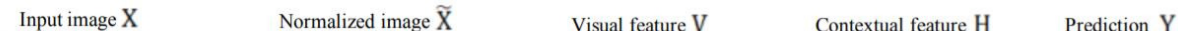
Input image X
Trans.
Normalized image X̃
Feat.
Visual feature V
Seq.
Contextual feature H
Pred.
Prediction Y
UNITED