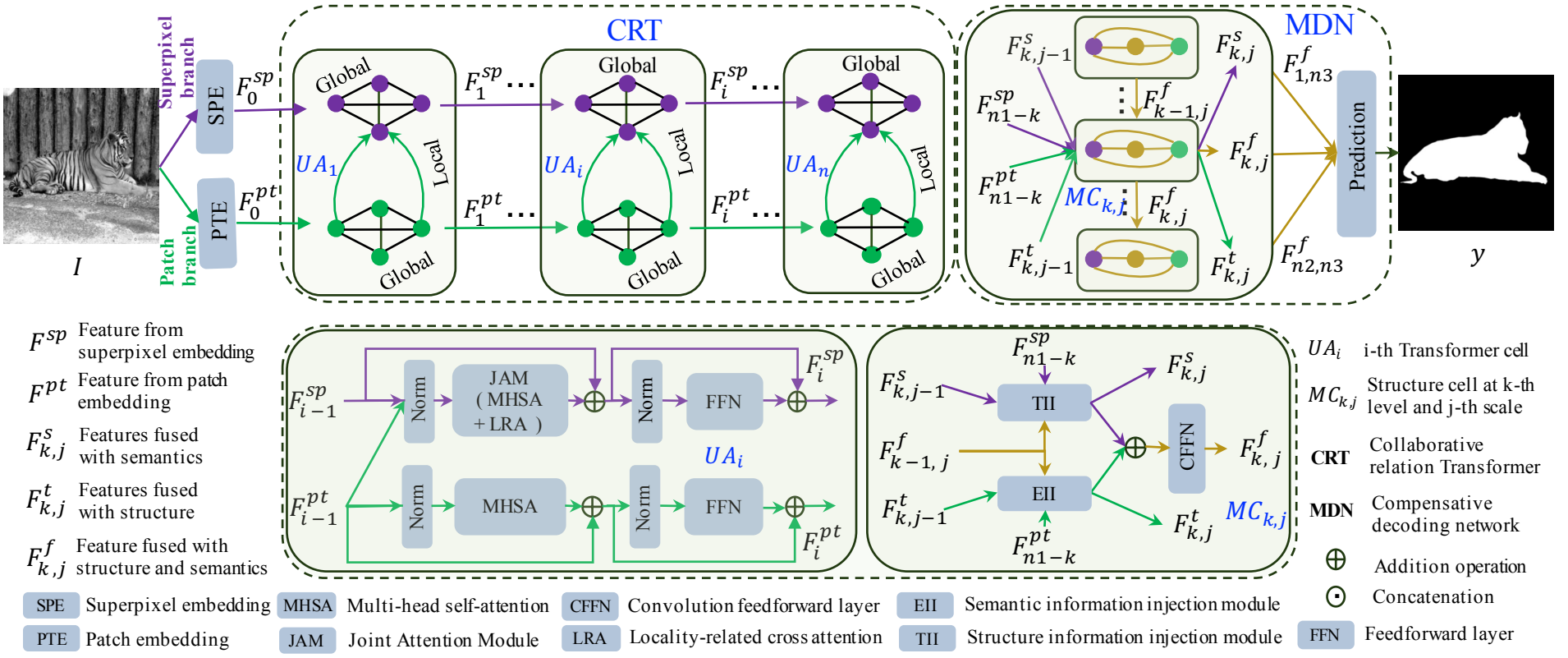
Superpixel branch
Patch branch
SPE
PTE
CRT
Global
Local
MDN
Prediction
JAM (MHSA + LRA)
MHSA
Norm
FFN
TII
EII
CFFN
I
y
F^{sp} Feature from superpixel embedding
F^{pt} Feature from patch embedding
$F^{s}_{k,j}$ Features fused with semantics
$F^{t}_{k,j}$ Features fused with structure
$F^{f}_{k,j}$ Feature fused with structure and semantics
UA_i i-th Transformer cell
$MC_{k,j}$ Structure cell at k-th level and j-th scale
CRT Collaborative relation Transformer
MDN Compensative decoding network
Addition operation
Concatenation
SPE Superpixel embedding
PTE Patch embedding
MHSA Multi-head self-attention
JAM Joint Attention Module
CFFN Convolution feedforward layer
LRA Locality-related cross attention
EII Semantic information injection module
TII Structure information injection module
FFN Feedforward layer