Cross Modal Distillation for Flood Extent Mapping

Shubhika Garg
Google Research

Ben Feinstein
Google Research

Shahar Timnat
Google Research

Vishal Batchu
Google Research

Gideon Dror
Google Research

Adi Gerzi Rosenthal
Google Research

Varun Gulshan
Google Research

Abstract

The increasing intensity and frequency of floods is one of the many consequences of our changing climate. In this work, we explore ML techniques that improve the flood detection module of an operational early flood warning system. Our method exploits an unlabelled dataset of paired multi-spectral and Synthetic Aperture Radar (SAR) imagery to reduce the labeling requirements of a purely supervised learning method. Past attempts have used such unlabelled data by creating weak labels out of them, but end up learning the label mistakes in those weak labels. Motivated by knowledge distillation and semi supervised learning, we explore the use of a teacher to train a student with the help of a small hand labeled dataset and a large unlabelled dataset. Unlike the conventional self distillation setup, we propose a cross modal distillation framework that transfers supervision from a teacher trained on richer modality (multi-spectral images) to a student model trained on SAR imagery. The trained models are then tested on the Sen1Floods11 dataset. Our model outperforms the Sen1Floods11 SAR baselines by an absolute margin of 4.15% pixel wise Intersection-over-Union (IoU) on the test split.

1 Introduction

Floods are one of the major natural disasters, exacerbated by climate change, affecting between 85 million to 250 million people annually and causing between \$32 to \$36 billion in economic damages [14, 11]. Some of these harms can be alleviated by providing early flood warnings, so that people can take proactive measures such as planned evacuation, move assets such as food and cattle and use sandbags for protection. One of the important user experience elements for an effective warning system is its overall accuracy, as false alerts lead to eroded trust in the system. Our work contributes towards improving the accuracy of flood warning systems such as [24], by increasing the accuracy of the inundation module. The inundation model in [24] learns a mapping between historical river water gauge levels and the corresponding flooded area which can be leveraged to predict future flooding extent based on forecast of the future river water gauge level. The accuracy of these forecasts is directly correlated with the accuracy of the underlying historical segmentation maps, hence we aim to improve the segmentation module through our contributions in this work.

In recent years, remote sensing technology has considerably improved and provided high resolution spatial and temporal satellite data. Sentinel-1 [30] and Sentinel-2 [10] satellites are commonly used to map the water surface because they provide open access to high spatial and temporal resolution data. Although Sentinel-2 is better for water segmentation as it shows high water absorption capacity in short wave infrared spectral range (SWIR) and near infrared (NIR) spectrum, it cannot penetrate cloud cover. This limits its application for mapping historical floods as cloud cover is highly correlated with flooding events. On the other hand, radar pulses readily penetrate clouds, making SAR satellites well suited for flood mapping [28, 22, 31].

Thresholding algorithms [20, 21, 6] are traditionally used to segment flooded regions from SAR images because of low back scatter intensity of water. Though techniques like Otsu thresholding [4]

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

work well for many cases, its failure modes include generating false positives for mountain shadows and generating excessive background noise due to speckle in SAR imagery. In recent years, Convolutional Neural Networks (CNN) have been used to segment flooded areas from satellite images. Unlike traditional pixel-wise methods, they can look at a larger context and incorporate spatial features from an image. A lot of work has focused on using opportunistically available cloud free Sentinel-2 images [23, 1]. Though these methods have good performance, their utility at inference time is limited because of the cloud cover issues mentioned above. Another line of work fuses Sentinel-1 and Sentinel-2 images [19, 9, 29, 3] to enhance surface water detection during flooded events. These methods not only a require a cloud free Sentinel-2 image, but also require that both images are taken at about the same time to avoid alignment issues. There has been some work done that uses multi temporal images [33, 27, 32] containing a pre-flood and a post-flood event. These methods can do change detection and exhibit better performance. In our work however, we focus on methods that only take a single Sentinel-1 timestamp image as input.

Sen1Floods11 [5] is a publicly available dataset with a small set of high quality hand labeled images and a larger set of weak labeled images. Most of the prior work that uses a single Sentinel-1 image as input [5, 12, 18, 16], used this weak labeled data to train their models. However, the limitation of using weak labeled data (despite using various regularization techniques), is that the model still learns the mistakes in those labels. Motivated by the semi-supervised methods in [25] and cross modal feature distillation [15], we use a teacher student setup that extracts information from a more informative modality (Sentinel-2) to supervise paired Sentinel-1 SAR images with the help of a small hand labeled and a large unlabelled data set. Similar to [15], we transfer supervision between different modalities. However, instead of supervising an intermediate feature layer like [15], we transfer supervision at the outputs. Our main contribution in this work are:

- We curate an additional large dataset (in addition to Sen1Floods11) from various flooding events containing paired Sentinel-1 and Sentinel-2 images and a weak label based on Sentinel-2 data.
- We propose a cross modal distillation framework and apply it for transfer supervision between different modalities using paired unlabelled data.

2 Datasets

2.1 Input imagery

Sentinel-1 image: Sentinel-1 [30] is an active remote sensing SAR satellite. We use the bands that consist of dual polarized data: Vertical Transmit-Vertical Receive (VV) and Vertical Transmit-Horizontal Receive (VH). These bands represent the log of the backscatter coefficient and are discriminative for detecting surface water as water reflects away all the emitted radiation from the satellite. The wavelength used for imaging is able to see through cloud cover.

Sentinel-2 image: Sentinel-2 [10] is a passive remote sensing satellite operating in visible and infrared wavelength. Its images are affected by atmospheric conditions and often contain significant cloud cover. In this work we use 4 bands: B2 (Blue), B3 (Green), B4 (Red) and B8 (NIR).

2.2 Datasets

Sen1Floods11 dataset: This is is a publicly available dataset [5], containing 4831 tiles from 11 flooding events across 6 continents. It contains paired Sentinel-1 SAR and Sentinel-2 multi-spectral image. Each image is 512×512 pixels at a resolution of 10m per pixel. All images are scaled to 16m per pixel input resolution and projected using to Universal Transverse Mercator (UTM) coordinate system. Due to the high cost of labeling, only 446 images out of 4831 are hand labeled by remote sensing experts to provide a good quality flood water labels. The authors provide an IID split of these hand labeled images, containing 252 training, 89 validation, and 90 test samples. The remaining 4,385 images have weak labels made by thresholding NDVI (Normalized Difference Vegetation Index) and MNDWI (Modified Normalized Difference Water Index) values. These weak labels are only used for training as they are not accurate enough to be used in validation or test.

External dataset (ED): We curated additional imagery by downloading closely acquired Sentinel-1 and Sentinel-2 images from Earth Engine [13] during flood events. An event is considered a flood if the river gauge measurement exceed the official warning level (data made available to us by external partners). We first search for a Sentinel-1 image that overlaps the flooding event duration. We then searched for Sentinel-2 images within 12hrs of Sentinel-1 timestamp and filtered only those that had less than 12% cloud cover (if no such image is found, then the data point was discarded). The data points were extracted from Bangladesh, Brazil, Colombia, India and Peru. These regions were chosen as they are also the regions of interest for final deployment. We also created a weak label for each image using the Normalized Difference Water Index (NDWI) band from the Sentinel-2 image. In total 23,260 image tiles of size 320 × 320, at a pixel resolution of 16m per pixel were created.

3 Methods

Our aim us to segment the flooded pixels using Sentinel-1 SAR image as an input. Formally, let $X_{S1} \in R^{H \times W \times 2}$ be SAR input space and let $Y \in R^{H \times W \times K}$ denote the pixel wise K class one hot label in the output space. Here, $K = 2$ as it contains 2 classes: dry and wet pixels. The paired Sentinel-2 image in training data is represented by $X_{S2} \in R^{H \times W \times 4}$. The hand labeled training set is denoted by $D_l = \left\{X_{S1}^i, X_{S2}^i, Y^i\right\}_{i=1}^{N_l}$ and the larger weak labeled training set as $D_{wl}{}^{**} = \left\{X_{S1}^i, X_{S2}^i, \hat{Y}^i\right\}_{i=1}^{N_{wl}}$. Here Y denotes a high quality label and $\hat{Y}$ denotes a noisy weak label. Our goal is to leverage both D_l and D_{wl} to train the segmentation network. The next section describes the supervised baseline followed by a cross modal distillation framework.

3.1 Supervised baseline

We train two supervised models for comparison. The first model is trained only on hand labeled data D_l. The second model is trained only on the larger weak labeled dataset D_{wl}, so that the network can generalize and avoid memorizing the label errors [26]. We use Deeplab v3+ [7] with an Xception 65 encoder [8] as the model architecture. Common regularization techniques like data augmentations (random crop with distortion, horizontal/vertical flips and colour jitter), dropout, weight decay and batch normalization are used to improve generalization. The network is trained to minimize the cross entropy loss. We apply edge based weighting to the cross entropy loss, which gives higher weights to the inner and outer edges of the binary label.

3.2 Cross modal distillation

In a cross modal distillation framework, the aim is to transfer supervision between two modalities. In our setup, a teacher is trained on stacked Sentinel-1 and Sentinel-2 images using D_l, and is used to supervise a Sentinel-1 only student model on the unlabelled dataset. The advantage of this method over binary weak labels is that the soft labels predicted by the teacher capture uncertainty better compared to the binary labels [2, 17]. Also compared to self distillation, cross modal distillation enables us to provide more accurate supervision by transferring information from a richer knowledge modality. Figure 1 summarizes the training setup used in our work. Both the teacher and the student have an exactly same architecture backbone. Let f_t and f_s represent the teacher and student network function respectively. The training is done in two stages as described below.

Stage 1: Training the teacher network Let X_{S1+S2}^i denote the stacked Sentinel-1 and Sentinel-2 image. $X_{S1+S2}^i \in D_l$ is used as input to the teacher network. The teacher is trained in the same manner as the supervised baseline described in Section 3.1. The training set is small but contains data from geographic locations spanning 6 different continents. This helps the teacher generalize well to different geographies in the unlabelled data seen during the next stage of training.

Stage 2: Training the student network The teacher weights from Stage 1 are kept frozen in this stage. We use paired Sentinel-1 and Sentinel-2 images from Sen1Floods11 hand labeled and ED weak labeled data as the unlabelled data to train the student network. The Sen1Floods11 weak labeled data is not used as the paired Sentinel-2 images were not provided in the dataset. The data in each batch is sampled equally from both the data sources to ensure equal weighting for the datasets. The stacked Sentinel-1 and Sentinel-2 image X_{S1+S2}^i is passed through the teacher to obtain the probabilities $p_t = \sigma(f_t(X_{S1+S2}^i))$ and the augmented paired Sentinel-1 image $\tilde{X}_{S1}^i = Aug(X_{S1}^i)$ is passed through the student to get the student probabilities $p_s = \sigma(f_s(\tilde{X}_{S1}^i))$. KL divergence loss ($L_{KD}$) is then minimized for $K = 2$ classes to update the student weights: $L_{KD} = -\sum_{i=1}^{K} p_t \log p_s$

**Paired Sentinel-2 images are only available for ED weak labeled dataset and are unavailable for Sen1Floods11 weak labeled daaset.

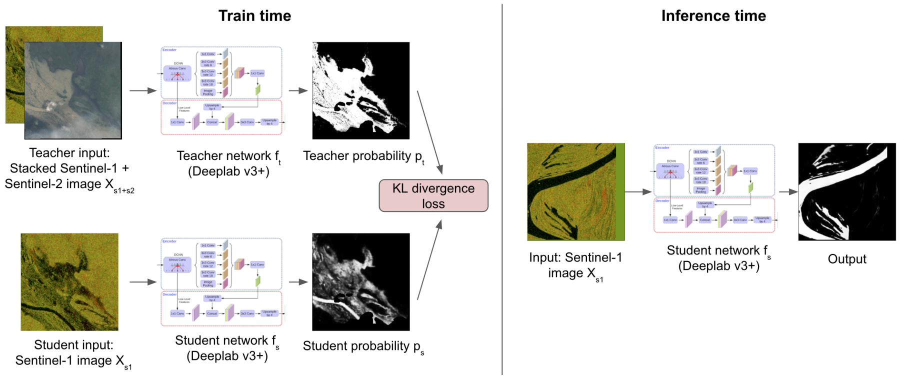

Figure 1: Overview of our cross modal distillation framework. During training a teacher model using Sentinel-1 and Sentinel-2 images is used to train a student using only the Sentinel-1 image. At inference time, only the student is used to make predictions.

4 Results

We use pixel-wise mean intersection over union (IoU) of the water class to validate our model performance. Figure 2 shows a qualitative comparison and Table 1 provides a quantitative comparison on Sen1Floods11 test split.

Method	IoU
Otsu Thresholding	54.58
Hand labeled supervised	67.63 ± 0.45
Weak labeled supervised: Sen1Floods11 weak	67.76 ± 2.41
Weak labeled supervised: Sen1Floods11 + ED weak	68.94 ± 1.11
Cross modal distillation	$\mathbf{71.91 \pm 0.41}$

Table 1: Result of Sentinel-1 supervised baseline models and our cross modal distillation framework on Sen1Floods11 handlabel test split. The numbers show the aggregated mean and standard deviation of IoU from 5 runs.

In Table 1 first two rows, it can be seen that a model trained on large set of Sentinel-2 weak label can match the performance of a model trained on small set of hand labeled data. This empirically verifies the claim that a large weak labeled dataset can act as a quick substitute for a small amount of costly hand label annotations. Including ED weak labeled data to Sen1Floods11 weak labeled data, further led to an increase in the model performance by 1.18% from Sen1Floods11 weak label baseline. This shows that there were still more gains to be had by increasing the weak label dataset size. Our cross distillation model performs better than all the models and exceeds Sen1Floods11 hand label baseline by 4.28% IoU and Sen1Floods11 weak label baseline by 4.15% IoU.

For comparison, we also report the performance of just the teacher model. The teacher model uses stacked Sentinel-1 and Sentinel-2 inputs and has a IoU of 79.25 ± 1.07 on the test set. As expected, it is much higher than the Sentinel-1 supervised hand label model mentioned above because Sentinel-2 image is a richer modality, but not suitable for inference because of cloud cover issues in Section 2.1.

5 Conclusion

We proposed a simple cross modal distillation framework to effectively leverage large amounts of unlabeled and paired satellite data and a limited amount of high quality hand labeled data. We distill knowledge from a teacher trained on the hand labeled images using the more informative modality as input. This helped us generate more accurate labels for the student network as compared to weak labels created by a simple thresholding technique. The student network trained this way outperforms both the supervised hand label and weak label baselines. A promising avenue for future research would be to include temporal imagery to improve performance.

Acknowledgements We would like to thank Yotam Gigi and John Platt for for reviewing the paper and providing invaluable feedback.

References

[1] Peri Akiva, Matthew Purri, Kristin Dana, Beth Tellman, and Tyler Anderson. H2o-net: Self-supervised flood segmentation via adversarial domain adaptation and label refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 111–122, 2021.

[2] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014.

[3] Yanbing Bai, Wenqi Wu, Zhengxin Yang, Jinze Yu, Bo Zhao, Xing Liu, Hanfang Yang, Erick Mas, and Shunichi Koshimura. Enhancement of detecting permanent water and temporary water in flood disasters by fusing sentinel-1 and sentinel-2 imagery using deep learning algorithms: Demonstration of sen1floods11 benchmark datasets. *Remote Sensing*, 13(11):2220, 2021.

[4] Linan Bao, Xiaolei Lv, and Jingchuan Yao. Water extraction in sar images using features analysis and dual-threshold graph cut model. *Remote Sensing*, 13(17):3465, 2021.

[5] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 210–211, 2020.

[6] Kyle M Brown, Crispin H Hambidge, and Jonathan M Brownett. Progress in operational flood mapping using satellite synthetic aperture radar (sar) and airborne light detection and ranging (lidar) data. *Progress in Physical Geography*, 40(2):196–214, 2016.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[9] Georgios I Drakonakis, Grigorios Tsagkatakis, Konstantina Fotiadou, and Panagiotis Tsakalides. Ombrianet—supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2341–2356, 2022.

[10] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

[11] Rebecca E Emerton, Elisabeth M Stephens, Florian Pappenberger, Thomas C Pagano, Albrecht H Weerts, Andy W Wood, Peter Salamon, James D Brown, Niclas Hjerdt, Chantal Donnelly, et al. Continental and global scale flood forecasting systems. *Wiley Interdisciplinary Reviews: Water*, 3(3):391–418, 2016.

[12] B Ghosh, S Garg, and M Motagh. Automatic flood detection from sentinel-1 data using deep learning architectures. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:201–208, 2022.

[13] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.

[14] D Guha-Sapir, R Below, and P Hoyois. Em-dat: international disaster database. 2015. *URL http://www. emdat. be*, 2015.

[15] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[16] Max Helleis, Marc Wieland, Christian Krullikowski, Sandro Martinis, and Simon Plank. Sentinel-1-based water and flood mapping: Benchmarking convolutional neural networks against an operational rule-based processing chain. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2023–2036, 2022.

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[18] Vaibhav Katiyar, Nopphawan Tamkuan, and Masahiko Nagai. Near-real-time flood mapping using off-the-shelf models with sar imagery and deep learning. *Remote Sensing*, 13(12):2334, 2021.

[19] Goutam Konapala, Sujay V Kumar, and Shahryar Khalique Ahmad. Exploring sentinel-1 and sentinel-2 diversity for flood inundation mapping using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180:163–173, 2021.

[20] Jiayong Liang and Desheng Liu. A local thresholding approach to flood water delineation using sentinel-1 sar imagery. *ISPRS journal of photogrammetry and remote sensing*, 159:53–62, 2020.

[21] Sandro Martinis and Christoph Rieke. Backscatter analysis using multi-temporal and multi-frequency sar data in the context of flood mapping at river saale, germany. *Remote Sensing*, 7(6):7732–7752, 2015.

[22] David C Mason, Sarah L Dance, and Hannah L Cloke. Floodwater detection in urban areas using sentinel-1 and worlddem data. *Journal of Applied Remote Sensing*, 15(3):032003, 2021.

[23] Gonzalo Mateo-Garcia, Joshua Veitch-Michaelis, Lewis Smith, Silviu Vlad Oprea, Guy Schumann, Yarin Gal, Atılım Güneş Baydin, and Dietmar Backes. Towards global flood mapping onboard low cost satellites with machine learning. *Scientific reports*, 11(1):1–12, 2021.

[24] Sella Nevo, Efrat Morin, Adi Gerzi Rosenthal, Asher Metzger, Chen Barshai, Dana Weitzner, Dafi Voloshin, Frederik Kratzert, Gal Elidan, Gideon Dror, et al. Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, 26(15):4013–4032, 2022.

[25] Sayak Paul and Siddha Ganju. Flood segmentation on sentinel-1 sar imagery with semi-supervised learning. *arXiv preprint arXiv:2107.08369*, 2021.

[26] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[27] Veda Sunkara, Nicholas Leach, and Siddha Ganju. Memory to map: Improving radar flood maps with temporal context and semantic segmentation. In *AGU Fall Meeting Abstracts*, volume 2021, pages NH35F–07, 2021.

[28] Angelica Tarpanelli, Alessandro C Mondini, and Stefania Camici. Effectiveness of sentinel-1 and sentinel-2 for flood detection assessment in europe. *Natural Hazards and Earth System Sciences*, 22(8):2473–2489, 2022.

[29] B Tavus, S Kocaman, HA Nefeslioglu, and C Gokceoglu. A fusion approach for flood mapping using sentinel-1 and sentinel-2 datasets. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:641–648, 2020.

[30] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012.

[31] Venkata Sai Krishna Vanama, Dipankar Mandal, and Yalamanchili S Rao. Gee4flood: rapid mapping of flood areas using temporal sentinel-1 sar images with google earth engine cloud platform. *Journal of Applied Remote Sensing*, 14(3):034505, 2020.

[32] Ritu Yadav, Andrea Nascetti, and Yifang Ban. Attentive dual stream siamese u-net for flood detection on multi-temporal sentinel-1 data. *arXiv preprint arXiv:2204.09387*, 2022.

[33] Meimei Zhang, Fang Chen, Dong Liang, Bangsen Tian, and Aqiang Yang. Use of sentinel-1 grd sar images to delineate flood extent in pakistan. *Sustainability*, 12(14):5784, 2020.

A Data pre-processing

For Sentinel-1 image normalization, we first clip the VV band from [-20, 0] and VH from [-30, 0] and then linearly scaled these values to the range [0,1]. For Sentinel-2 image, we clipped the 4 bands from [0, 3000] range and then linearly scaled them to [0, 1] range.

B Training details

For all the models, we use Deeplab v3+ model [7] with Xception 65 [8] as the backbone encoder. The skip connection from the encoder features to the decoder is applied at stride 4 and 2. We use a batch size of 64 with input image shape of $(321, 321, C)$ (here $C = 2$ for Sentinel-1 images and $C = 4$ for Sentinel-2 images). For optimization, Momentum optimizer is used with momentum set to 0.9. The learning rate is decayed with a polynomial schedule from initial value to zero with a power of 0.9. The models are trained for 30k steps. A learning rate and weight decay grid search hyperparameter tuning is done by choosing a learning rate from $\{$0.3, 0.1, 0.003, 0.001$\}$ and weight decay from $\{$1e-3, 1e-4, 1e-5, 1e-6$\}$. All the hyperparameter tuning and best model checkpoint selection is then done on the validation split. After the best checkpoint selection, the model is frozen and all the results are reported on the test split.

C Qualitative results

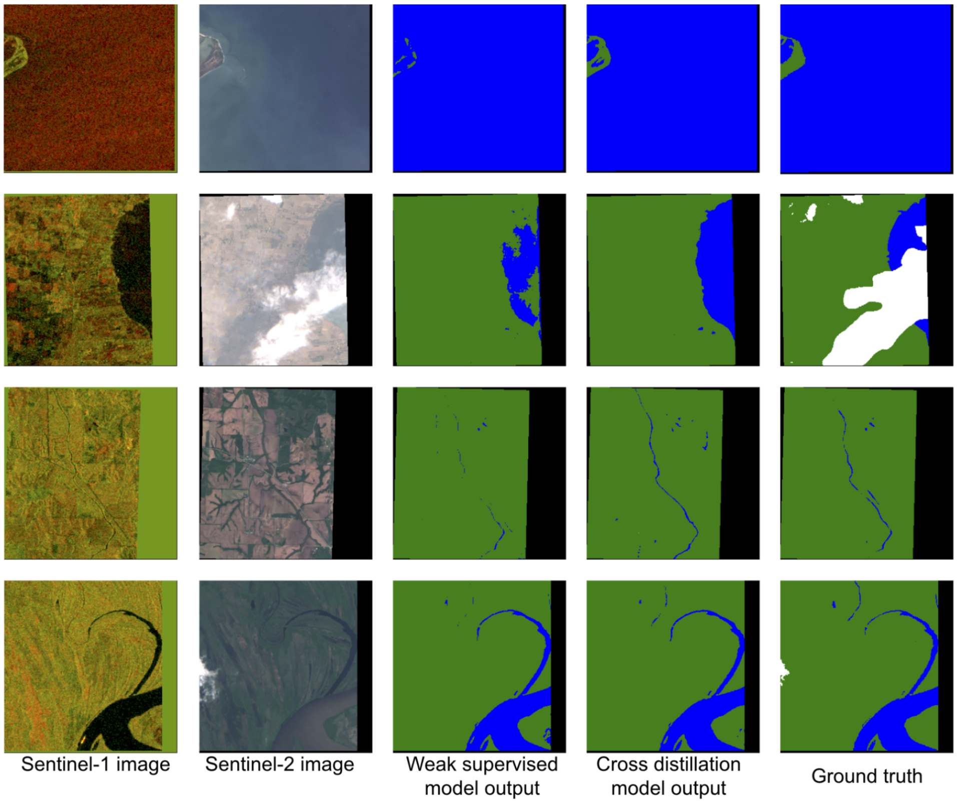

Figure 2: Model inference comparison on Sen1Floods11 handlabel test split. In the output predictions dry pixels are shown in green, water pixels in blue and invalid pixels in black. The ground truth is labeled on the Sentinel-2 image and can contain clouds which are masked in white color. It can be seen that cross modal distillation produces sharper and more accurate results. Weak labeled supervised baseline on the other hand sometimes misses big parts of river due to mistakes learnt from the training data.