SEMANTIC SEGMENTATION OF HIGH-RESOLUTION UAV IMAGES REPRESENTING POST-FLOOD SCENES

Pietro Brugnolo, s230254

Technical University of Denmark
Repository: https://github.com/PietroBrugnolo

ABSTRACT

Decision support systems based on visual scene understanding represent an area of great interest in post-disaster damage assessments. Most of the available datasets are composed of satellite images characterized by low spatial resolutions and high revisit periods resulting inadequate for systems to be used for effective and rapid intervention. However, this is not the case with FloodNet, a high-resolution image dataset acquired after Hurricane Harvey in 2017 with DJI Mavic Pro quadcopters. The authors developed it to address the issue of a decision support system to be used in unmanned aerial vehicle (UAV) applications. FloodNet represents urbanised and rural flood-affected areas and poses challenges such as the recognition of flooded streets and buildings. This work investigates how different deep neural architectures perform in a semantic segmentation task, using both labeled and unlabeled images from FloodNet.

1. INTRODUCTION

Floods represent a very large part of all natural disasters, according to [1] floods have been the 41% of all the weather-related disasters in Europe between 2001 and 2020. This leads to the need of developing increasingly advanced decision support systems to be used during these emergencies. Unmanned aerial vehicles (UAV) are particularly useful in these contexts because of their ability to acquire real-time information quickly, ensuring a fast response time. It is therefore critical to have suitable visual scene understanding systems. This kind of systems have been extensively investigated on public datasets of ground-based images such as Cityscapes [2], ImageNet [3] and Microsoft COCO [4]. Although some datasets for post disaster damage assessments exist [5] [6] [7], they mainly contain low resolution satellite images or highly noisy social media images. Only with the development of FloodNet [8], composed of high definition UAV images, recent works [9] [10] [11] [12] [13] have begun to focus on UAV-based visual understanding systems for post-flood scenes.

The two main challenges that FloodNet poses are the high class imbalance and the small number of labeled images. The dataset is mostly made up of unlabeled images which may be useful to increase the system generalization ability when used in a self-supervised context. With this work we investigate the semantic segmentation performance of well-known deep neural architectures. We train with a fully-supervised approach a U-Net, a DeepLabV3 and a PSPNet and moreover we investigate a semi self-supervised approach on DeepLabV3 and PSPNet. In addition, in 5 out of 6 of the experiments, networks have been pre-trained with a fully-supervised approach on LoveDA [14], a remote sensing land-cover dataset which includes rural and urban scenes. Below we describe the methodology, the characteristics of datasets and the experiments with a final comment on results.

2. METHODOLOGY

As explained above we investigate the performance of our networks using a fully-supervised approach and then using a semi self-supervised approach aiming at enhancing the generalization capability.

2.1. Fully-supervised approach

The fully-supervised approach is based exclusively on the available labeled data. We resize the images to 512×512 and we apply data augmentation based on random vertical and horizontal flips to avoid overfitting. Images are then feed into the network, outputs and corresponding labels are used to calulate the loss. Finally weights are updated through backpropagation. We use *Cross Entropy* as loss function and *AdamW* as optimizer.

2.2. Semi self-supervised approach

The semi self-supervised approach relies on same resizing and same data augmentation of the fully-supervised one. Inspired by what is proposed in [9] and in [15] we then train our networks for a total of $N = N_{fs} + N_{ss}$ epochs where during the first N_{fs} only labeled images are exploited and for the remaining N_{ss} both labeled and unlabeled images are used. The loss function is given by:

$$Loss = \begin{cases} \frac{1}{b}\sum_i^b L(y_i, p_i), & n < N_{fs}, \\ \frac{1}{b}\sum_i^b L(y_i, p_i) + \alpha\frac{1}{b}\sum_j^b L(f(p_j), p_j)) & n > N_{fs} \end{cases} \tag{1}$$

Where $0 < n < N$ identifies the current epoch, b is the batch size, L is the *Cross Entropy* loss, p_i is the network's prediction and y_i the corresponding given label. Note that p_i has dimensions (C, H, W) while y_i has (H, W) where C is the number of classes, H is the resized height and W is the resized width. The pseudo labels exploited in the self supervised part are defined through the function f as:

$$f(p_j) = argmax(p_j) \tag{2}$$

where $argmax$ is computed along the first dimension, corresponding to the classes. This approach was first proposed and well explained in [15]. Finally, the α coefficient is defined as:

$$\alpha = \frac{n - N_{fs}}{N_{ss}} \tag{3}$$

When $n > N_{fs}$ increases also α increases making the self-supervised part of the loss more and more important ending with $\alpha = 1$. i.e. a perfect balance in the loss between fully-supervised part and self-supervised part. The idea is that the more the semi self-supervised training proceeds the more the network learns to generalize creating pseudo labels more and more accurate. It is important to note that, after N_{fs} fully-supervised epochs, the $argmax$ is used to improve the separation among different decision regions. However, the mistakes the network makes in creating these regions are also enhanced, resulting in a model that becomes more confident in making wrong decisions. For this reason it is critical to mitigate this effect using a loss with the supervised loss added to the self-supervised weighted by α.

2.3. Validation

To be sure that the network is learning a validation is performed on unseen data at each epoch. If the validation metric is better than those calculated in all previous epochs the network weights are saved. The validation metric employed is the DICE score:

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Where TP, FP and FN are respectively true positives, false positives and false negatives. For the final evaluation we also use the $Recall$:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

3. DATASET

As already mentioned, in this work 2 different datasets are used, LoveDA in a fully-supervised pre-training and FloodNet for the final training. The idea is to exploit LoveDA to enhance networks' ability to understand and generalize ground geometries from aerial/satellite images aiming for better performance on FloodNet. Note that both datasets present high class imbalance.

3.1. LoveDA

LoveDA [14] is a high-resolution remote sensing land-cover spaceborne dataset acquired over the cities of Nanjing, Changzhou and Wuhan in July 2016. LoveDA is domain-adaptive, in fact images are grouped in rural and urban. Although ground classes are the same in the two groups, they present different distributions. However, since FloodNet presents both rural and urban scenes together, in our pre-training we are not interested in the domain-adaptive task and we group all the images together. LoveDA presents 7 different classes: background, building, road, water, barren, forest and agricolture. As shown in table 1 we used respectively 2522 and 1669 labeled images for the pre-training and the validation.

3.2. FloodNet

FloodNet [8] is a high-resolution aerial dataset for post-flood scene understanding acquired with small UAVs (DJI Mavic Pro quadcopters) in Texas after Hurricane Harvey in 2017. FloodNet was originally developed to cover tasks of classification, semantic segmentation and vision questioning answering and presents 10 different classes: background, building flooded, building non-flooded, road flooded, road non-flooded, water, tree, vehicle, pool, grass. As already explained, this work investigates the semantic segmentation task, where 398 labeled images are available. As shown in table 1 we divided them in 2 sets for training and validation, composed respectively of 360 and 38. Regarding the self-supervised step we used a total of 1495 unlabeled images. In figure 1 we can observe 2 examples of images with the corresponding labels.

4. EXPERIMENTS

We used three well-known deep semantic segmentation architectures for our task: a U-Net [16], a DeepLabV3 [17] and a PSPNet [18].

The U-Net is based on a contracting path to capture context and a symmetric expanding path that enables precise localization. It can be seen as a encoder-decoder structure. Within the contracting path, encoder layers capture contextual details and diminish the spatial resolution of the input. In contrast,

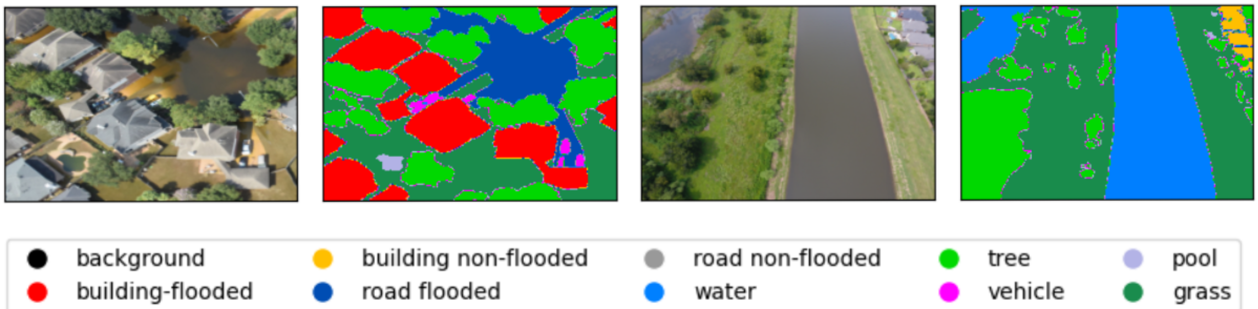


Fig. 1. FloodNet, images and corresponding labels.

Dataset	*Image resolution*	*Train labeled images*	*Val labeled images*	*Unlabeled images*
LoveDA	1024×1024	2522	1669	-
FloodNet	$\sim 4000\times \sim 3000$	360	38	1495

Table 1. Data description.

the expansive path consists of decoder layers that decode the encoded data. Additionally, skip connections leverage information from the contracting path to generate a segmentation map.
The DeepLabV3 is based on a backbone network, a ResNet50 in our case, that extracts features where atrous convolution is used in the last few blocks of the backbone to control the size of the feature map. As final step, an ASPP network is added on top of the backbone..
Finally, the PSPNet is also based on a backbone network, a ResNet34 in our case, used to extract the feature map and then a pyramid pooling module is applied above it. The ASPP network and the pyramid pooling module both aim to collect varied sub-region representations across various scales. Through a final concatenation, for both the DeepLabV3 and the PSPNet, an improved final map is generated.

4.1. Experimental setup

Nvidia Tesla V100-16GB GPUs were used both for the pre-training and for the final training. In all approaches images were resized to 512×512. Pre-training was based on $N = 200$ epochs of fully-supervised training on LoveDA. The fully-supervised training on FloodNet consisted of 200 epochs and the semi self-supervised training consisted of $N_{fs} = 50$ fully-supervised epochs and $N_{ss} = 150$ semi self-supervised epochs. The learning rate was $lr = 0.001$ and batch size was set to $bs = 8$ for all models. Considering the semi self-supervised epochs on FloodNet, due to the lack of memory on GPUs it was impossible to increase the batch size of the unlabeled samples (this is why in equation 1 the batch size for labeled and unlabeled samples is the same). This resulted in a limited number of unlabeled samples that our models were able to see at each epoch, forcing it to be equal to the number of labeled samples, i.e. 360. To overcome this issue and try to exploit all the information available a random shuffling of unlabeled samples was applied at each epoch. In table 2 are shown the setup parameters.

Training	*Resized res*	N	N_{fs}	N_{ss}	lr	bs
Fully-sup	512×512	200	-	-	0.001	8
Semi self-sup	512×512	200	50	150	0.001	8

Table 2. Training setup.

4.2. Results

4.2.1. LoveDA pre-training

Table 3 shows the performance obtained on the LoveDA validation set after pre-training. The reported DICE score is the average of all classes DICE scores. Considering that we

Model	*DICE*
U-Net	0.4
DeepLabV3	0.51
PSPNet	0.44

Table 3. LoveDA validation performance.

grouped together rural and urban images from LoveDA, obtaining highly inhomogeneous sets, results are satisfactory. There is certainly room for improvement but a detailed analysis of the segmentation performance on LoveDA goes beyond the scope of this work.

4.2.2. *FloodNet*

We performed 6 experiments with different configurations. Configurations and results are shown in table 4, were the validation DICE score for each class, the average DICE score and the average Recall are reported. First we tested the U-Net with and without pre-training, using a fully-supervised approach in both cases. The average DICE score remained the same, but the Recall improved. This indicates that the pre-training on LoveDA helped improving the final performance on FloodNet. The corresponding confusion matrices are shown in figure 2. It is clear that the main problem of the U-Net is the recognition of flooded roads, which are often confused with water. It is interesting to notice that the pre-trained U-Net has significantly higher recall values for road classes (both flooded and non-flooded).

We proceeded testing the fully-supervised and semi self-supervised approach on pre-trained PSPNet and DeepLabV3. We observed an improvement of the DICE score and same Recall for the PSPNet and an imporvement on both DICE and Recall for DeepLabV3. This suggests that the idea of joining the fully-supervised approach to the self-supervised approach improved the performance. In figure 3 confusion matrices for DeepLabV3, respectively for the two approaches, are shown. We can observe that the recall of each class, except for vehicle, comparing the fully-supervised approach to the semi self-supervised either improved or remained the same, with a good improvement for the class road-flooded. However, given the small size of the validation set from which the validation metrics have been derived, and the fact that we are talking about small overall improvements, the challenge of finding effective methods to exploit unlabeled data remains largely open. There is a risk that the small improvements observed on the few images of the validation set are not reflected on much larger data. Unfortunately, due to the limited availability of labeled images and the need to use as many of them as possible for training, there were few remaining unseen labeled images to use for validation. A k-fold cross-validation performing multiple training on different subsets and averaging the validation metrics could solve this issue, however this approach would have required a large number of trainings resulting in a timing not in line with the current project. It will be important, in future works, to apply cross-validation to make sure that observed performance differences are significant and to investigate further techniques that use unlabeled data, moving from the fact that there is much room for improvement.

As expected the most difficult classes to recognize for our models are vehicle and pool, less present in our data, and road-flooded that is often misclassified as water probabily because of their closeness in the feature space. For future developments an idea that could reduce the issue of class imbalance is to weigh more the classes less present when computing the loss.

In figure 4 some visual qualitative results are shown.

Here you have a short video that shows how best models work.

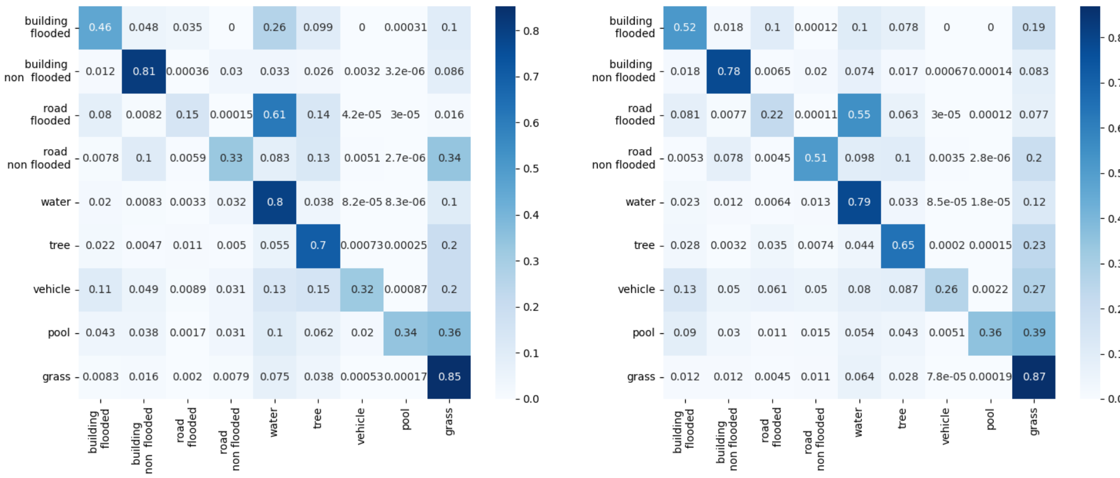
5. CONCLUSIONS

In this work we investigated how some well-known deep architectures perform on a semantic segmentation task on FloodNet. Specifically, we investigated how a pre-train on a high-resolution remote sensing dataset, LoveDA, can help to improve U-Net's performance on FloodNet in a fully-supervised training, and how a semi self-supervised training on pre-trained PSPNet and DeepLabV3 can improve performance respect to a fully-supervised training. Results are encouraging and suggest that, with appropriate pre-train and appropriate self-supervised methods performance can improve further. The challenge of finding methods to proper exploit the unlabeled data to improve generalization capabilities is still open.

For future developments, if more memory for training is available, we think it might be interesting to investigate a W-Net unsupervised pretraining [19] on FloodNet unlabeled images and a subsequent supervised W-Net encoder training on labeled images. As suggested in [9] also self-supervised Vision Transformers [20] could be a valid topic of research for this task.

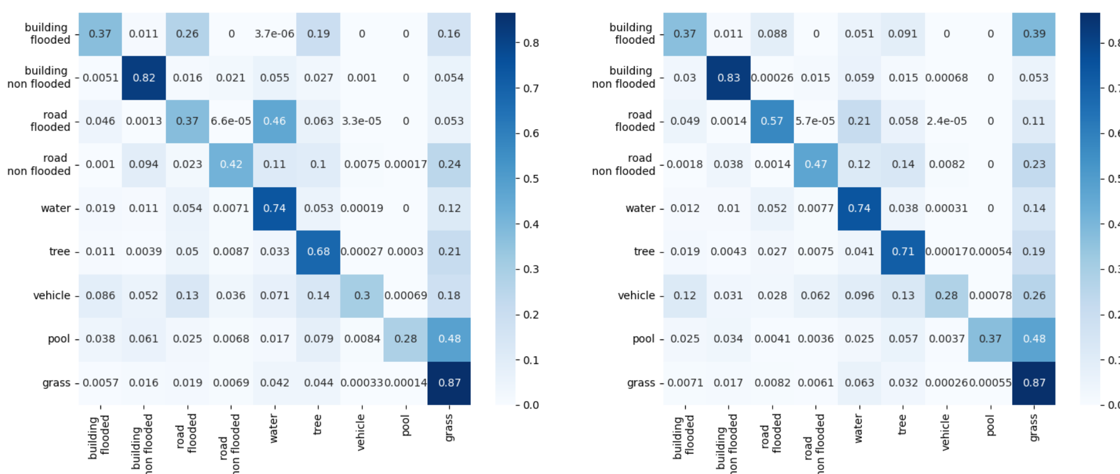
Model	*Pre trained*	*Training*	*Building flooded*	*Building non flooded*	*Road flooded*	*Road non flooded*	*Water*	*Tree*	*Vehicle*	*Pool*	*Grass*	*DICE*	*Recall*
U-Net	No	Fully-sup	0.29	0.49	0.06	0.27	0.43	0.64	0.1	0.3	0.73	0.37	0.53
U-Net	Yes	Fully-sup	0.2	0.48	0.08	0.27	0.5	0.58	0.17	0.3	0.74	0.37	0.55
PSPNet	Yes	Fully-sup	0.4	0.45	0.2	0.35	0.53	0.66	0.11	0.43	0.76	0.43	0.53
DeepLabV3	Yes	Fully-sup	0.38	0.52	0.17	0.38	0.53	0.65	0.21	0.47	0.74	0.45	0.54
PSPNet	Yes	Semi self-sup	0.4	0.52	0.23	0.37	0.54	0.66	0.18	0.68	0.73	0.48	0.53
DeepLabV3	Yes	Semi self-sup	0.32	0.64	0.35	0.41	0.6	0.66	0.3	0.46	0.74	0.5	0.58

Table 4. FloodNet validation performance.



(a) U-Net, not pre-trained, fully-supervised. (b) pre-trained U-Net, fully-supervised.

Fig. 2. U-Net confusion matrices normalized by row



(a) pre-trained DeepLabV3, fully-supervised. (b) pre-trained DeepLabV3, semi self-supervised.

Fig. 3. DeepLabV3 confusion matrices normalized by row.

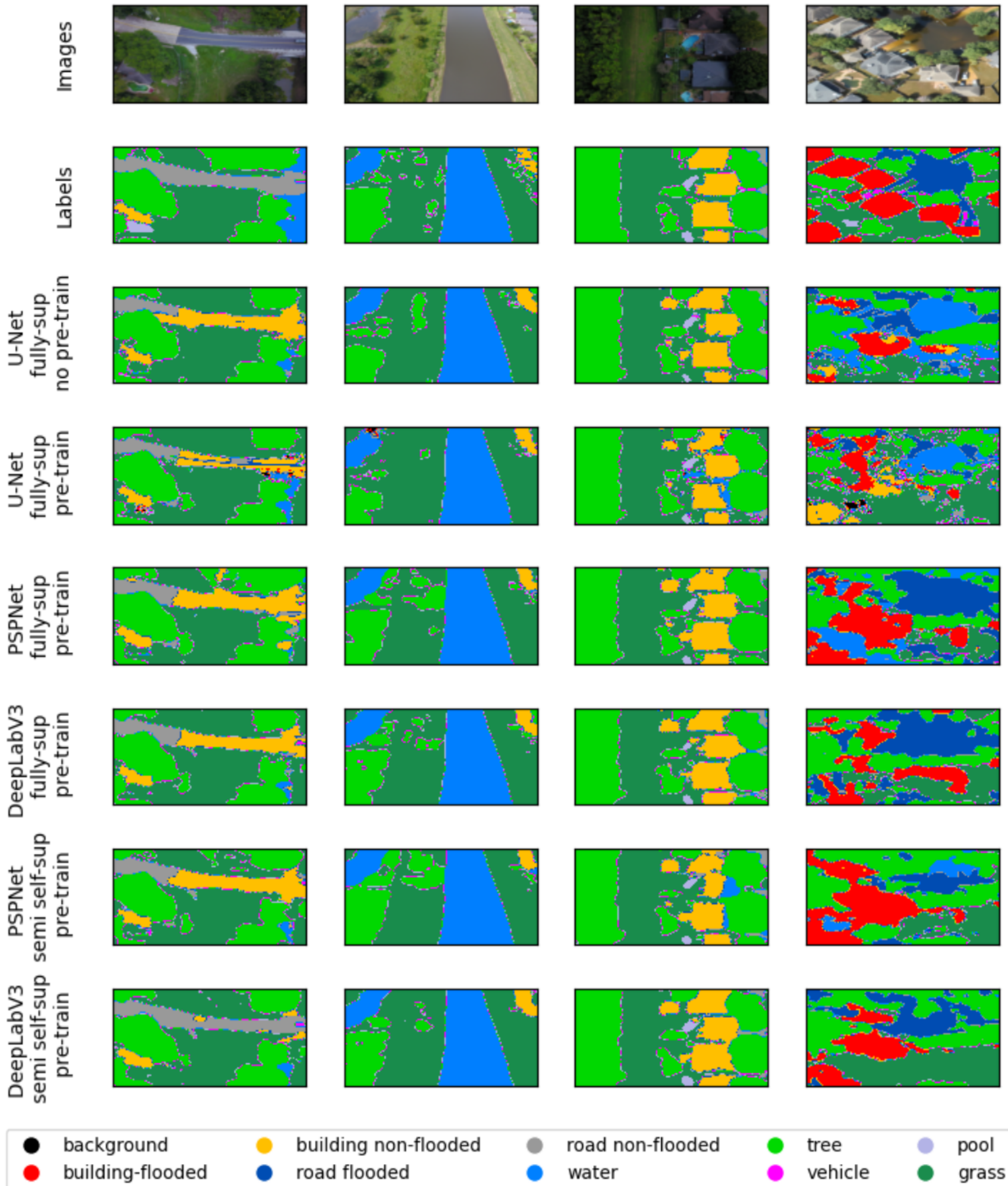


Fig. 4. Qualitative results.

6. REFERENCES

[1] "Distribution of weather-related disaster incidents in europe between 2001 and 2020, by type," https://www.statista.com/statistics/1269886/most-common-natural-disasters-in-europe/, Online; accessed 5 November 2023.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211 – 252, 2014.

[4] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[5] Benjamin Bischke, Patrick Helber, Zhengyu Zhao, Jens De Bruijn, and Damian Borth, "The multimedia satellite task at mediaeval 2018: Emergency response for flooding events," in *MediaEval 2018 - Multimedia Benchmark Workshop*, Martha Larson, Piyush Arora, Claire-Hélène Demarty, Michael Riegler, Benjamin Bischke, Emmanuel Dellandrea, Mathias Lux, Alastair Porter, and Gareth J.F. Jones, Eds. 2018, CEUR Workshop Proceedings, pp. 1–3, CEUR-WS.org, 2018 Working Notes Proceedings of the MediaEval Workshop, MediaEval 2018 ; Conference date: 29-10-2018 Through 31-10-2018.

[6] Ritwik Gupta, Bryce Goodman, Nirav N. Patel, Richard Hosfelt, Sandra Sajeev, Eric T. Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew E. Gaston, "xbd: A dataset for assessing building damage from satellite imagery," *ArXiv*, vol. abs/1911.09296, 2019.

[7] Ethan Weber, Nuria Marzo, Dim P. Papadopoulos, Aritro Biswas, Àgata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba, "Detecting natural disasters, damage, and incidents in the wild," *ArXiv*, vol. abs/2008.09188, 2020.

[8] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[9] Sahil Khose, Abhiraj Tiwari, and Ankita Ghosh, "Semi-supervised classification and segmentation on high resolution aerial images," *ArXiv*, vol. abs/2105.08655, 2021.

[10] Ziquan Wang, "Disaster remote sensing image semantic segmentation model with boundary constraints based on segnext," in *2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, 2023, pp. 734–737.

[11] Farshad Safavi and Maryam Rahnemoonfar, "Comparative study of real-time semantic segmentation networks in aerial images during flooding events," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 15–31, 2023.

[12] Riskyana Dewi Intan Puspitasari, Fadhilah Qalbi Annisa, and Danang Ariyanto, "Flooded area segmentation on remote sensing image from unmanned aerial vehicles (uav) using deeplabv3 and efficientnet-b4 model," in *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2023, pp. 216–220.

[13] Sushant Lenka, Bhavam Vidyarthi, Neil Sequeira, and Ujjwal Verma, "Texture aware unsupervised segmentation for assessment of flood severity in uav aerial images," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 7815–7818.

[14] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," Oct. 2021.

[15] Dong-Hyun Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," 2013.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, vol. abs/1505.04597, 2015.

[17] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017.

[18] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, 2016.

[19] Xide Xia and Brian Kulis, "W-net: A deep model for fully unsupervised image segmentation," *ArXiv*, vol. abs/1711.08506, 2017.

[20] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.