU**	**mAcc**	**OA**	**head**	**body**	**tail**
CEL	37.9	52.4	73.1	**73.7**	64.5	17.5
WCEL	33.2	**63.0**	66.0	65.8	**74.1**	**50.4**
CBL	36.7	59.5	71.9	70.7	66.6	40.4
SL	37.4	54.0	71.3	72.6	66.6	21.9
MCE-2	**39.0**	55.8	72.8	70.5	73.0	24.3
MCE-3	38.8	61.6	**73.3**	70.8	73.3	41.2
Oracle	49.1	71.9	77.2	72.3	91.4	55.0

compared to all other approaches. A small decrease in the performance on head classes and overall accuracy is observed compared to the CEL. However, this trade-off between head-tail accuracy is shared among all re-balancing methods, with the MCE-2 and MCE-3 models offering the smallest or second smallest deterioration in performances for the frequent classes. The seesaw loss performance also offers a generally good compromise between all classes, however, the MCE models introduce a better performance for tail classes, mIoU and mAcc. With the increase in the number of experts from MCE-2 to MCE-3, the performance remains consistent between the head and body classes, but the significant enhancement in the performance of the tail classes underlines the effectiveness of the third expert in the network.

Similar results are observed for models trained on the FLAIR dataset (Table II). The MCE approaches obtain the best metrics for mIoU and OA, and the second-best results for the mAcc, body and tail accuracy. The results are surprisingly good with the WCEL loss, which obtained the best body and tail accuracy. Compared to other re-balancing methods, MCE approaches consistently exhibit a superior trade-off, effectively enhancing the accuracy of both body and tail classes while minimizing the compromise in accuracy for head classes. It is important to note that the results reported in FLAIR [43] focus on the 13 most common classes and discard the 6 'tail' classes, thus the numerical value of the metrics should not be compared.

B. Oracle model and accuracy per class

As expected, the Oracle outperforms all other models in terms of performance, since we introduce prior knowledge about the specific expert to look at for each pixel, and thus it forms a performance upper bound. The results with MCE-2 and MCE-3 models are very close to the oracle upper bounds for the head classes, however, the body and tail classes exhibit much lower values. This indicates the head expert predictions are well incorporated into the network output, however, the predictions of body or tail experts could be better taken into account.

Table III provides more details on individual expert predictions for the MCE-3 model, where we look at each expert's probabilities before the aggregation module. The high level of accuracy of each expert on their set of target classes demonstrates that our training strategy is effective in specializing each expert on a given subset of classes. Expert 1 closely aligns with the results observed in Table I for the CEL network for all head, body and tail classes, where the CEL is commonly seen as an upper boundary regarding overall accuracy and head classes. Each expert focuses on the most frequent classes among its target classes, and the accuracy is reduced for underrepresented samples for each expert(i.e. body and tail classes for the head expert), illustrating that models better learn from smaller but balanced sets of classes. The complementary loss pushes the expert's predictions for non-target classes toward zero, however, the logits of the latter are not exactly zeros, leading to some infrequent but correct predictions of non-target classes (expert 2 on head, expert 3 on head and body classes). In light of these results, enhancing the network capabilities seems to depend upon a more effective aggregation of expert logits, which is further analysed in Section V-D.

C. Study of the training strategy

The ablation study presented in Table IV compares the effectiveness of the complementary loss (L_{com}) and the adapted learning rate for each expert (adapt-lr) on a MCE model with 3 experts. While very effective on body and tail classes, a model without adapted learning rate and complementary loss significantly lowers its recognition ability on head classes,

TABLE III: Performance in terms of accuracy per class from the MCE-3 model with the prediction from each expert, the MCE-3 output and the oracle model. The oracle relies on the ideal aggregation of individual expert predictions to generate the network output, achieving this with the prior knowledge of which expert to look at for each pixel.

	Head	Scree	Bedrock	Glacier	Forest	**Body**	Scree with grass	Water	Bedrock with grass	**Tail**	Large blocks	Large blocks with grass
Expert 1	**90.6**	**86.8**	**86.5**	**92.7**	**96.5**	59.2	42.6	94.0	40.9	23.0	46.1	0.0
Expert 2	5.4	5.9	5.6	4.1	5.9	**92.7**	**93.3**	**99.6**	**85.3**	68.2	90.3	46.1
Expert 3	9.2	8.8	11.9	8.0	7.9	5.0	3.0	0.4	11.5	**91.1**	**97.5**	**84.7**
MCE-3	89.3	83.8	85.2	92.6	95.6	72.8	65.4	96.7	56.4	62.0	75.2	48.9
Oracle	90.6	86.8	86.5	92.7	96.5	92.7	93.3	99.6	85.3	91.1	97.5	84.7

TABLE IV: Effects of the complementary loss (L_{com}) and the adapted learning rate (adapt-lr) on the MCE-3 network.

adapt-lr	**Lcom**	**mIoU**	**mAcc**	**head**	**body**	**tail**
-	-	37.9	**72.1**	72.9	**87.1**	**84.1**
✓	-	40.5	69.7	79.6	83.1	64.8
✓	✓	**53.6**	70.0	**89.3**	72.8	62.0

TABLE V: Comparison of aggregation methods for MCE-3.

Aggregation	**mIoU**	**mAcc**	**OA**	**head**	**body**	**tail**
MCE-3 (Mean)	53.6	70.0	87.2	89.3	72.8	62.0
Max-pool	51.5	**72.1**	86.4	88.4	**77.4**	**67.6**
zero-non target z_i	**54.6**	68.3	**87.6**	**89.8**	69.9	57.3
MLP	51.6	59.5	86.0	88.9	56.8	34.5
CNN	50.0	67.8	85.5	88.1	63.9	67.2

leading to poor overall performance. The adapted learning rate mechanism upholds an elevated level of accuracy in classifying head categories while preserving a satisfactory degree of recognition for less prevalent classes. The incorporation of the complementary loss further amplifies this effect.

D. Comparison of aggregation methods

We study different alternatives for aggregating the experts' output, as presented in Table V. We run inference passes based on the MCE-3 model weights and only modify the aggregation methods: instead of mean of logits (mean), we use maximum-pooling of logits (max-pool), or we set to zero the logits coming from non-target classes for each expert (zero non-target logits). We also experimented with the aggregation of the experts' logits with learnable layers. We train a one-layer CNN, and respectively a small 2-hidden layer multi-layer perceptron (MLP), to aggregate the experts' logits based on the output of the frozen best-trained MCE-3 model. We trained these smaller networks for 50 epochs each with a learning rate of $5e^{-4}$ for the CNN, resp. $1e^{-5}$ for the MLP, decaying by a factor 0.1 every 10 epochs.

The results indicate that max-pooling pushes again the head-tail ratio toward the rarest classes, by compromising slightly the performances on head classes. The zero-ing of the non-target classes logits for each expert favours slightly the head classes, but the performance for the body and tail classes drops, indicating that the network benefits from the logits coming from all experts, even if the expert is not specialised in the subset of classes, showing that the experts have a truly complementary action. The two learnable aggregation methods based on the frozen MCE-3 model with mean aggregation obtain lower performance in all metrics, illustrating the difficulty of the aggregation task. In summary, the mean logit aggregation method appears to be effective for our network, as the network autonomously learned to balance and aggregate the sometimes conflicting experts' predictions.

This analysis of the aggregation strategy is complemented by Figure 7, which illustrates how the MCE-3 network handles diverse expert predictions through mean aggregation. Each expert produces diverse LC predictions, focusing on their target LC classes. When the experts produce conflicting predictions, the aggregation via the mean of the logits allows to determine whether to pick the predictions from one of the experts (rows a,d,e) without disruption from other experts or to locally select one expert's predictions (rows b or c). When the classes from the labels are not among one expert target classes (which is often the case for the Expert 3), the expert predictions are misguiding and meaningless, but also do not impact the aggregated output (row a,b,e).

VI. LIMITATIONS

One notable limitation of the MCE models is that while they show improved performance on rare classes, overall accuracy might become slightly compromised, most probably because of the lower recall on the majority classes; this is observed on the TLM dataset with both backbones, but not on the FLAIR dataset. For image classification tasks, many re-balancing methods improve the performance on the rare classes by sacrificing the overall accuracy, typically re-weighting or data augmentation methods [12], [6]. However, the recent multi-expert methods seem to better integrate these constraints and several multi-expert works are advertised as improving both minority and majority classes accuracy on benchmarking datasets, such as ACE [14], RIDE [48] or LFME [17], yet in a real-world scenario, it might not always be the case as we observe it here on the TLM dataset. The drop in overall accuracy in that dataset is less than 2%, for a gain of 60% in accuracy on the minority classes. The choice between emphasizing rare classes or optimizing overall accuracy depends on the specific task objectives and should align with specific application requirements. In the real-world scenario of national mapping, the updating of land cover maps

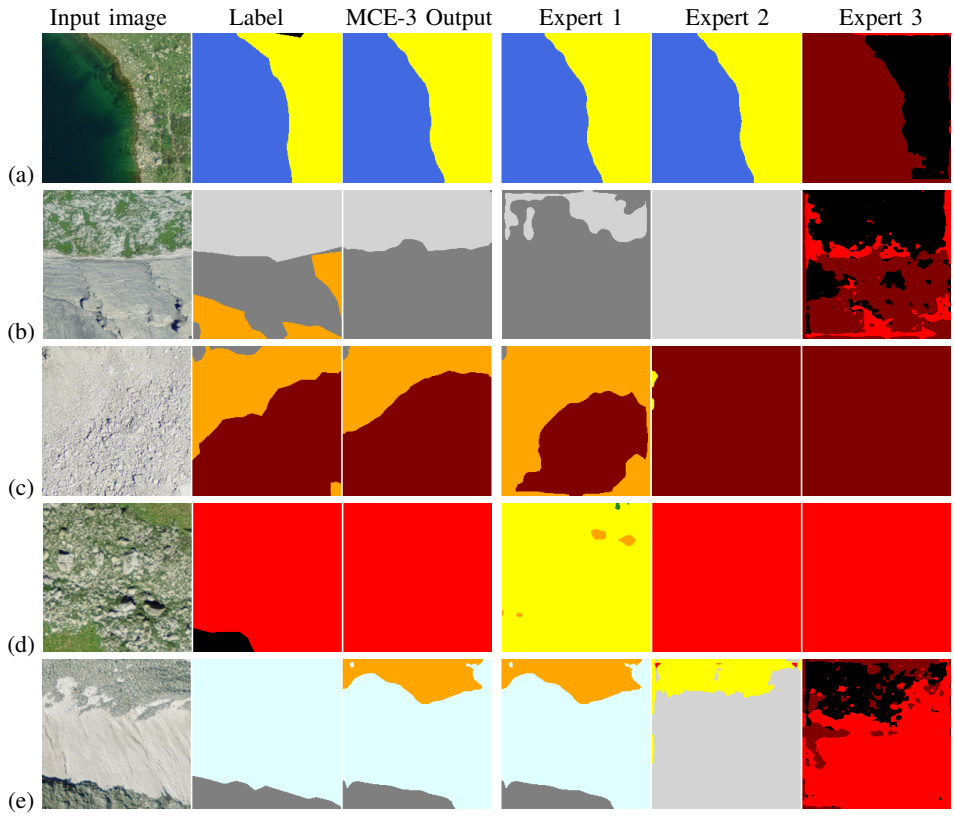


Fig. 7: Example of input images, labels, final network output and individual expert predictions for the MCE-3 model. Labels for semantic segmentation maps: ■ background, ■ water, ■ glacier ■ bedrock, ■ bedrock with grass, ■ large blocks, ■ large blocks with grass, ■ scree, ■ scree with grass.

still commonly resorts to manual verification of the predictions due to stringent accuracy requirements. In such cases, the priority shifts towards recall of these rare classes that were maybe forgotten, even if it results in a compromise and slightly reduces the performances of more prevalent classes, which already meet very high accuracy standards. Furthermore, these investigations are based on the assumption that the reference data is correct. In reality, however, the interpreters who have collected this reference data are also subject to a certain error rate, leading to reference data with a certain amount of noise. Greater uncertainties occur with complex land cover classes, such as mixed classes ('scree with grass'), as opposed to more straightforward classes such as water. This a priori error cannot be determined in practice. The results obtained must be viewed with this restriction.

Through our experiments, we observed that each expert obtained high accuracy on their designated set of classes, but the aggregation of their predictions seemed to be the sensitive part of the network. Even though several approaches have been studied to better aggregate the individual expert's predictions (e.g. using an MLP), the learning of the rare categories in the aggregation layer remains sensitive to the abundant negative gradients from frequent classes and does not beat the simple average of each expert logit.

VII. Conclusion

This work has presented a multiple complementary expert model that effectively addresses the class imbalance problem in semantic segmentation. These problems are common in several real-world applications involving remote sensing data, from land cover mapping to ecosystem classification or species distribution models. In all those cases, long-tail distributed classes require specialised approaches to detect and model rare classes correctly. Extensive experiments conducted on two land cover datasets have led to the development of an efficient model and training strategy. Our approach involves training several experts in a complementary manner, each specializing in a balanced subset of classes. Through an ablation study, we have demonstrated the effectiveness of adaptive learning rates and the complementary loss function, enabling an advantageous head-tail class trade-off. Overall, our MCE approach surpasses the performance of commonly used methods for handling class imbalance in terms of mean Intersection-over-Union, mean accuracy and rare classes accuracy, showcasing the ability of our experts to learn distinctive features tailored to specific class subsets. These findings underscore the potential of our approach in advancing semantic segmentation for real-world applications by mitigating class imbalance effects.

References

[1] D. Tuia, D. Marcos, K. Schindler, and B. Le Saux, *Deep Learning-based Semantic Segmentation in Remote Sensing*, John Wiley & Sons,

Ltd, 2021.

[2] D. Tuia, K. Schindler, B. Demir, et al., "Artificial intelligence to advance earth observation: a perspective," *arXiv preprint arXiv:2305.08413*, 2023.

[3] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 96–107, 2018.

[4] J. Feranec, T. Soukup, G. Hazeu, and G. Jaffrain, *European Landscape Dynamics: CORINE Land Cover Data*, CRC Press, 2016.

[5] B. Kellenberger, M. Volpi, and D. Tuia, "Fast animal detection in UAV images using convolutional neural networks," in *IEEE international geoscience and remote sensing symposium (IGARSS)*, 2017, pp. 866–869.

[6] Y. Zhang, B. Kang, B. Hooi, et al., "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10795–10816, 2023.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] L. Yang, H. Jiang, Q. Song, and J. Guo, "A survey on long-tailed visual recognition," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1837–1872, 2022.

[9] G. Van Horn and P. Perona, "The Devil is in the Tails: Fine-grained Classification in the Wild," *arXiv preprint arXiv:1709.01450*, Sept. 2017, arXiv: 1709.01450.

[10] P. Zimmermann, E. Tasser, G. Leitinger, and U. Tappeiner, "Effects of land-use and land-cover pattern on landscape-scale biodiversity in the european alps," *Agriculture, Ecosystems & Environment*, vol. 139, no. 1, pp. 13–22, 2010.

[11] T.-Y. Lin, P. Goyal, R. Girshick, et al., "Focal loss for dense object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2017, pp. 2980–2988.

[12] Y. Cui, M. Jia, T.-Y. Lin, et al., "Class-balanced loss based on effective number of samples," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 9268–9277.

[13] J. Tan, C. Wang, B. Li, et al., "Equalization Loss for Long-Tailed Object Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 11662–11671.

[14] J. Cai, Y. Wang, and J.-N. Hwang, "ACE: Ally Complementary Experts for Solving Long-Tailed Recognition in One-Shot," in *IEEE International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 112–121, IEEE.

[15] T. Li, P. Cao, Y. Yuan, et al., "Targeted supervised contrastive learning for long-tailed recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6918–6928.

[16] B. Kang, S. Xie, M. Rohrbach, et al., "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019.

[17] L. Xiang, G. Ding, and J. Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, Aug. 2020, pp. 247–263.

[18] S. Pavlitskaya, C. Hubschneider, M. Weber, et al., "Using mixture of expert models to gain insights into semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 342–343.

[19] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 9719–9728.

[20] Z. Zhong, J. Cui, Y. Yang, et al., "Understanding imbalanced semantic segmentation through neural collapse," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19550–19560.

[21] S. Lu, F. Gao, C. Piao, and Y. Ma, "Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data," in *International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, Oct. 2019, pp. 230–233.

[22] S. Ma, Z. Zhao, Z. Hou, and X. Yang, "Image semantic segmentation algorithm based on a multi-expert system," *Journal of Electronic Imaging*, vol. 32, no. 03, June 2023.

[23] E. S. Aimar, A. Jonnarth, M. Felsberg, and M. Kuhlmann, "Balanced Product of Calibrated Experts for Long-Tailed Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sept. 2018, pp. 801–818.

[27] T.-A. Nguyen, B. Kellenberger, and D. Tuia, "Mapping forest in the swiss alps treeline ecotone with explainable deep learning," *Remote Sensing of Environment*, vol. 281, pp. 113217, 2022.

[28] S. Srivastava, J. E. Vargas-Munoz, and D. Tuia, "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution," *Remote sensing of environment*, vol. 228, pp. 129–143, 2019.

[29] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[30] J. Wang, W. Zhang, Y. Zang, et al., "Seesaw loss for long-tailed instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 9695–9704.

[31] A. K. Menon, S. Jayasumana, A. S. Rawat, et al., "Long-tail learning via logit adjustment," *arXiv preprint arXiv:2007.07314*, 2020.

[32] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[33] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," *arXiv preprint arXiv:1912.02757*, 2019.

[34] Y. Li, T. Wang, B. Kang, et al., "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[35] J. Cui, S. Liu, Z. Tian, et al., "Reslt: Residual learning for long-tailed recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3695–3706, 2022.

[36] Z. Zhou, C. Zheng, X. Liu, et al., "A dynamic effective class balanced approach for remote sensing imagery semantic segmentation of imbalanced data," *Remote Sensing*, vol. 15, no. 7, pp. 1768, 2023, Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[37] Y. Wang, J. Fei, H. Wang, et al., "Balancing logit variation for long-tailed semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19561–19573.

[38] J. Cui, Y. Yuan, Z. Zhong, et al., "Region rebalance for long-tailed semantic segmentation," *arXiv preprint arXiv:2204.01969*, June 2022.

[39] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning (ICML)*, July 2010, pp. 807–814.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, July 2015.

[41] Y. Li, T. Wang, B. Kang, et al., "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 10991–11000.

[42] P. Goyal, P. Dollár, R. Girshick, et al., "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[43] A. Garioud, S. Peillet, E. Bookjans, et al., "Flair 1: semantic segmentation and domain adaptation dataset," 2022.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[45] J. Deng, W. Dong, R. Socher, et al., "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 248–255.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1026–1034.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2017.

[48] X. Wang, L. Lian, Z. Miao, et al., "Long-tailed recognition by routing diverse distribution-aware experts," *arXiv preprint arXiv:2010.01809*, 2020.

Valerie Zermatten (Student Member, IEEE) received the MSc degree from the Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, where she is currently pursuing the PhD degree, in collaboration with the Federal Office of Topography swisstopo. Her research interest include deep learning for processing and understanding remote sensing images with multi-modal data.

Xiaolong Lu received the BSc degree from Sun Yat-sen University, Guangzhou, China and the MSc degree from the Swiss Federal Institute of Technology (ETHZ), Zurich, Switzerland. He is currently a research assistant with the Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany. His research interests include multi-modal learning and long-tailed learning.

Javiera Castillo-Navarro received the Ph.D. degree from Université Bretagne Sud, Vannes, France, in collaboration with the Office National d'Études et Recherches Aérospatiales, Université Paris-Saclay. She is currently a post-doctoral researcher with Ecole Polytechnique Federale de Lausanne (EPFL), Sion, Switzerland. Her research focuses on deep learning for scene understanding and Earth observation applications.

Tobias Kellenberger is head of Innovation within the department of Topography at the Swiss Federal Office of Topography (swisstopo). He holds a master and a Ph.D. degree in Geography / Remote Sensing from the University of Zürich (Switzerland). From 1996 to 2009 he worked as a scientific researcher, lecturer and as head of the research group LA-COMMLab (Land Cover Mapping and Monitoring) at the Remote Sensing Laboratories (RSL), Department of Geography, University of Zürich. From 1996 to 2003 he acted as secretary and until 2010 as president of the Swiss Remote Sensing Commission (SRSC). 2004– 2009 he was head of the scientific National Point of Contact (NPOC) for satellite images at the University of Zürich and until 2019 of the NPOC at swisstopo. In 2009 he joined swisstopo. He is involved in the 3D Special Interest Group of EuroSDR (European Spatial Data Research) and is heading the Geoscience coordination Group (GKG AG) at Federal and cantonal level. His research focus is on multispectral and multi-sensor data extraction; radiometric sensor calibration; VGI, 3D standards, modelling and visualization of 3D Geodata within mapping agencies; VGI; routing; Geodata Science (AI).

Devis Tuia (Fellow Member, IEEE) received the PhD degree from the University of Lausanne, Lausanne, Switzerland, in 2009. He was a postdoc with the University of Valencia, the University of Colorado, Boulder, CO, and EPFL Lausanne. From 2014-2017, he was an assistant professor with the Department of Geography, University of Zurich. He was then professor of Geo-information science with Wageningen University, The Netherlands. Since 2020, he is an associate professor with Ecole Polytechnique Federale de Lausanne (EPFL), in Sion, Switzerland. He is interested in algorithms for information extraction and data fusion of remote sensing images using machine learning. For more information please visit: https://www.epfl.ch/labs/eceo/