Consistency-Aware Graph Network for Human Interaction Understanding (Supplementary Material)

Zhenhua Wang[†], Jiajun Meng[†], Dongyan Guo[†], Jianhua Zhang[♯], Javen Qinfeng Shi[‡], Shengyong Chen[♯]
[†]Zhejiang University of Technology, [♯]Tianjin University of Technology, [‡]The University of Adelaide

1. Layer-wise Input and Output Shapes of TOGN

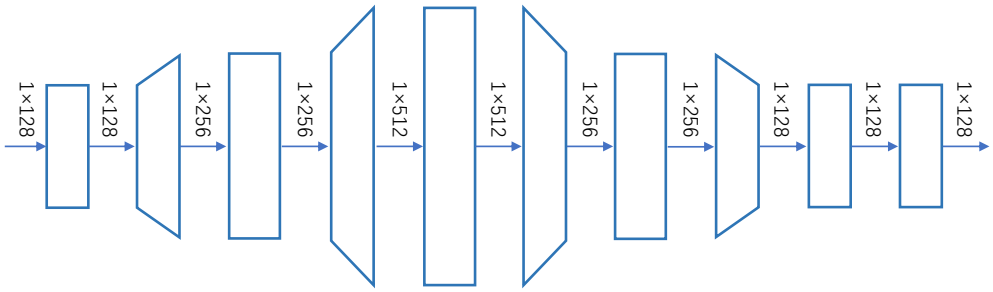


Figure 1. Input and output shapes of a 10-layer TOGN.

There are 10 layers in total in our proposed third-order graph network (TOGN). We give the shapes of inputs and outputs for each TOGN layer in Figure 1. We would like to note that in each layer, the factor feature and the node feature share the identical shape, *i.e.* $\mathbf{f}^l \in \mathbb{R}^{D_l}$ and $\mathbf{g}^l \in \mathbb{R}^{D_l}$, where $\mathbf{f}^l$ denotes the node feature, $\mathbf{g}^l$ denotes the factor feature, and D_l denote the length of the input feature vector in the l-th layer. Shapes of edge features over different layers are fixed being 1×16.

2. Visualization of Learned λ^c

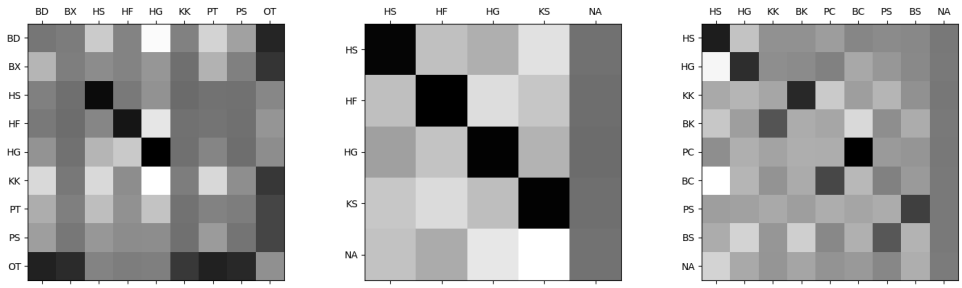
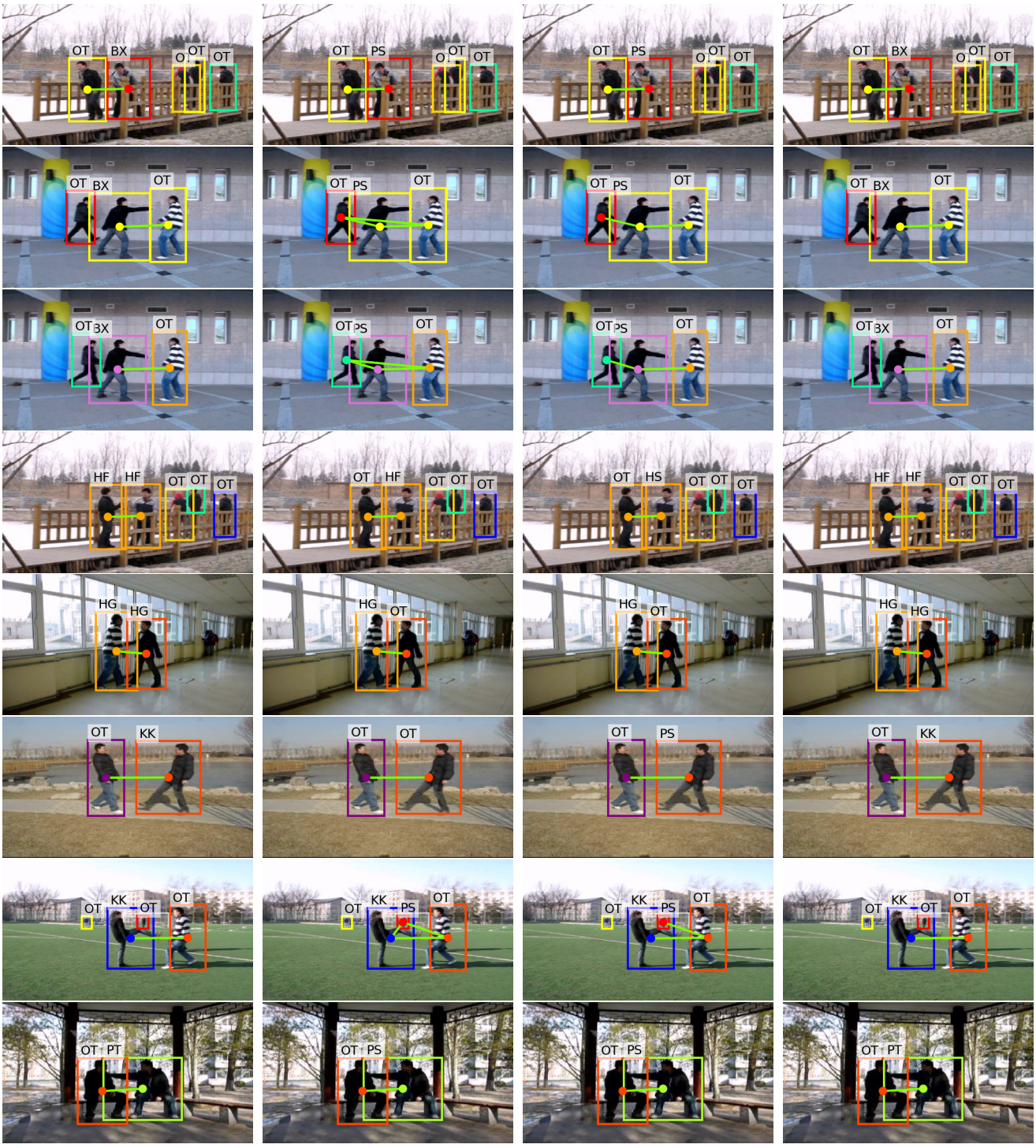


Figure 2. Visualization of λ^C as a matrix on BIT, TVHI, UT. Darker cells correspond to lower penalties. BD: bend, BX: box, HS: handshake, HF: high-five, HG: hug, KK: kick, PT: point, PS: push, OT: others, KS: kiss, BK: be-kicked, PC: punch, BC: be-punched, BS: be-pushed, NA: no-action.

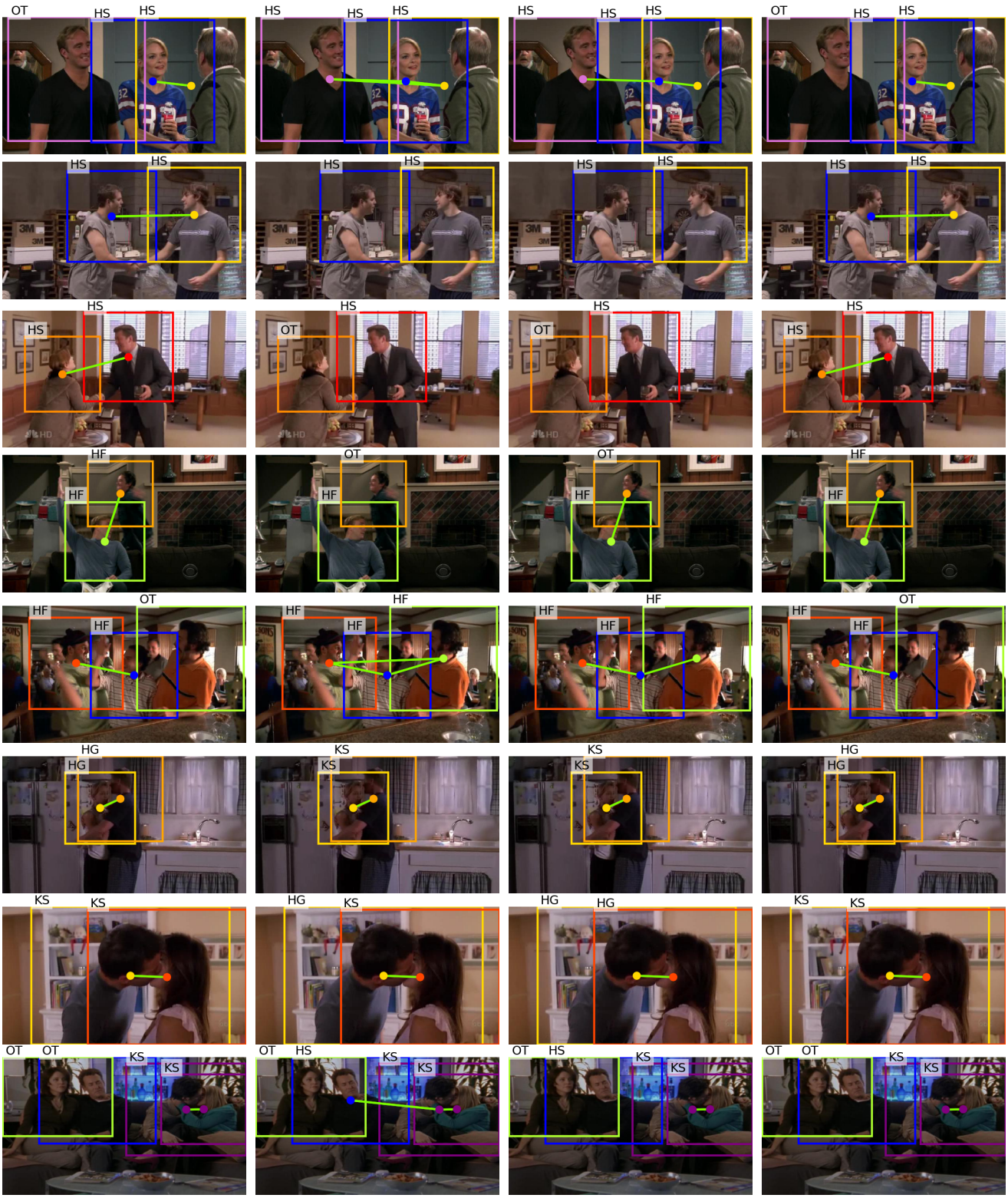
Here we visualize the learned $\lambda^{\mathcal{C}}(y_i, y_j)\ \forall (y_i, y_j) \in \mathcal{Y}^2$ as a matrix on both BIT and TVHI (Figure 2). One can see that our proposed CAR module is able to learn suitable $\lambda^{\mathcal{C}}$ automatically from data. For example, on BIT the labeling $(\mathrm{KK}, \mathrm{HG})$, *i.e.* (*kick*, *hug*) is not encouraged according to the left matrix in Figure 2 since they are not compatible.

3. More Qualitative HIU Results



(a) Groundtruth (b) Base model (c) Modified GN (d) CAGNet

Figure 3. More qualitative HIU results on BIT. Each row shows an example. Columns from left to right correspond to results of *groundtruth*, *base-model*, *Modified GN*, and *CAGNet*. Green lines denote predicted interactive pairs ($z = 1$). Texts present predicted individual actions (y variables), where *BX, PS, PT, HS, HG, KS, HF, KK, OT* mean *box, push, point, handshake, hug, kiss, high-five, kick, others* respectively. Note that the predictions of CAGNet (the rightmost column) always obey the two oracles defined in Section 4.3.



(a) Groundtruth (b) Base model (c) Modified GN (d) CAGNet

Figure 4. More qualitative HIU results on TVHI. Each row shows an example. Columns from left to right correspond to results of *groundtruth*, *base-model*, *Modified GN*, and *CAGNet*. Green lines denote predicted interactive pairs ($z = 1$). Texts present predicted individual actions (y variables), where *HS, HF, HG, KS, OT* mean *handshake, high-five, hug, kiss, others* respectively. Note that the predictions of CAGNet (the rightmost column) always obey the two oracles defined in Section 4.3.