Supplementary Material

1. NETWORK ARCHITECTURE

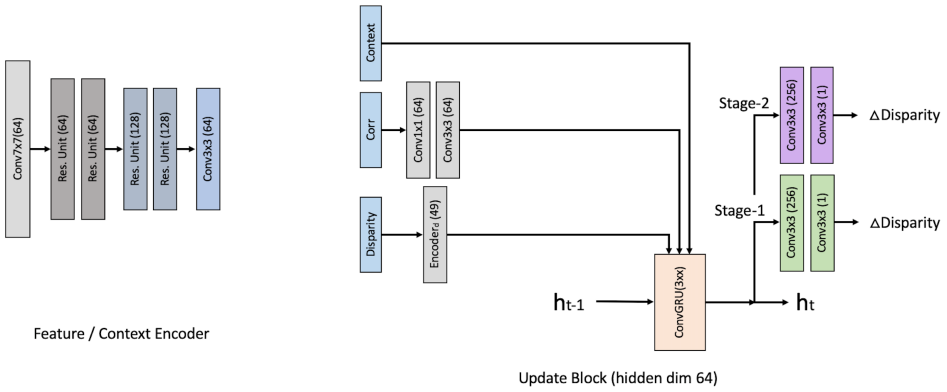

Fig. S1. Network architecture details. The context and feature encoders have the same architecture, the only difference is that the feature encoder uses instance normalization while the context encoder uses batch normalization.

2. DISCUSSION ON COMPUTATIONAL COST

Fig. S2 compares our method with others on reconstruction quality versus computational cost on the test set of Tanks-and-Temples. We plot two versions of our method (the proposed version in the main paper and a lightweight one with lower input resolution and fewer neighbor views). We see that our method has comparable or better quality-cost tradeoffs than prior works on the intermediate set of Tanks-and-Temples, but is significantly more memory-efficient on the advanced set of Tanks-and-Temples.

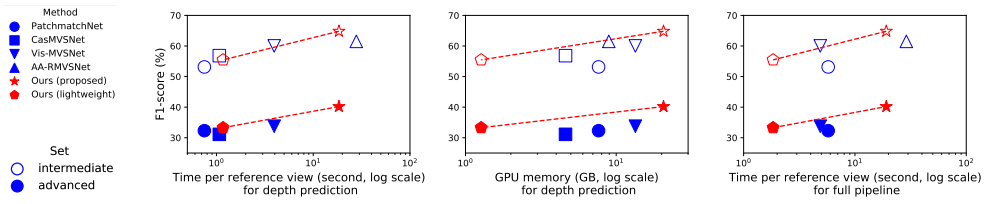

Fig. S2. Reconstruction quality versus computational cost. Time is divided by the number of reference views following convention. All methods use the same set of reference views. Memory is measured by **torch.cuda.max_memory_allocated**, which returns the peak allocated memory. All results are measured on an A6000 GPU. We are unable to measure CasMVSNet for subfigure (c) and EPP-MVSNet because certain technical difficulties in installing dependencies and running certain parts of their code on our hardware could not be resolved yet.