Dataset	Number of Train Scenes	Number of Train Episodes	Number of Val Scenes	Number of Val Episodes
IndoorEnv	990	898267	905	99
MP3D	61	5000000	495	11
Gibson	72	4932479	1000	16

Table A1. **Dataset Statistics**

Appendix A. Dataset Details

We constructed the Point-Nav datasets for each of IndoorEnv, MP3D, and Gibson environments using a sampling-based method which filtered out easy episodes (those with $\frac{\text{euclidean distance}}{\text{geodesic distance}} < 1.1$). We additionally filter out episodes where there is no path between the start and goal location. The start points from these episodes were additionally used for the Exploration and Flee tasks, but the goal locations were ignored. Per-environment statistics are listen in Table A1.

Appendix B. Cumulative Performance based on Starting Distance

To examine the balance of starting distances, Figure F1 shows a breakdown of performance on all episodes which start closer than N meters. This additionally shows that 50% of the episodes start more than 9 meters away and nearly 25% start greater than 15 meters away.

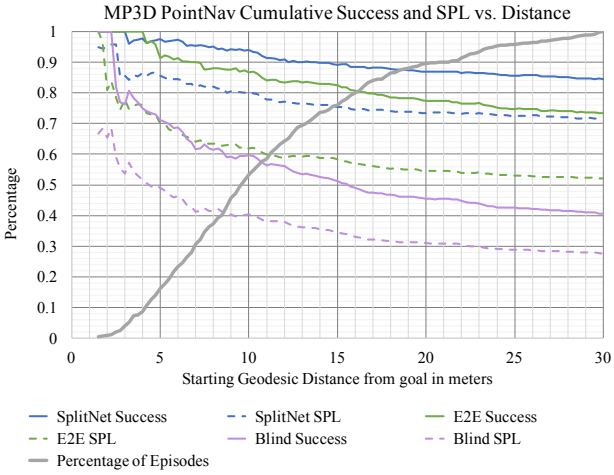

Figure F1. Cumulative Accuracy and SPL on MP3D dataset.

Appendix C. Network Architecture

Figure F2 shows the encoder-decoder architecture of SplitNet. The E2E method trains the blue and orange portions, and the blind agent trains only the orange. Additionally the Motion layers are only trained for SplitNet. Those are omitted for simplicity due to them operating on multiple timesteps. The Egomotion

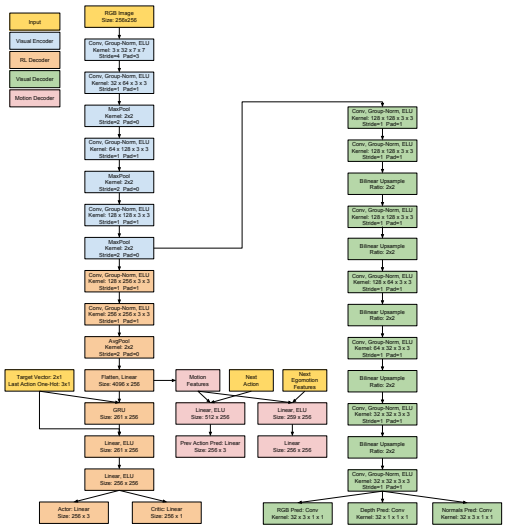

Figure F2. **SplitNet Architecture**

Appendix D. Auxiliary Outputs

To verify that our network learns to encode geometry and appearance information, we show the output of the RGB, depth, and surface normal decoders on test environments in Figure F3. Learning these encoder-decoders, especially for depth, improves the network's ability to codify visual information into actionable representations. The decoders also allow us to see where the network makes mistakes as a method of debugging the failure modes.

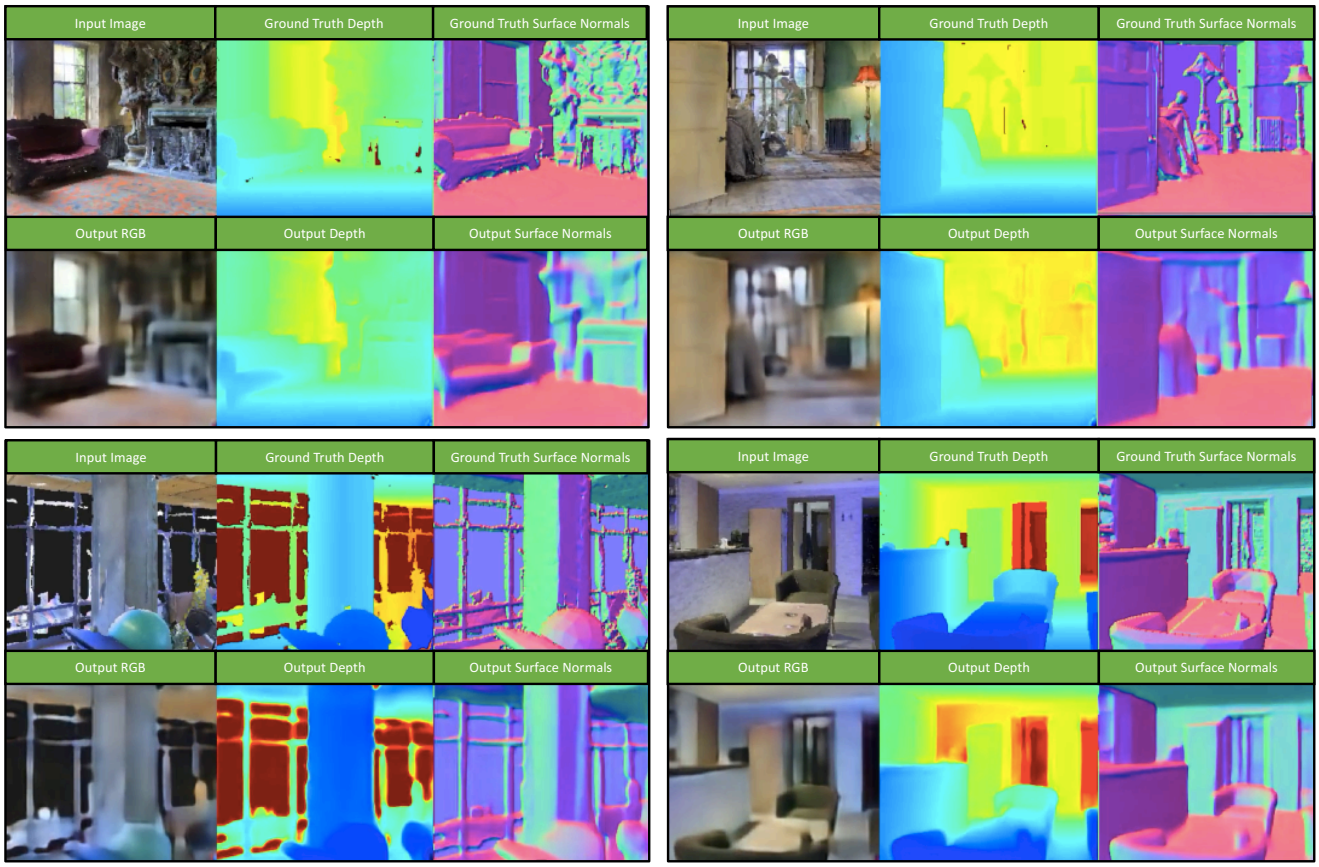

Figure F3. Example predictions of auxiliary outputs on unseen MP3D test environments.