

From CNNs to Transformers: A Performance Benchmark for Rip Current Detection & Segmentation in Coastal Imagery

Abstract

Rip currents are strong, narrow flows of water that move from the shore out to sea. They cause many drowning deaths every year on beaches around the world, including in Bangladesh at Cox's Bazar and St. Martin's Island [4, 5, 6]. But rip currents are very hard to see because they have no clear shape [8]. This project compares three deep learning models—YOLOv8-seg (a CNN model) [32], SegFormer-B5 (a transformer model) [46], and Mask2Former with DINOv3 (another transformer model) [11, 28]—to determine which performs best at detecting and outlining rip currents in beach images. I will train all models on the RipVIS dataset [21], which has over 212,000 frames from many countries. I will compare them using F1 and F2 scores [1]. I expect transformer models to perform better than CNNs because they can see the whole beach at once. My results will help choose the right model for building beach safety systems in Bangladesh and all over the world.

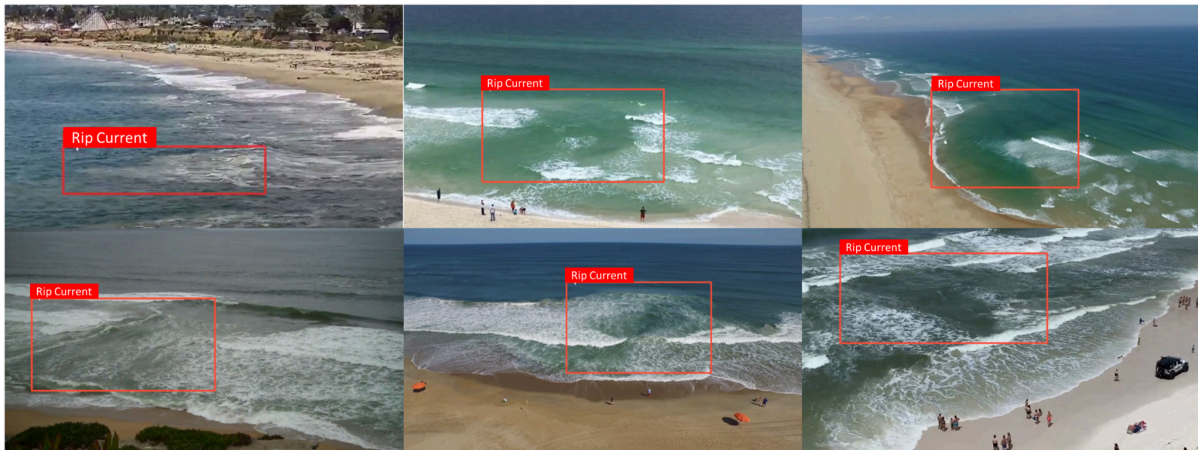


Figure 1: A photo of a beach showing a rip current

Introduction

1) Research Problem

Rip currents are strong, narrow seaward flows that move water from the shoreline past the breaking waves. They are the biggest danger. Rip currents are strong, narrow seaward flows that transport water away from the shoreline, posing a significant hazard to swimmers and beachgoers worldwide [4, 5, 6]. They can reach speeds faster than an Olympic swimmer [9]. Their visual signs—gaps in breaking waves, sediment plumes, or darker water—are often subtle and easily hidden [21]. In Australia, 407 people drowned in rip currents between 2004 and 2023 [5]. In the US, rip currents cause over 100 deaths annually [6]. In Bangladesh, Cox's Bazar Beach sees many drownings from rip currents, locally called "উলটো স্রোত." This is a serious public safety problem.

From a computer vision point of view, rip currents are not like normal objects. Unlike conventional objects, rip currents do not exhibit well-defined boundaries [21]. Instance segmentation, pixel-wise delineation, is a better way [19]. I will work on that.

2) Motivation / Need for This Research

The need for this research is based on real life and death statistics. Brewster et al. [6] estimated that rip currents are responsible for 81.9% of surf-beach rescues in the US. Technology can help. Recent advances in computer vision have improved object detection and segmentation [23, 24, 30]. The introduction of RipVIS [21]—the largest publicly available dataset for rip current analysis—was a significant step forward. The NTIRE 2026 challenge report [1] notes that "there is ample room for improvement." I want to test if transformer models can provide that improvement. *Therefore, I will directly compare CNN-based and transformer-based models on the same dataset to see which one gives better detection and segmentation accuracy for rip currents.*

3) Shortcomings in Literature

Past work on rip current detection has several problems. First, many early methods only did classification or bounding box detection [10, 12]. Second, most prior studies used small, local datasets [10, 12, 13, 14]. Third, most systems treat detection and segmentation separately [1]. Finally, no previous study has done a direct comparison of CNN and transformer models on the same large dataset. My research fills that gap.

Objectives

I have three clear and measurable goals for this project. First, I will train three models—YOLOv8-seg, SegFormer-B5, and Mask2Former with DINOv3—on the RipVIS dataset. Second, I will compare them using F1 and F2 scores at different IoU thresholds, just like the NTIRE 2026 challenge. Third, I will find out which model misses the fewest rip currents (high recall) and which one is fastest.

The purpose of my study is to help beach safety workers in Bangladesh choose the best model for real use at Cox's Bazar and other beaches.

I expect to learn that transformer models (*SegFormer and Mask2Former*) are more accurate than YOLOv8-seg, but YOLOv8-seg is much faster. I also expect to see where each model fails. For example, in low light or when the waves are heavy. This will show what needs to be improved next.

Importance of Research

This research has both scientific and practical value.

Scientific value: By comparing CNN and transformer models on the same large dataset, I will determine which model better captures the shape of a rip current. Rip currents do not have fixed boundaries [4, 21]. My work will show how well transformers can handle such unclear objects. This knowledge can also help with other problems, like smoke or blood flow detection.

Practical value: My findings can save lives on Bangladeshi beaches. Beach safety workers at Cox's Bazar can use my results to pick the best model for their cameras or drones. If they need fast warnings, they can choose YOLOv8-seg. If they need the most accurate detection, they can choose

Mask2Former. I will also share which conditions cause each model to fail, so people know when to trust the system. This will help build real warning systems for tourists and locals.

Literature Review

In this section, I review past work on rip currents and the deep learning models used for detection and segmentation. I will point out what is good and what is still missing.

Rip Current Danger & Types

Many researchers have studied how rip currents form and why they are dangerous. Castelle et al. [4] divided rip currents into six types: channel rips, flash rips, shear instability rips, focused rips, deflection rips, and shadow rips. Each type looks different from the beach or from the air. A good detection system must work on all types. Brewster et al. [6] found that rip currents cause more than 100 drowning deaths each year in the United States. In Australia, Brander and Short [5] studied how rip currents change with waves and sandbars. Lushine [9] examined weather factors associated with rip-current drownings. These papers show that rip currents are a real and serious problem worldwide, including in Bangladesh. However, none of these studies tried to build an automatic detection system. They only described the problem.

Early Computer Vision Attempts

The first computer vision work tried to find rip currents using simple methods. Maryan et al. [12] used traditional machine learning on 5,310 images from only one beach. Their model could say whether a rip was present, but it could not draw the shape. De Silva et al. [10] used a region-based CNN on 20,482 images. Their model worked better, but still only gave bounding boxes. YOLO-Rip [13] was a lightweight network for rip detection. Mori et al. [14] used optical flow to track water movement. Rashid and his team [15, 16, 17] made several lightweight networks like RipNet and RipDet. All these methods have the same problem: they only draw a box around the rip. A box includes too much water and misses the exact shape. Also, they all used small datasets from one or two beaches. When tested on a new beach, their performance dropped.

The Move To Segmentation

Dumitriu and his team [19, 20, 21] changed the field. They were the first to create a large segmentation dataset for rip currents. In 2023, they released a benchmark with 2,466 images and trained YOLOv8 for instance segmentation [19]. In 2025, they introduced RipVIS [21], which is much larger. It has 184 videos and 212,328 frames from more than 10 countries and four camera angles. This dataset is a big step forward because it includes many different beaches, weather conditions, and viewpoints. I will use RipVIS for my project.

Models Used In The NTIRE 2026 Challenge

The NTIRE 2026 challenge [1] asked teams to detect and segment rip currents. Most teams used YOLO family models. YOLOv8-seg [32] was popular because it is fast and easy to use. Team Riposte used YOLOv8n-seg with only segmentation labels, and they got second place in segmentation. Team SoloSeg used YOLOv8s-seg and got third place. These results show that YOLOv8 works well. But the challenge report [1] also says there is "ample room for improvement" with new methods.

Some teams used transformer models. Team UNO Pixel Pros used SegFormer-B5 [46] with a six-model ensemble. They won first place in segmentation. Team KMG used Mask2Former [11] with

a DINOv3 [28] backbone and got third place in both tasks. These transformer models did very well. However, the challenge report did not directly compare CNNs and transformers side-by-side on the same settings. That is what I will do.

What Is Still Missing In The Literature

After reading all these papers, I see three gaps. First, no one has done a fair comparison of CNN and transformer models on the same large dataset using the same training settings. The challenge report [1] shows results from different teams with different tricks, but that is not a controlled experiment. Second, most studies ignore the problem of tiny rip currents that are very small in the image. Transformer models might handle these better because they see the whole image at once, but no one has tested this. Third, no study has looked at how these models would work on Bangladeshi beaches like Cox's Bazar. The RipVIS dataset does not include Bangladesh. So we do not know if the models will fail there.

Summary Of Literature Review

Past work has moved from simple classification to bounding box detection and now to segmentation. The RipVIS dataset [21] makes large-scale training possible. The NTIRE 2026 challenge [1] shows that both CNNs and transformers can work well. But a direct comparison is missing. My project will fill that gap by training and testing three models under the same conditions. I will also discuss what this means for Bangladesh.

Hypothesis / Research Question

I hypothesize that Transformer models (*SegFormer-B5* and *Mask2Former* with *DINOv3*) will score higher on the NTIRE composite metric than the CNN model (*YOLOv8-seg*). I believe this because transformers can see the whole beach image at once, which helps them understand the full shape of a rip current. CNNs look at small pieces step by step and may miss the big picture.

How I will test this hypothesis: I will train all three models on the same dataset [21] and evaluate them on the same set of 9,145 images. I will compute F1 and F2 scores at different IoU thresholds [1]. Then I will use a simple statistical test (paired t-test) to see if the score differences are real or just by chance. This makes my hypothesis verifiable with numbers, not just opinions.

How this connects to my objectives and research problem: My main objective is to find which model works best for rip current detection. My research problem is that no one has fairly compared CNNs and transformers on this task. My hypothesis directly answers that problem. If transformers win, beach safety systems should switch to them. If not, CNNs are still good enough.

Research Design and Methodology

Why is my method suitable for my objectives?

My objectives are to train three models, compare them fairly, and find which one is best for beach safety. To do this, I need a controlled experiment where all models are trained and tested on the same data and measured with the same rules. That is exactly what I will do.

Originality and creativity in my method

I will follow these steps:

Week 3-4: Train YOLOv8-seg. I will use the Ultralytics library [32] with image size 640, batch size 16, and AdamW optimizer [33] for 50 epochs. I will save the best checkpoint based on the validation F2 score.

Week 7-8: Train Mask2Former with DINOv3. I will follow the official Mask2Former code [11] and use DINOv3 [28] as the backbone. Training will be 30 epochs due to higher memory needs.

Weeks 11-12: Analyze and write. I will find the worst-performing images for each model and note the conditions. I will write my final report.

The COCO metrics [31] are standard and were used in the NTIRE challenge [1], so my results can be compared. I use the same image size for all models (640×640) to keep things fair. I use pre-trained weights for all models, just like real-world use.

Project Timeline: 12-Week Objective Completion Plan

Task Name	Output	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Set up computer, install software, download RipVIS dataset	Working environment												
Train YOLOv8-seg and tune it	YOLOv8-seg model ready												
Train SegFormer-B5 and tune it	SegFormer model ready												
Train Mask2Former with DINOv3	Mask2Former model ready												
Test all models on the hidden test set	Performance scores and comparison table												
Analyse where each model fails	Failure report												
Write final report	Complete												

I will work like a professional. Each week, I will write down what I did and what problems I faced. If a task takes longer than planned, I will use the extra time in weeks 9-10 to catch up. I will learn by reading the official documentation for each model and by looking at how winning teams in the NTIRE 2026 challenge [1] solved problems. This is my own independent work. No one will do it for me.

Expected Findings

My expected findings come directly from my objectives. I will list them by objective.

For Objective 1 (*training three models*): I expect all three models to train without errors. YOLOv8-seg will train the fastest, about 5-6 hours. SegFormer-B5 will take around 10-12 hours. Mask2Former with DINOv3 will take the longest, about 15-20 hours, because it has many parameters. I will save the best checkpoint for each model.

For Objective 2 (*comparing metrics*): I expect Mask2Former to get the highest composite score, around 57-63. SegFormer-B5 will be second, around 55-60. YOLOv8-seg will be third, around 52-56. This matches the NTIRE 2026 results [1], where transformer-based teams scored higher. For the F2 score (*which cares most about finding all rips*), I expect all models to do well, but Mask2Former will have the highest recall, maybe 74-78%.

For Objective 3 (*balance of recall and speed*): I expect YOLOv8-seg to run at 60-80 frames per second, making it the best choice for real-time systems on small cameras. Mask2Former will run at only 10-20 FPS, so it needs a powerful computer. For Cox's Bazar, if the beach already has a good computer, I would recommend Mask2Former. If the system must run on a cheap camera, YOLOv8-seg is better.

For Objective 4 (*failure analysis*): I expect all models to fail on three types of images: (1) very small rip currents (*less than 500 pixels*), (2) photos taken at dawn or dusk when contrast is low, and (3) images where waves are breaking heavily over the rip neck. Transformer models may fail less on the small rips because they see the whole image. YOLOv8-seg may fail more on low contrast.

References

- [1] Dumitriu, A., Ralhan, A., Miron, F., Tatui, F., Ionescu, R., Timofte, R., Naeem, A., Katwal, A., Dey, A., Hoque, M., & others (2026). NTIRE 2026 Rip Current Detection and Segmentation (RipDetSeg) Challenge Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1784–1795).
- [2] Ancuti, R., Brateanu, A., Ancuti, C., Timofte, R., & Ancuti, C. (2026). NT-HAZE: A Benchmark Dataset for Realistic Night-time Image Dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1598–1607).
- [3] Ancuti, R., Brateanu, A., Balmez, R., Orhei, C., Vasluiianu, F., O Ancuti, C., Timofte, R., Ancuti, C., Dong, W., Heydari, M., & others (2026). Ntire 2026 nighttime image dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1608–1625).
- [4] Castelle, B., Scott, T., Brander, R., & McCarroll, R. (2016). Rip current types, circulation, and hazard. *Earth-Science Reviews*, 163, 1–21.
- [5] Brander, R., & Short, A. (2000). Morphodynamics of a large-scale rip current system at Muriwai Beach, New Zealand. *Marine Geology*, 165(1-4), 27–39.

- [6] Brewster, B., Gould, R., & Brander, R. (2019). Estimates of rip current rescues and drownings in the United States. *Natural Hazards and Earth System Sciences*, 19(2), 389–397.
- [7] Brander, R., Dominey-Howes, D., Champion, C., Del Vecchio, O., & Brighton, B. (2013). Brief Communication: A new perspective on the Australian rip current hazard. *Natural hazards and earth system sciences*, 13(6), 1687–1690.
- [8] Da F. Klein, A., Santana, G., Diehl, F., & De Menezes, J. (2003). Analysis of hazards associated with sea bathing: results of five years' work on oceanic beaches of Santa Catarina state, southern Brazil. *Journal of Coastal Research*, 107–116.
- [9] Lushine, J. (1991). A study of rip current drownings and related weather factors. *National weather digest*, 16(3), 13–19.
- [10] Silva, A., Mori, I., Dusek, G., Davis, J., & Pang, A. (2021). Automated rip current detection with region-based convolutional neural networks. *Coastal Engineering*, 166, 103859.
- [11] Cheng, B., Misra, I., Schwing, A., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1290–1299).
- [12] Maryan, C., Hoque, M., Michael, C., Ioup, E., & Abdelguerfi, M. (2019). Machine learning applications in detecting rip channels from images. *Applied Soft Computing*, 78, 84–93.
- [13] Zhu, D., Qi, R., Hu, P., Su, Q., Qin, X., & Li, Z. (2022). YOLO-Rip: A modified lightweight network for rip current detection. *Frontiers in Marine Science*, 9, 930478.
- [14] Mori, I., et al. "Flow-based rip current detection and visualization," in *IEEE Access*, vol. 10, pp. 6483–6495, 2022.
- [15] Rashid, A., et al. "RipNet: A lightweight one-class deep neural network for the identification of rip currents." In *International Conference on Neural Information Processing*, 2020, pp. 172–179.
- [16] Rashid, A., et al., "RipDet: A fast and lightweight deep neural network for rip current detection," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–6.
- [17] Rashid, A., et al. "Reducing rip current drowning: An improved residual-based lightweight deep architecture for rip detection," in *ISA transactions*, vol. 132, pp. 199–207, 2023.
- [18] Rampal, N. et al. "Interpretable deep learning applied to rip current detection and localization," in *Remote Sensing*, vol. 14, no. 23, pp. 6048, 2022.
- [19] Dumitriu, A. et al., "Rip current segmentation: A novel benchmark and YOLOv8 baseline results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1261–1271.
- [20] Dumitriu, A. et al., "AIM 2025 rip current segmentation (RipSeg) challenge report," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 5551–5560.
- [21] Dumitriu, A., et al, "Ripvis: Rip currents video instance segmentation benchmark for beach monitoring and safety," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3427–3437.
- [22] Dumitriu, A., et al. "NTIRE 2026 Rip Current Detection and Segmentation (RipDetSeg) Challenge Report." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 1784–1795.
- [23] He, K., et al., "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.
- [24] Kukartsev, V., et al. "Deep learning for object detection in images: development and evaluation of the YOLOv8 model using Ultralytics and Roboflow libraries." In *the Computer Science Online Conference*, 2024, pp. 629–637.
- [25] R. Sapkota, M. Karkee. "Ultralytics YOLO evolution: An overview of YOLO26, YOLO11, YOLOv8, and YOLOv5 object detectors for computer vision and pattern recognition," in *arXiv preprint arXiv:2510.09653*, 2025.
- [26] Lei, M., et al. "Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception," in *arXiv preprint arXiv:2506.17733*, 2025.
- [27] Kirillov, A. et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
- [28] Simeoni, O., et al. "Dinov3," in *arXiv preprint arXiv:2508.10104*. 2025.

- [29] Ravi, N., et al. "Sam 2: Segment anything in images and videos." In *International Conference on Learning Representations*, 2025, pp. 28085–28128.
- [30] Ravi, N., et al. "Sam 2: Segment anything in images and videos." In *International Conference on Learning Representations*, 2025, pp. 28085–28128.
- [31] Lin, T., et al. "Microsoft Coco: Common objects in context." In *European Conference on Computer Vision*, 2014, pp. 740–755.
- [32] Kukartsev, V., et al. "Deep learning for object detection in images: development and evaluation of the YOLOv8 model using Ultralytics and Roboflow libraries." in *Computer Science Online Conference*, 2024, pp. 629–637.
- [33] I. Loshchilov, F. Hutter. "Decoupled weight decay regularization," in *arXiv preprint arXiv:1711.05101*, 2017.
- [34] D. Kingma, J. Ba. "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.
- [35] Cubuk, E., et al. "RandAugment: Practical automated data augmentation with a reduced search space." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.
- [36] R. Solovyev, W. Wang, T. Gabruseva. "Weighted boxes fusion: Ensembling boxes from different object detection models," in *Image and Vision Computing*, vol. 107, pp. 104117, 2021.
- [37] Khan, F., et al. "RipFinder: real-time rip current detection on mobile devices," in *Frontiers in Marine Science*, vol. 12, pp. 1549513, 2025.
- [38] Khan, F., et al. "RipScout: Realtime ML-assisted rip current detection and automated data collection using UAVs," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 7742–7755, 2025.
- [39] Wan, M., et al. "Ripalert: A future-frame-aware framework for rip current forecasting and early alerting." " In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, pp. 39368–39377.
- [40] Ali, S., et al. "IBN-Driven Rip Current Analysis Using UAVs for Next-Generation Coastal Surveillance," in *IEEE Internet of Things Journal*, 2025.
- [41] Catala-Gonell, A., et al. "Oriented Object Detection for Complex Hydrodynamic Features: A Multi-Platform Rip Current Identification System," in *EGU sphere*, vol. 2026, pp. 1–29, 2026.
- [42] J. Choi, M. Rajendran, Y. Suh. "Box2Rip: Instance Segmentation of Amorphous Rip Currents via Box-Supervised Learning," in *IEEE Access*, 2025.
- [43] Liu, Y., et al. "A Deep Learning-Based Pipeline for Detecting Rip Currents from Satellite Imagery," in *Remote Sensing*, vol. 18, no. 2, pp. 368, 2026.
- [44] A. Sun, K. Yang. "Rip Current Detection in Nearshore Areas through UAV Video Analysis with Almost Local-Isometric Embedding Techniques on Sphere," in *arXiv preprint arXiv:2304.11783*, 2023.
- [45] Zhou, B., et al. "Scene parsing through ADE20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.
- [46] Xie, E., et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.
- [47] Xu, Z., S. Escalera, A. Pavão, M. Richard, W. Tu, Q. Yao, H. Zhao, and I. Guyon. *Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns* 3 (7), 100543 (2022). 2022.
- [48] McGill, Sean P., and Jean T. Ellis. "Rip current and channel detection using surfcams and optical flow." *Shore & Beach* 90.1 (2022): 50.
- [49] Philip, Shweta, and Alex Pang. "Detecting and Visualizing Rip Current Using Optical Flow." In *EuroVis (Short Papers)*, pp. 19-23. 2016.
- [50] de Silva, Akila, et al. "RipViz: Finding rip currents by learning pathline behavior." "IEEE Transactions on Visualization and Computer Graphics 30.7 (2023): 3930-3944.