MSTA-based Feature Extractor
Masked Spatial-Temporal Attention
× L
Layer Norm
Masked Temporal Attn
Layer Norm
Masked Spatial Attn
Layer Norm
Feed-Forward Network
Positional Encoding
Embedding Tokens
Spatial-Temporal Correlation
Tokens at each time
Masked Self Attention
Temporal Correlation
Tokens at each position
Masked Self Attention
Spatial Correlation
Target Patch
Temporal Auxiliary
Spatial Auxiliary
Masked Patch
Output of Target