Unveiling Submarine Cable Paths: A Self-Supervised Contrastive Learning Approach

Riya Ponraj[‡], Alagappan Ramanathan[*], Sangeetha Abdu Jyothi[*],
Amreesh Phokeer[Δ], Yu Wang[‡] and Ramakrishnan Durairajan[‡†]
[‡] University of Oregon, [Δ] Internet Society, [*] UC Irvine, [†] Link Oregon

Overview. Our digital lives depend critically on the seamless interplay between the logical (traffic) and physical (infrastructure) layers of submarine cable networks (SCNs), which carry over 99% of intercontinental traffic across more than 1.8 million km of fiber infrastructure [15]. Understanding the interplay between these layers is essential for analyzing global routing dependencies and over-reliance, optimizing performance, enhancing security, and improving resilience [6, 11, 13].

Challenges. Despite their importance, accurately mapping traffic to physical infrastructure presents significant challenges due to noise at three distinct levels. At the infrastructure level, SCNs exhibit diverse architectural designs, including linear, trunk-and-branch, and festoon topologies. These structural differences render common approximations (e.g., Haversine or great-circle distance calculations) between cable landing stations (CLS) unreliable. Effective cartography must account for this heterogeneity to avoid oversimplifications that obscure real-world complexity. At the traffic level, modern routing mechanisms such as MPLS-based virtualization, multi-cable routing, and dynamic load balancing obfuscate the actual physical paths taken by packets [8]. Compounding these issues is data-level noise: public maps (e.g., from Telegeography) omit critical details such as architectural heterogeneity, while measurement platforms like CAIDA's Ark or RIPE Atlas may exhibit bias due to vantage point placement.

Limitations of State-of-the-Art. Prior efforts (e.g., [12, 13]) have made significant progress in mapping traceroutes to submarine cables. While they are as compelling as ever, like other physical-to-logical mapping efforts (e.g., [5, 9, 10]), they predominantly rely on *heuristics*. More importantly, we note that these methods may not fully capture the real-world complexity and operational pragmatics of SCNs and do not comprehensively handle the above multiple levels of noise. What is critically lacking is a principled and data-driven framework that can address these issues holistically.

Approach & Insights. We introduce **Oceanus** (see Figure 1), a self-supervised contrastive learning framework for SCN cartography that moves beyond heuristic techniques by addressing the aforementioned challenges. At its core, Oceanus is grounded in three key insights.

(1) To capture the topological uniqueness and focus on learning structurally consistent patterns that reflect the diversity and complexity of actual SCN deployments, Oceanus framework first filters out physically impossible paths. To this end, we enrich cable topology information with architectural layouts, such as trunk-and-branch or festoon configurations. We then use the published cable lengths to calculate the Speed-of-Light (SoL) latency, allowing us to filter out paths with Round-Trip-Time (RTT) difference exceeding physically plausible thresholds (i.e., the SoL latency) [3]. These filtering steps enable Oceanus to focus on meaningful traceroutes that carry valid RTT signals, providing cleaner temporal supervision for the self-supervised learning training [4, 7] in the next stage.

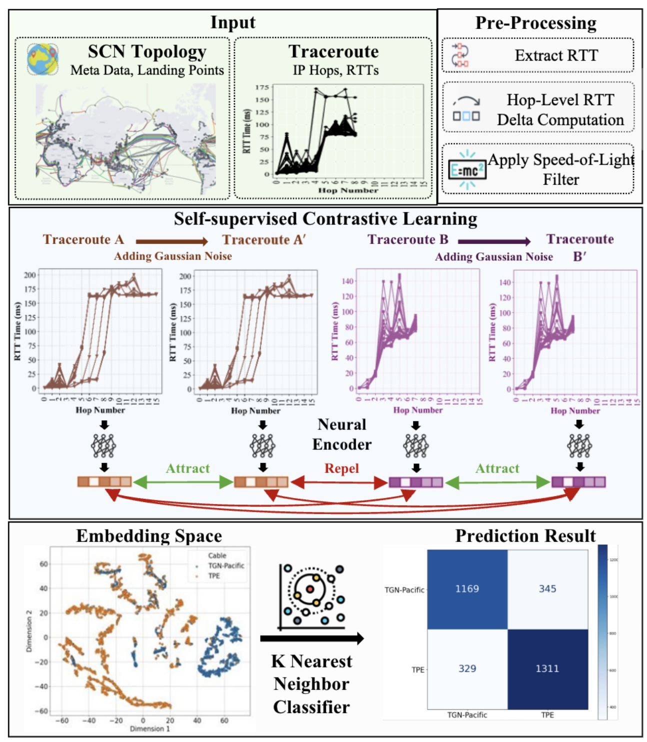

Figure 1: Oceanus Pipeline.

(2) After removing noisy signals, we model traceroutes as sequential time-series data and design a Multi-Layer Perceptron (MLP)-based neural network encoder [14] to extract temporal patterns from the RTT differences. The key intuition is that RTT variations observed along submarine cables reflect their properties, such as geolocation, physical characteristics, and length. These patterns can be captured as latent feature embeddings that are informative for downstream classification tasks. Our 4-layer MLP-based neural network encoder processes the input RTT sequence with

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

non-linear transformations and dropout regularization to capture temporal interactions among RTT difference measurements, ultimately generating the informative latent feature embedding as shown in Figure 1. Based on this latent feature embedding, we then apply a K-Nearest Neighbor (KNN) classifier to predict the corresponding cable label. This approach allows the model to remain robust in the face of routing variability introduced by techniques like load balancing and route virtualization.

(3) The labels between traceroutes and their corresponding physical cables are naturally scarce due to the opaque nature of routing policies, MPLS tunneling, and limited access to proprietary infrastructure data (e.g., given current operational practices). Furthermore, the RTT measurements (e.g., jitter and small fluctuations) are inherently unstable. Hence, we design a self-supervised contrastive learning strategy to train our previously designed 4-layer MLP-based neural network encoder in Oceanus, enhancing its generalization for limited traceroute-cable labeling mapping and robustness to noisy RTT measurements. Specifically, for each labeled traceroute, we generate augmented versions by adding slight Gaussian noise. The positive pair consists of a traceroute and its own augmented version, while the negative pair consists of a traceroute and an augmented version of a different traceroute. This self-supervised contrastive learning allows Oceanus to capture the inherent latency behavior, measurement biases, and physical characteristics of SCNs while removing those features caused by slight jitters, essentially considering the third level of noise.

Evaluation Setup. We evaluate Oceanus using datasets from Nautilus [12], which provides traceroutes probabilistically labeled with candidate submarine cables. Although noisy (mainly due to geolocation inaccuracies), these weak labels are sufficient for contrastive learning, which tolerates labeling imprecision by focusing on relative similarity. Each traceroute is processed to extract RTTs and compute differences between consecutive hops, revealing latency jumps typically indicative of submarine cable transitions.

For our preliminary evaluation, we focus on two trans-Pacific submarine cables with linear topologies: (1) Tata TGN-Pacific from Emi, Japan to Hillsboro, USA [1]; and (2) Trans-Pacific Express (TPE) from Maruyama, Japan to Nedonna Beach, USA [2].

To ensure physical plausibility, we apply SoL filtering to compare observed RTT differences against theoretical minimums as follows. For cables with exactly two landing points (e.g., Unity/EAC-Pacific), we use TeleGeography's cable length and assume signal propagation at two-thirds the speed of light in fiber. For cables with multiple landing points (e.g., Tata TGN-Pacific), we use the segment-unfolding method described in [13] to isolate the relevant path segment first and then do the above step for each segment.

We then train a 4-layer self-supervised MLP encoder within the SimCLR framework [4] to generate embeddings from RTT differences. Inputs consist of 2D vectors: submarine hop index and corresponding RTT differences. Each hidden layer is followed by ReLU activation, batch normalization, and dropout (rates: 0.2 for the first two layers, 0.1 for the third). The final layer outputs a 16-dimensional embedding per traceroute segment. Training uses the NT-Xent contrastive loss. Gaussian noise with standard deviation $\sigma = 0.2$ is applied for augmentation. The model is optimized using Adam with the learning rate 1×10^{-3}, batch size 64, over 30 epochs.

Preliminary Results. To evaluate the quality of the learned embedding space, we first assess the ability of the model to cluster traceroutes traversing the same cable. Using only 20% of the labeled data, we train a k-NN classifier to predict cable labels. We measure the performance using accuracy, precision, and recall on a held-out set, which are reported with respect to the weak cable labels from Nautilus. For qualitative analysis, we use t-SNE to visualize the embedding space and observe cluster formation. Oceanus achieves 78.63% accuracy, 78.60% precision, and 78.58% recall. These results indicate that Oceanus effectively learns cable-specific latency signatures, even with weak supervision. As the shown in the t-SNE visualization of Figure 1, traceroute segments for Tata TGN-Pacific and Trans-Pacific Express form coherent and distinct clusters.

To evaluate the effectiveness of Oceanus, we next analyze ~160,000 traceroutes labeled by Nautilus as traversing the Tata TGN-Pacific or TPE cables. Nautilus assigns 12,565 traceroutes uniquely to each cable and marks 134,870 as belonging to both, due to shared landing points and heuristic limitations. In contrast, Oceanus assigns 30,369 traceroutes to Tata TGN-Pacific and 129,631 to TPE, eliminating overlap entirely. This represents a significant increase in confidently classified routes using Oceanus (30,369 > 12,565 for Tata TGN-Pacific, and 129,631 > 12,565 for TPE), resolving ambiguity inherent in heuristic methods.

Although Oceanus achieves 80% classification accuracy, this is based on a filtered dataset of ~25,000 less ambiguous examples (down from 160,000 traceroutes), selected to improve signal clarity during training. The model currently uses only two RTT-based features, yet still outperforms the baseline. Future work will incorporate more features and supervision to further improve accuracy.

Acknowledgements. We thank Internet Society Foundation and NSF CNS-2145813 for funding this project.

REFERENCES

[1] 2002. TGN-Pacific (formerly VSNL Transpacific). Submarine cable system connecting Emi, Japan to Hillsboro, USA.
[2] 2008. Trans-Pacific Express (TPE) Cable System. Submarine cable linking Maruyama, Japan to Nedonna Beach, USA.
[3] AFRINIC. 2021. Unintended Consequences of Submarine Cable Deployment on Internet Routing. https://afrinic.net/20210330-unintended-consequences-of-submarine-cable-deployment-on-internet-routing/amp. Accessed: 2025-07-09.
[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.
[5] Anderson et al. 2022. iGDB: connecting the physical and logical layers of the internet. In *ACM IMC*.
[6] Bischof et al. 2018. Untangling the world-wide mesh of undersea cables. In *ACM HotNets*. 78–84.
[7] Chakraborty et al. 2020. G-SimCLR: Self-supervised contrastive learning with guided projection via pseudo labelling. In *ICDMW*. IEEE.
[8] Carisimo et al. 2023. A hop away from everywhere: A view of the intercontinental long-haul infrastructure. *ACM SIGMETRICS* (2023).
[9] Durairajan et al. 2013. Internet Atlas: A Geographic Database of the Internet. In *ACM HotPlanet Workshop*.
[10] Durairajan et al. 2015. InterTubes: A Study of the US Long-haul Fiber-optic Infrastructure. In *ACM SIGCOMM Conference*.
[11] Liu et al. 2020. Out of sight, not out of mind: A user-view on the criticality of the submarine cable network. In *ACM IMC*.
[12] Ramanathan et al. 2023. Nautilus: A framework for cross-layer cartography of submarine cables and ip links. *ACM SIGMETRICS* (2023).
[13] Wang et al. 2025. Threading the Ocean: Mapping Digital Routes Across Submarine Cables using Calypso. In *ACM SIGCOMM*.
[14] Martin Riedmiller and A Lernen. 2014. Multi layer perceptron. *Machine learning lab special lecture, University of Freiburg* 24 (2014), 11–60.
[15] TeleGeography. 2023. Submarine Cable Map. https://www.submarinecablemap.com/