BENCHMARK OF SELF-SUPERVISED MODELS FOR VIDEO UNDERSTANDING ON THE CATER DATASET

THÉO MOUTAKANNI & ETIENNE BOISSEAU

January 25, 2021

INTRODUCTION

Figure 1: Examples of scene bias (Choi et al., 2019).

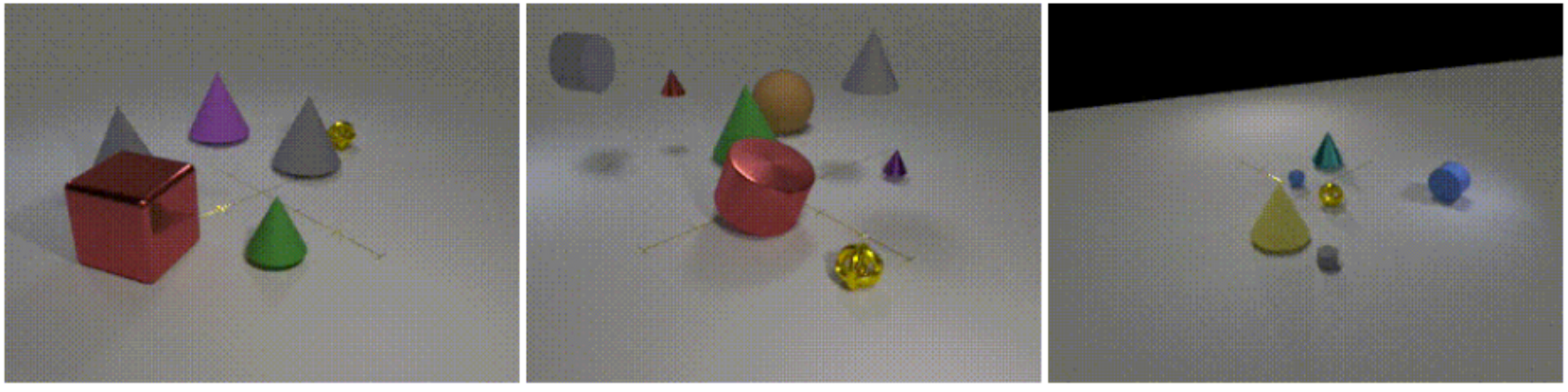

Figure 2: The CATER Dataset

PROBLEM DEFINITION

Our goal: provide a benchmark of self-supervised learning methods for videos on a dataset which does not suffer from any scene bias.

Selected methods:

1. Self-supervised spatiotemporal learning via video clip order prediction (Xu et al., 2019)
2. Memory-augmented dense predictive coding for video representation learning (Han, Xie, and Zisserman, 2020)
3. Learning correspondence from the cycle-consistency of time (Wang, Jabri, and Efros, 2019)

SELF-SUPERVISED SPATIOTEMPORAL LEARNING VIA VIDEO CLIP ORDER PREDICTION (XU2019SELF)

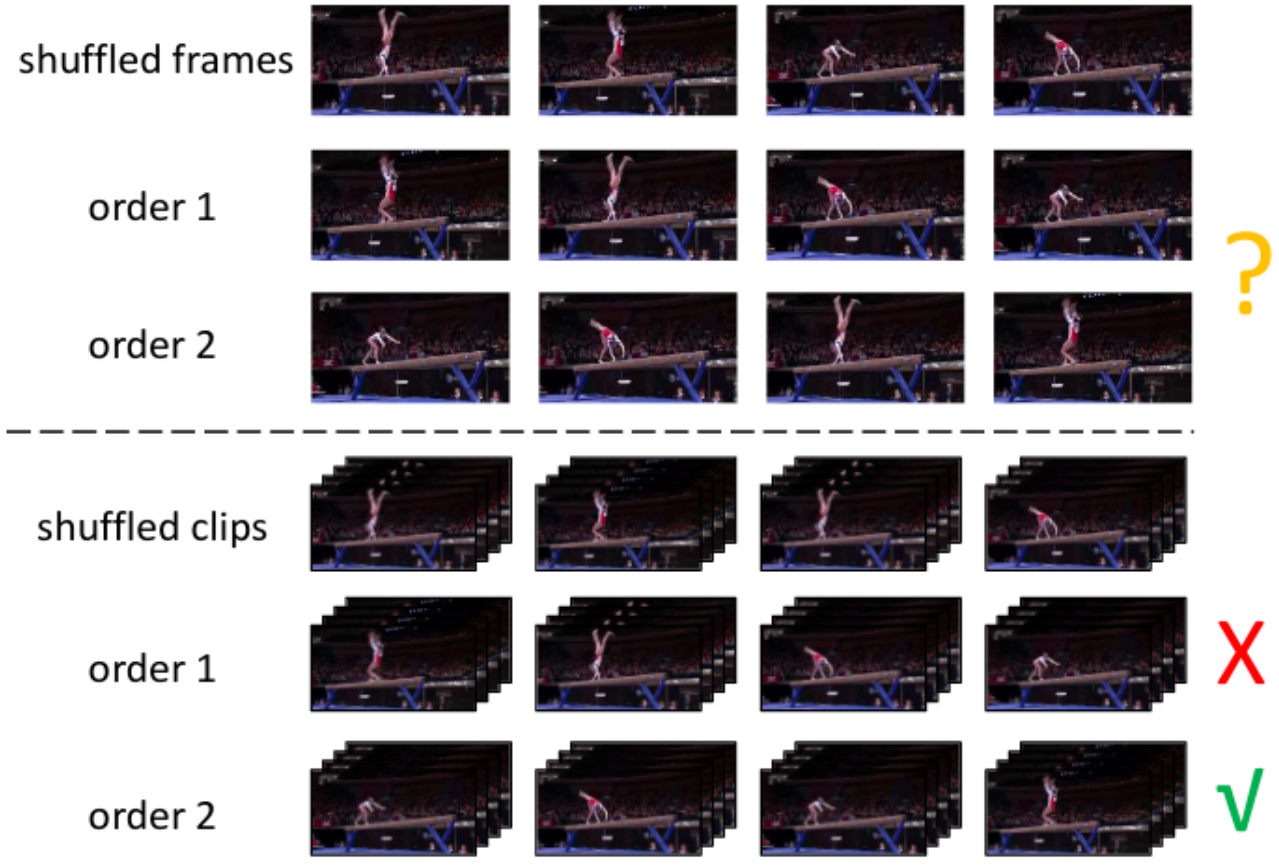

Figure 3: The model learns to sort video clips instead of single frames.

SELF-SUPERVISED SPATIOTEMPORAL LEARNING VIA VIDEO CLIP ORDER PREDICTION (XU2019SELF)

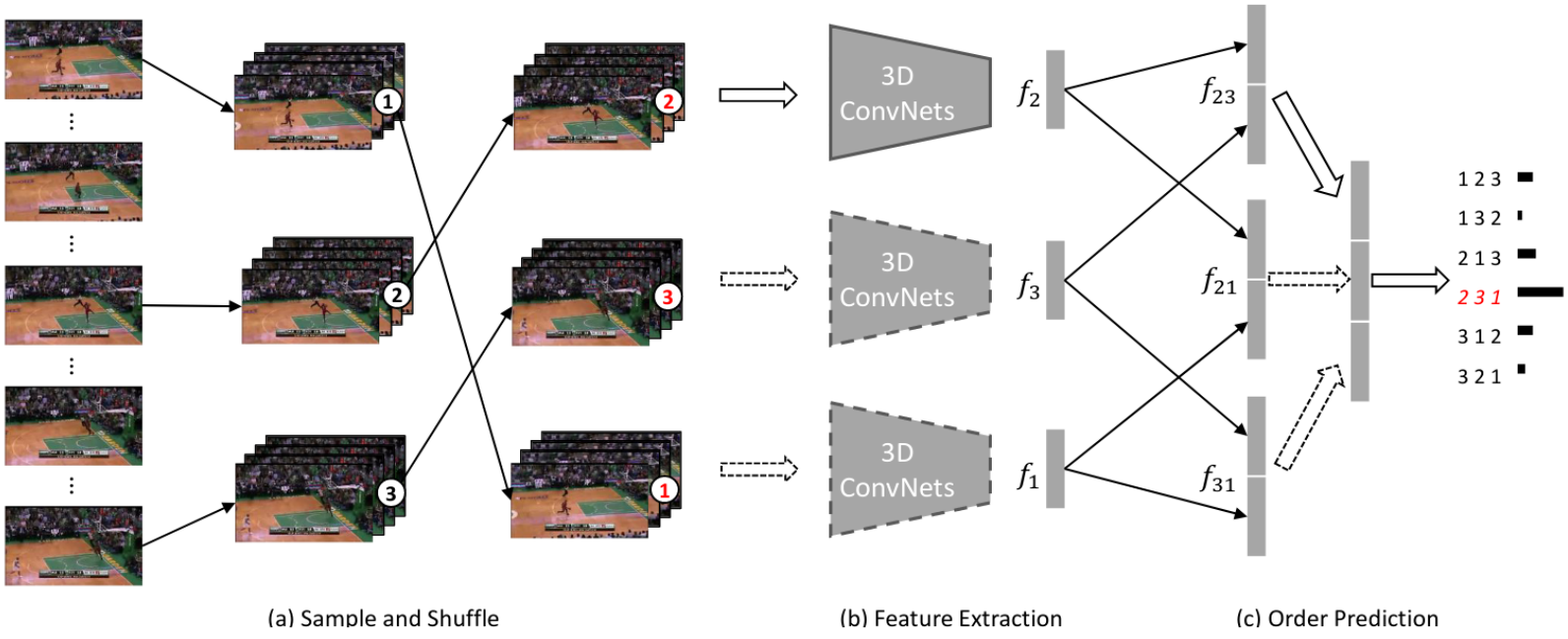

Figure 4: Model architecture

MEMORY-AUGMENTED DENSE PREDICTIVE CODING FOR VIDEO REPRESENTATION LEARNING (HAN20)

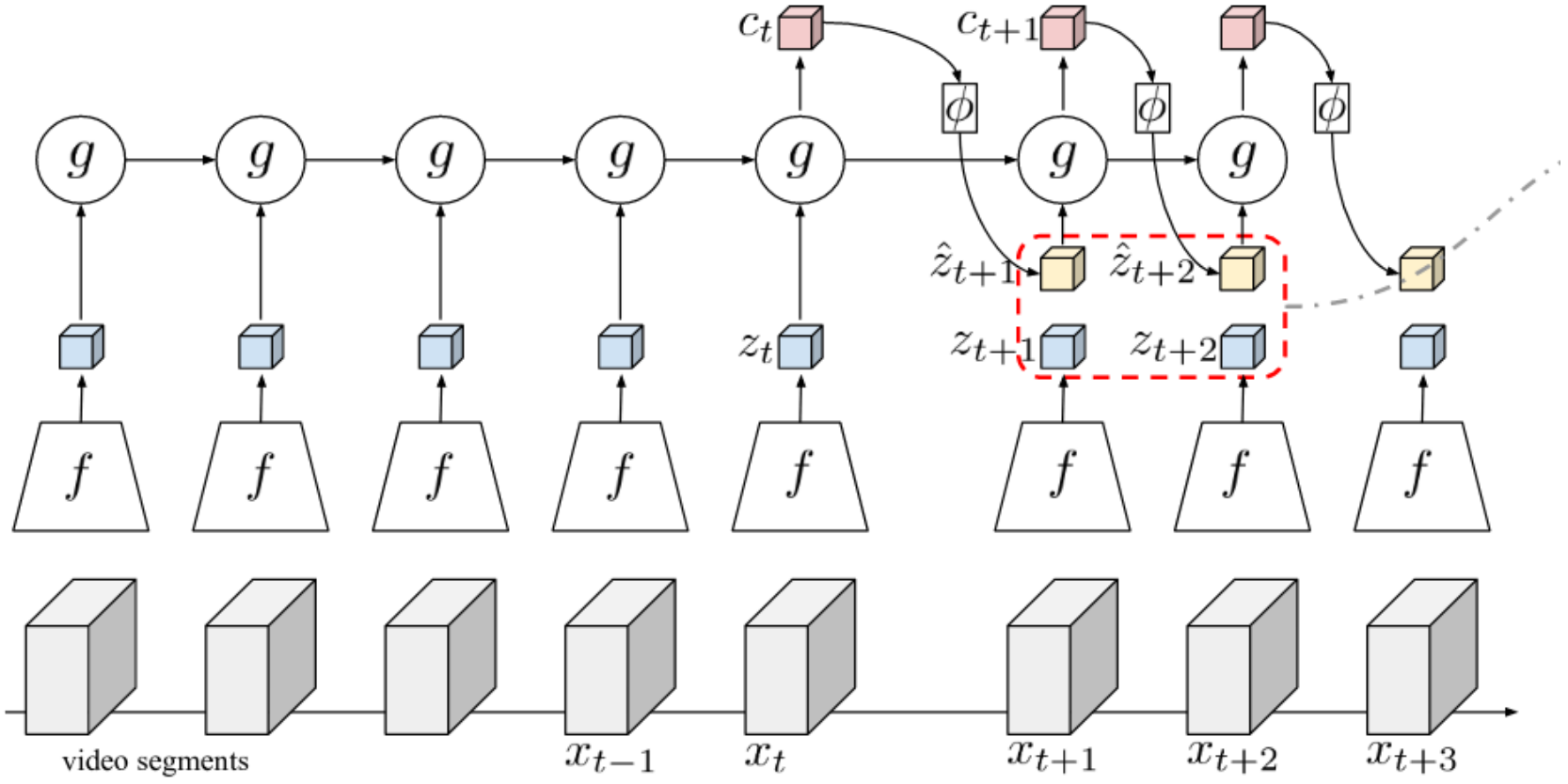

Figure 5: Original DPC architecture

MEMORY-AUGMENTED DENSE PREDICTIVE CODING FOR VIDEO REPRESENTATION LEARNING (HAN20)

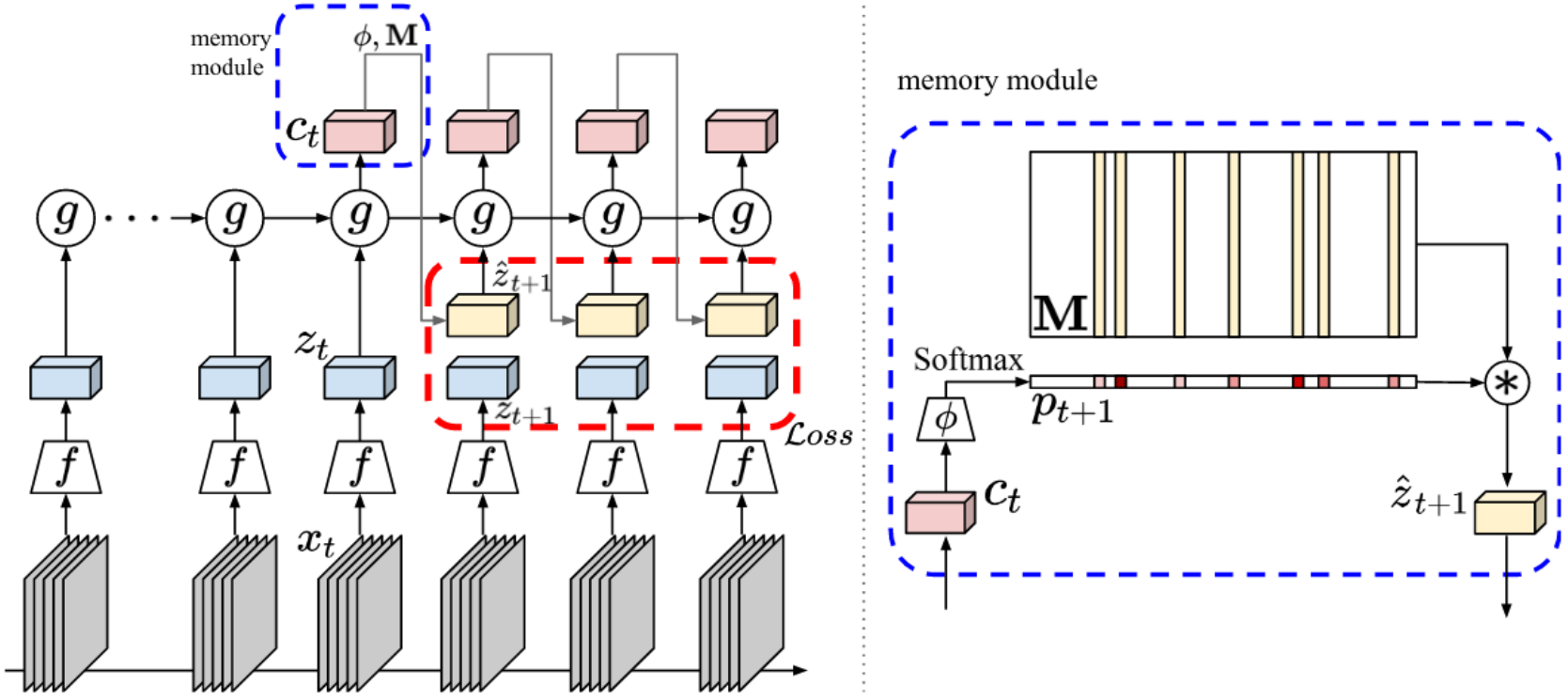

Figure 6: MemDPC architecture, including a memory module

LEARNING CORRESPONDENCE FROM THE CYCLE-CONSISTENCY OF TIME (CVPR2019˙CYCLETIME)

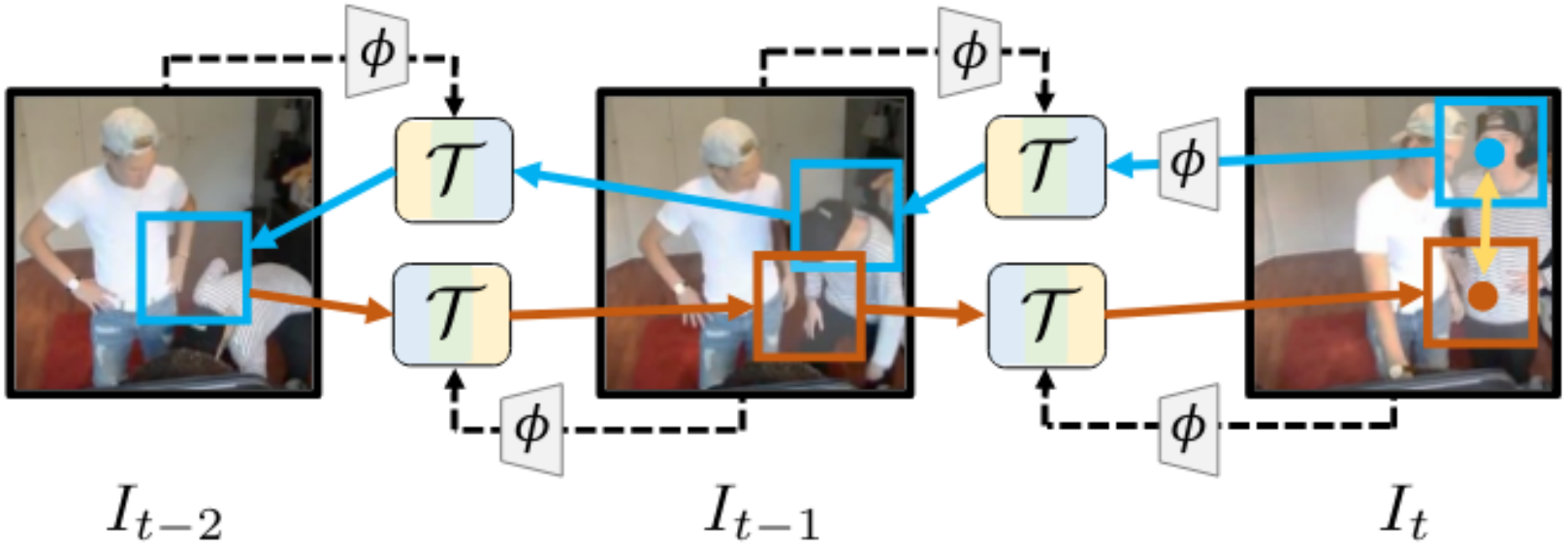

Figure 7: Cycle-consistency model architecture

LEARNING CORRESPONDENCE FROM THE CYCLE-CONSISTENCY OF TIME (CVPR2019˙CYCLETIME)

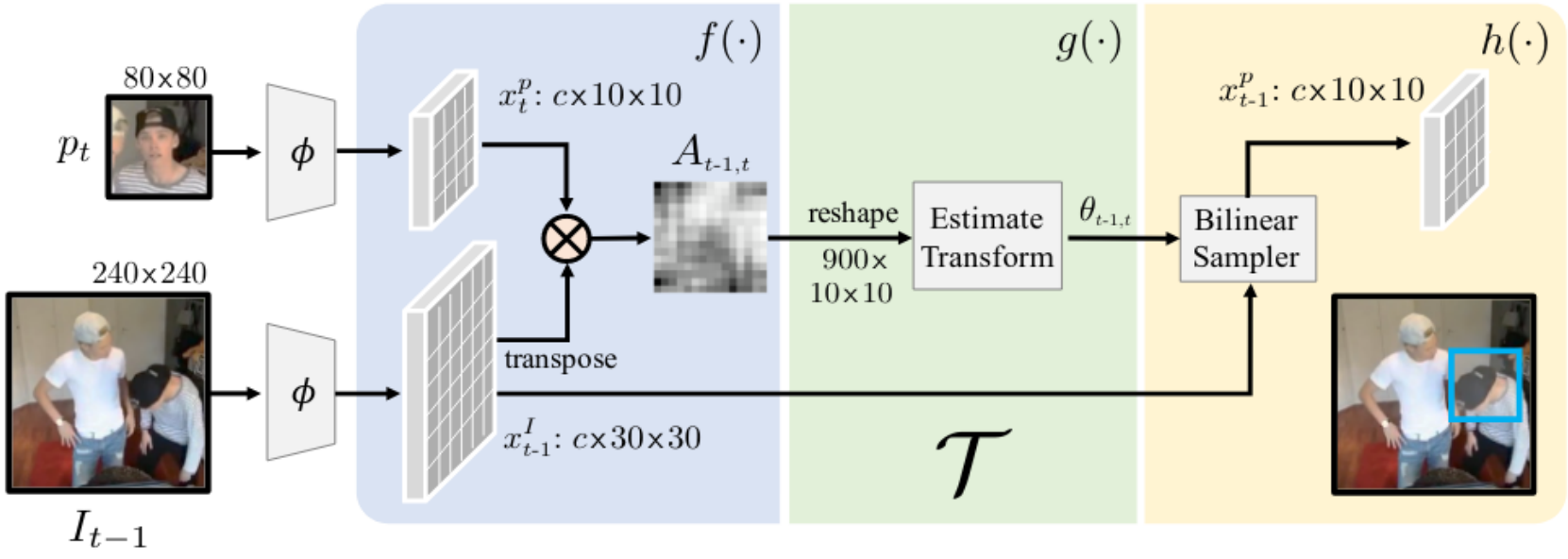

Figure 8: Tracker function architecture

RESULTS

- Method 1 (VCOP): Could not learn
- Method 2 (MemDPC): 68% mAP on action recognition, 38% on Compositional action recognition
- Method 3 (Cycle-consistency): 67% mAP on action recognition, 40% on Compositional action recognition

CONCLUSION AND FURTHER WORK

- Method 1 (VCOP): More experiments are needed, including more epochs for a fair comparison.
- Method 2 (MemDPC) and Method 3 (Cycle-consistency): More hyperparameters and computation time could be given to the models, as well as more complicated classification heads (MLPs, LSTMs)