SEA-Flow3D: Simplified, Efficient, and Accurate Scene Flow via Spatial Vector Sampling and Multi-scale Refinement

Supplementary Material

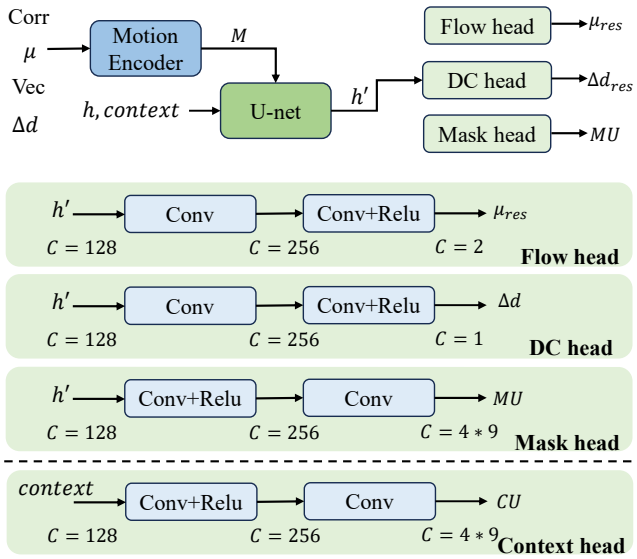

Figure 7. **Optimization pipeline.** Motion features (Corr, μ, Vec, Δd) are first processed by the decoupled motion encoder and then fused with the hidden state h and context features ($\mathbf{C}$) before being fed into the RNN optimizer (U-Net). The optimizer outputs the updated hidden state h', from which three lightweight convolution heads generate the required upsampling masks and residual updates. Here, C denotes the number of feature channels.

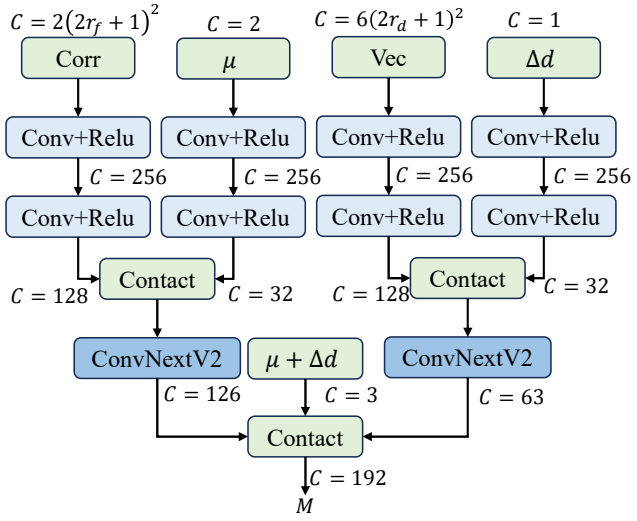

Figure 8. **Architecture of Split Motion Encoder.**

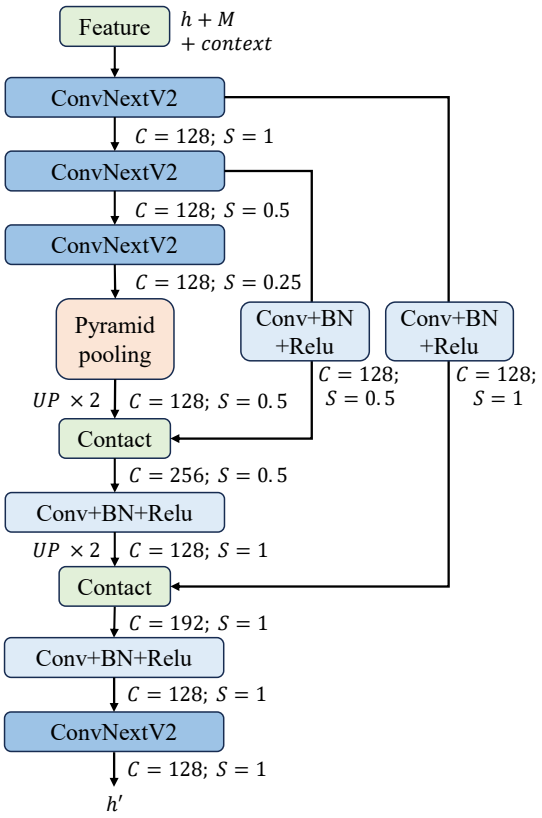

Figure 9. **Structure of the U-Net-style optimizer.** S denotes the feature-scale resolution, and $UP \times 2$ indicates bilinear upsampling by a factor of 2.

5. Network Architecture Details.

Figs. 7, 8 and 9 provide the complete structural specification of our optimization network. In the U-Net–style optimizer, we incorporate a pyramid pooling module [37] to aggregate global contextual information through multi-scale average pooling.