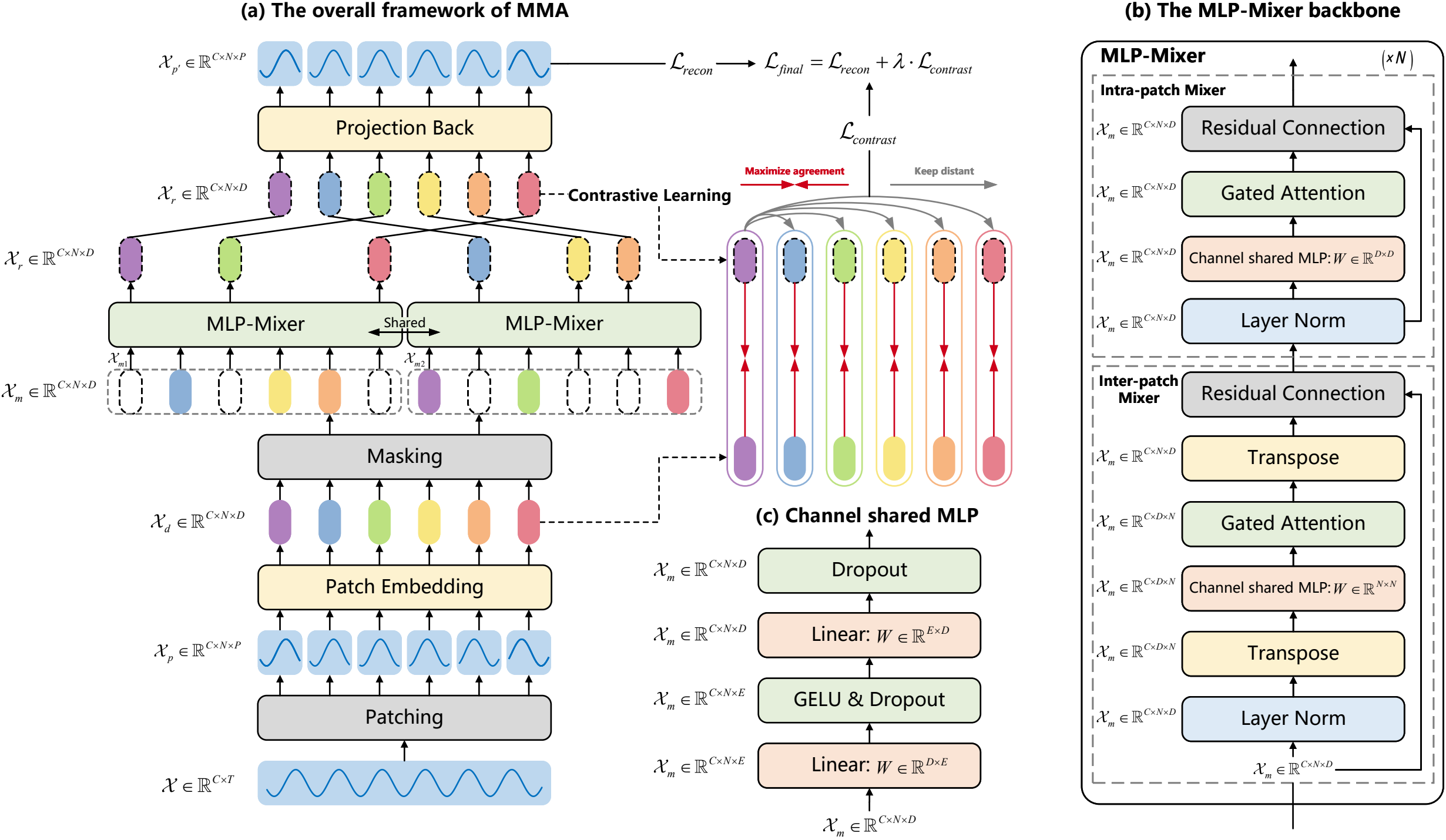
(a) The overall framework of MMA
$\mathcal{X}_{p'} \in \mathbb{R}^{C\times N\times P}$
$\mathcal{L}_{recon}$
$\mathcal{L}_{final} = \mathcal{L}_{recon} + \lambda \cdot \mathcal{L}_{contrast}$
Projection Back
$\mathcal{X}_r \in \mathbb{R}^{C\times N\times D}$
$\mathcal{L}_{contrast}$
Contrastive Learning
Maximize agreement
Keep distant
$\mathcal{X}_r \in \mathbb{R}^{C\times N\times D}$
MLP-Mixer
Shared
MLP-Mixer
$\mathcal{X}_{m1}$
$\mathcal{X}_{m2}$
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Masking
$\mathcal{X}_d \in \mathbb{R}^{C\times N\times D}$
Patch Embedding
$\mathcal{X}_p \in \mathbb{R}^{C\times N\times P}$
Patching
$\mathcal{X} \in \mathbb{R}^{C\times T}$
(c) Channel shared MLP
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Dropout
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Linear: $W \in \mathbb{R}^{E\times D}$
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times E}$
GELU & Dropout
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times E}$
Linear: $W \in \mathbb{R}^{D\times E}$
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
(b) The MLP-Mixer backbone
MLP-Mixer
$(\times N)$
Intra-patch Mixer
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Residual Connection
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Gated Attention
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Channel shared MLP: $W \in \mathbb{R}^{D\times D}$
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Layer Norm
Inter-patch Mixer
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Residual Connection
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Transpose
$\mathcal{X}_m \in \mathbb{R}^{C\times D\times N}$
Gated Attention
$\mathcal{X}_m \in \mathbb{R}^{C\times D\times N}$
Channel shared MLP: $W \in \mathbb{R}^{N\times N}$
$\mathcal{X}_m \in \mathbb{R}^{C\times D\times N}$
Transpose
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$
Layer Norm
$\mathcal{X}_m \in \mathbb{R}^{C\times N\times D}$