Transformer-based Flood Scene Segmentation for Developing Countries

Ahan M R∗
BITS Pilani Goa Campus
ahanmr98@gmail.com

Roshan Roy∗
BITS Pilani
rroshanroy@gmail.com

Shreyas Sunil Kulkarni
BITS Pilani Hyderabad Campus
sskshreyas@gmail.com

Vaibhav Soni
MANIT Bhopal
vaibsoni@gmail.com

Ashish Chittora
BITS Pilani Goa Campus
ashishc@goa.bits-pilani.ac.in

Abstract

Floods are large-scale natural disasters that often induce a massive number of deaths, extensive material damage, and economic turmoil. The effects are more extensive and longer-lasting in high-population and low-resource developing countries. Early Warning Systems (EWS) constantly assess water levels and other factors to forecast floods, to help minimize damage. Post-disaster, disaster response teams undertake a Post Disaster Needs Assessment (PDSA) to assess structural damage and determine optimal strategies to respond to highly affected neighborhoods. However, even today in developing countries, EWS and PDSA analysis of large volumes of image and video data is largely a manual process undertaken by first responders and volunteers. We propose FloodTransformer, which to the best of our knowledge, is the first visual transformer-based model to detect and segment flooded areas from aerial images at disaster sites. We also propose a custom metric, Flood Capacity (FC) to measure the spatial extent of water coverage and quantify the segmented flooded area for EWS and PDSA analyses. We use the SWOC Flood segmentation dataset and achieve 0.93 mIoU, outperforming all other methods. We further show the robustness of this approach by validating across unseen flood images from other flood data sources.

1 Introduction and Context

The Center for Research on the Epidemiology of Disasters, in affiliation with the World Health Organization (WHO), reported that natural disasters accounted for 1.3 million deaths and over USD 2 trillion in economic damage — all between 1998 and 2017 [19]. Flooding related damage is a factor in most of them [4] and frequent the list of most expensive disasters [17]. Developing economies of Asia are disproportionately affected and are the worst-hit by floods, accounting for 44% of all flood disasters from 1987-1997 [18]. India alone registers 1/5th of global deaths from floods [11]. Rapid urbanization, global climate change, and rising sea water levels will expose 1.47 billion more people to flood risk, with 89% of them living in low-middle income countries [5]. Flood Segmentation technology is instrumental for Disaster Prediction and Response is critical to save lives and livelihoods.

Flood Response: Typically, disaster management teams complete a Post Disaster Needs Assessment (PDSA) and rapidly develop infrastructure based on this report on the collected data [6]. Unmanned Aerial Vehicles (UAVs) are deployed to collect large volumes of image and video data in affected areas.
regions. PDSA is essential for identification of submerged regions, sanity check of large building structures, debris identification, and search-and-rescue (S&R) operations.

**Flood Forecasting:** Flood segmentation techniques can be critically important for flooding-related Early Warning Systems (EWS). According to research in [12], Indians given a flood warning are twice as likely to evacuate safely than Indians without any notice, which require constant monitoring of river or sea water levels. Comparison of current levels with historical evidence of flood-prone water levels can help understand when to trigger warnings appropriately.

**Constraints:** Developing countries are plagued by resource and economic constraints. Failure of macro- and micro-infrastructure planning in Nicaragua led to re-construction on top of an earthquake faultline [7]. Weak social safety and insurance policies inflate recovery time [27]. Economic vulnerability renders countries like Haiti, Ethiopia, Nepal, El Salvador in a near-permanent state of emergency alert [7]. In these countries, processing and analysis of large-scale visual data from UAVs for PDSA in Flood response is a manual process that requires multi-team intervention, which poses a serious bottleneck in search-and response speed. Deployment of EWSs is infeasible because human monitoring of video feeds is too cumbersome and expensive.

**AI Technology:** To reduce the burden of manual analysis on crisis responders, Deep learning is well-suited to scale, automate and expedite these operations. The last few years have witnessed a tremendous rise in CNN-based image classification and segmentation research [21]. However, CNNs suffer from a well-known problem — large inductive biases. Concretely, CNNs assume locality and translation equivariance, which hurt the interpretability of pure CNN-based algorithms. Recently, visual transformers have garnered attention for image classification, segmentation and object detection tasks [2, 20, 26, 3] for challenging these assumptions with comparable accuracy.

**Contributions:** In this work, we propose a hybrid fused CNN-Transformer: **FloodTransformer** to tackle flood water segmentation on the Water Segmentation Open Collection (WSOC) dataset [22]. First, we achieve state-of-the-art results and are the first work (to the best of our knowledge) to apply new transformer-driven research to the flood data domain. Second, our approach is extendable — we demonstrate the ability of our model to generalize well on unseen data sources. Further local calibration, if required at all, simply requires weight fine-tuning with previous, region-specific, flood scene data. Third, our model does not suffer from data scarcity - it only requires image data input and not complex sensor data which is hard to collect [13]. Last, the transformer-based encoder applies recent DL innovations to the flood data domain. Although the hybrid method still uses CNNs in the decoder network, the aforementioned spatial inductive biases no longer occur throughout the entire network. Dependencies between patch embeddings are learnt from scratch. This improves the robustness of our approach.

## 2 Methodology

To achieve the Flood Scene Understanding, we introduce a Deep Learning model for Flood image segmentation and quantify the impact of flooding with a custom metric called **Flooding Capacity**.

### 2.1 Method

Inspired by Zhang et al. [23], we propose **FloodTransformer** to solve segmentation for the flood data domain. It is a fusion architecture of Visual Transformer [25] and Convolution Neural Networks (CNNs) and its model architecture is displayed in Figure 1.

Complex flooding imagery may contain heterogeneous objects, flooding patterns and backgrounds. Using the self-attention module of the visual Transformer module from [25] and global vector representation learned from the CNN network, **FloodTransformer** fuses the trained embeddings to learn long-term spatial relationships between the aforementioned entities in images of flood affected areas. Using Hadamard bilinear product [23], the fusion module fuses information via embeddings from both parallel streams into a dense representation. The combination of multi-level fusion maps generates the segmentation output of the model. We summarize each component below, per Zhang et al. [23].

**Transformer Module:** We use the encoder-decoder network using Visual Transformer [25]. The input image $x \in \mathbb{R}^{H \times W \times 3}$ is sliced into $N$ patches, where $N = \frac{H}{F} \times \frac{W}{F}$ and $F$ is usually set to 16 or
32. These patches are flattened into a linear projection layer to generate an image embedding, added to a trainable position embedding, and passed into $L$ layers of the Transformer encoder’s Multi-head Self Attention (MSA) mechanism. Every layer updates the embedding with $[q, k, v]$ triplets and the projection matrix. The encoded sequence $z^L$ is passed to the Transformer decoder, which aims to recover spatial context and outputs feature maps $t^0, t^1, t^2$ of different scales.

**CNN Module**: Per [23] and [15], we implement three Res-blocks as a shallow CNN network; Transformers already provide global context and retain rich local information and relationships. In parallel to the transformer module, the three CNN blocks output downsampled feature maps $v_0, v_1, v_2$ of different scales.

**Fusion Module**: Outputs $v_0, t_0 \in \frac{H}{2} \times \frac{W}{2}$, $v_1, t_1 \in \frac{H}{4} \times \frac{W}{4}$ and $v_2, t_2 \in \frac{H}{8} \times \frac{W}{8}$ from both modules are fused via the Fusion module from Zhang et al. [23]. The fused vector $f_i$ is generated by implementing the following operations:

$$\hat{t}^i = ChannelAttn\left(t^i\right), \quad \hat{g}^i = SpatialAttn\left(g^i\right), \quad \hat{b}^i = Conv\left(t^iW_1 \odot g^iW_2\right)$$

where $\odot$ is the Hadamard product as described in [23].

### 2.2 Metric

We introduce a custom metric called *Flooding Capacity (FC)* to quantify the impact of flooding in an image. It is the percentage of flooding in an image, and is calculated from the segmented mask output of the image, as described in Equation 1.

$$FloodingCapacity(FC) = \frac{1}{\sum_{i=0}^{k} \sum_{j=0}^{k} (p_{ij} == 0)\ Sum((p_{ij} == 1))} \sum_{i=0}^{k} \sum_{j=0}^{k} (p_{ij} == 1)$$  \hspace{1cm} (1)$$

$p_{ij}$ are the pixel values of the binary mask output and can have values 0, 1. We calculate the flooding area as the sum of all $p_{ij} == 1$ flooding pixels. We divide flooding area by total area, which is the sum of all $p_{ij} == 0$ non-flooding and $p_{ij} == 1$ flooding pixels.

Flood Scene Segmentation, as described in Section 2.1, used with *Flooding Capacity* calculation from Equation 1 can be used on images pre/post flooding for complete Scene understanding. For Flood Response, this helps optimize operational planning. For Flood Forecasting, this can quantify the water-cap in a particular region over time separated image inputs.

### 3 Dataset Description

**Water Segmentation Open Collection (WSOC)** dataset [22] consists of 11900 flood or water body related images, composed from multiple open-source source datasets such as COCO [9], Semantic...
### Table 1: Average and Standard deviation of Mean Intersection over Union (mIoU) and Pixel Accuracy (PA) on WSOC dataset

| Model                  | Backbone | WSOC Dataset | mIOU [%] | PA [%] |
|------------------------|----------|--------------|----------|--------|
|                        |          |              | μ        | σ      | μ      | σ      |
| Tiramisu [8]           | None     |              | 0.38     | 0.17   | 0.73   | 0.24   |
| SegNet [1]             | ResNet50 |              | 0.85     | 0.08   | 0.94   | 0.01   |
| UNet [14]              | ResNet50 |              | 0.88     | 0.06   | 0.94   | 0.03   |
| FCN32 [16]             | ResNet50 |              | 0.79     | 0.10   | 0.87   | 0.08   |
| PSPNet [24]            | ResNet50 |              | 0.83     | 0.08   | 0.90   | 0.07   |
| SegNet [1]             | VGG16    |              | 0.82     | 0.09   | 0.90   | 0.05   |
| UNet [14]              | VGG16    |              | 0.5      | 0.26   | 0.67   | 0.15   |
| FCN32 [16]             | VGG16    |              | 0.76     | 0.13   | 0.91   | 0.02   |
| PSPNet [24]            | VGG16    |              | 0.82     | 0.09   | 0.92   | 0.02   |
| FloodTransformer (Ours)| Transformer |              | 0.93     | 0.03   | 0.96   | 0.02   |

Drone Dataset [10] and other datasets containing flood-related images [22]. These images were shot from a variety of sources such as fixed surveillance cameras, drones (UAVs), crowdsourced in-field observations (on foot, boat, vehicle) and from social media streams. Some sample images are shown in Figure 2.

Image resolutions vary from 147x150 pixels for low-quality social media streams to 2448x3264 pixels for sophisticated aerial imagery. Sampling from disparate sources is the reason for significant variations in visual appearance, illumination, water body, and geography; this makes water segmentation challenging. Newly-added social media images undergo a three-step manual annotation process to generate binary ground truth masks for pixel-wise classification between water or no-water.

### 4 Results

We follow the WSOC method [22] and perform a 90-10 train-test split of the data. Qualitative results of output segmentation masks are shown in Figure 2. We evaluate the performance of our hybrid Flood Transformer model with mean intersection-over-union (mIOU) and Percentage Accuracy (PA) scores, on WSOC, as shown in Table 1. We compare our result to that of other benchmark models on WSOC [22], in Table 1.

As shown, FloodTransformer comfortably achieves state-of-the-art segmentation mIOU of 0.93. It is based on Zheng et al [23]’s TransFuse architecture, which comfortably beats traditional benchmarks such as UNet [14]. The Flood Capacity (FC) from Equation 1 of predicted mask in Fig 2(i)(c) is 0.53 and of mask in Fig 2(ii)(c) 0.47.

To demonstrate the robustness of our model with further qualitative results, we perform inference on out-of-dataset images, as shown in Figure 3. The Flood Capacity (FC) from Equation 1 for the predicted mask in Fig 3(i)(b) is 0.18 and for mask in Fig 3(ii)(b) is 0.26.

### 5 Conclusions and Impact

In this work we propose, to the best of our knowledge, the first hybrid transformer-CNN model for flood water segmentation: FloodTransformer. We achieve state-of-the-art segmentation results on the WSOC dataset and empirically demonstrate the power and robustness of visual transformers, deviating from pure CNN-based approaches.

We further validate our model’s performance by demonstrating competitive qualitative results on out-of-dataset aerial images, indicating competitive performance on images in the wild. With the increasing number and variability of flooding-related disasters in developing countries (varying weather conditions, water flow, geography etc.) and variability of visual data collected from UAVs (varying mode of capture, camera technology etc.), this generalization capacity is crucial.
The flood water segmentation methodology adapted in this paper is useful for Flood Prediction and Response. Given the high incidence of flooding-related disasters, our method is an important step towards improving flood warning systems and covering billions of people in high-risk developing countries. It can also aid in the first step of disaster management: Post Disaster Needs Assessment (PDSA). In particular, in both cases it reduces the need for manual analysis; this is crucial in developing countries with constrained resources and high populations, where speed and efficiency of response is of paramount importance.
References

[1] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. 05 2015.

[2] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. ArXiv, abs/2103.14586, 2021.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, pages 213–229, Cham, 2020. Springer International Publishing.

[4] Melissa Denchak. Flooding and climate change: Everything you need to know. 2020.

[5] JUN ERIK and RENTSCHLER MELDA SALHAB. 1.47 billion people face flood risk worldwide: for over a third, it could be devastating. 2020.

[6] FEMA.gov. Preliminary damage assessment guide. 2016.

[7] United States Government. U.s. disaster assistance to developing countries: Lessons applicable to u.s. domestic disaster programs: A background paper. 1980.

[8] Simon Jégou, Michal Drozdzal, David Vázquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1175–1183, 2017.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 740–755, Cham, 2014. Springer International Publishing.

[10] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 165:108–119, 2020.

[11] Chaitanya Mallapur. India accounts for 1/5th of global deaths from floods, as climate change warning is sounded. 2021.

[12] Berman Micah, Jagnani Maulik, Nevo Sella, Pande Rohini, and Reich Ofir. Using technology to save lives during india’s monsoon season. 2021.

[13] Sella Nevo, Gal Elidan, Avinatan Hassidim, Guy Shalev, Oren Gilon, Grey Nearing, and Yossi Matias. ML-based flood forecasting: Advances in scale, accuracy and reach, 11 2020.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pages 234–241, Cham, 2015. Springer International Publishing.

[15] Roshan Roy, MR Ahan, Vaibhav Soni, MANIT Bhopal, and Ashish Chittora. Towards automatic transformer-based cloud classification and segmentation.

[16] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):640–651, 2017.

[17] BRINK Editorial Staff. The 10 most costly natural disasters of the century. 2019.

[18] United Nations University. Two billion vulnerable to floods by 2050; number expected to double or more in two generations. 2004.

[19] Pascaline Wallemacq, UNISDR, and CRED. Economic losses, poverty and disasters 1998-2017, 10 2018.

[20] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, J. Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. ArXiv, abs/2105.15203, 2021.

[21] Mutlu Yapıcı, Adem Tekerek, and Nurettin Topaloglu. Literature review of deep learning research areas. 5:188–215, 12 2019.

[22] Mirko Zaffaroni and Claudio Rossi. Water segmentation dataset, December 2019.

[23] Yundong Zhang, Huuye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation, 2021.

[24] 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), pages 6230–6239, 2017.

[25] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer, 2021.
[26] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In International Conference on Learning Representations, 2021.

[27] Matija Zorn. Natural Disasters and Less Developed Countries, pages 59–78. 01 2018.