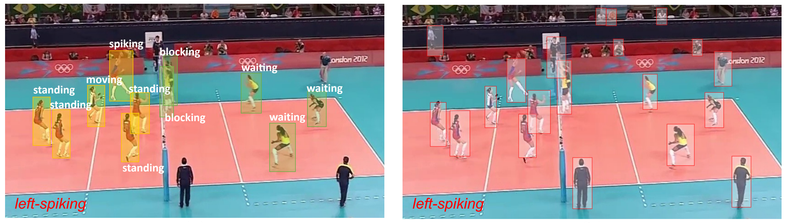
Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition

Mingfei Han[1]*, David Junhao Zhang[2]*, Yali Wang[3]*, Rui Yan[2], Lina Yao[4], Xiaojun Chang[1,5], Yu Qiao3,6 ✉

[1] ReLER Lab, AAII, UTS [2] NUS [3] SIAT, CAS [4] UNSW [5] RMIT [6] Shanghai AI Lab

Group Activity Recognition:

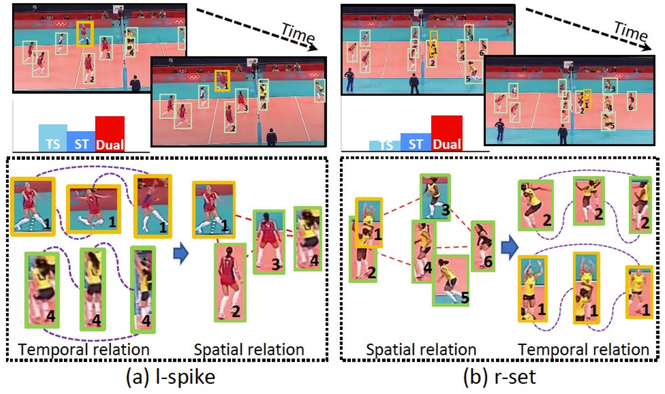


(a) (b)

Fig.1 Center frame for *left-spiking*. (a) Full supervision: Accurate positions, individual action and group activity labels (b) Weak supervision: Detected positions and group activity label.

- GAR contains multi-level interactions, *i.e.*, actor-actor and actor-group.
- Given a video containing several actors, the model is required to infer the group activity, i.e., *left-spiking* in Fig.1, by learning the multi-level relations.
- Actor positions and group activity labels are the minimum requirement for our method. *One model for multiple data settings*!

Motivation:



(a) I-spike (b) r-set

Fig.2 Example of *I-spike* and *r-set*. TS path and ST path are skilled at different classes.

- Fig. 2 (a) refers to the *I-spike* activity, where the hitting player (actor 1) and the defending player (actor 4) move fast to hit and block the ball, while other players (e.g., actor 2 and actor 3) stand without much movement. Hence, it's better to model temporal dynamics first.
- On the contrary, Fig. 1 (b) refers to the *r-set* activity in the volleyball, where most players on the right-side team are moving cooperatively to tackle the ball falling in different positions. Hence, it is better to reason spatial actor interaction first.

With spatial and temporal modeling applied in different orders, ST path and TS path are skilled at different classes

Framework:

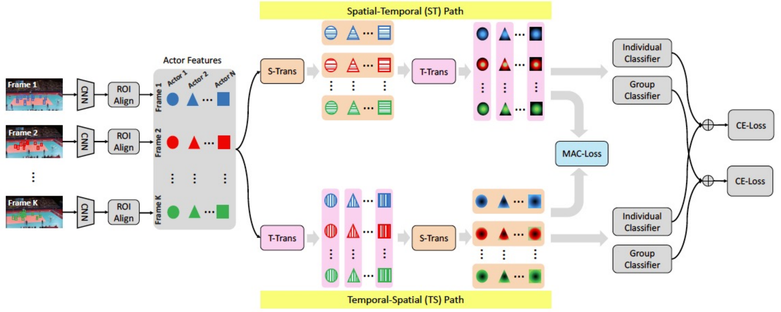


Fig.3 Framework. Dual-path actor interaction is achieved by switching the order of spatial and temporal relation modules. Different colors, shapes and patterns mean different frames, actors and TS/ST respectively.

- Given a video containing K frames, we can obtain features $\mathbf{X} \in \mathbb{R}^{K\times N\times C}$ for N actors by RoI Align.
- Actor features are then fed into two complementary spatiotemporal modeling paths for actor evolution, *i.e.*, ST and TS.

$$\mathbf{X}_{\mathrm{ST}} = \text{T-Trans}(\mathbf{X} + \text{MLP}(\text{S-Trans}(\mathbf{X}))$$
$$\mathbf{X}_{\mathrm{TS}} = \text{S-Trans}(\mathbf{X} + \text{MLP}(\text{T-Trans}(\mathbf{X}))$$

- A concise Multi-scale Actor Contrastive Loss is performed on $\mathbf{X}_{\mathrm{ST}}$ and $\mathbf{X}_{\mathrm{TS}}$ for inter-path interaction. See paper Sec. 3.3 for details.

Performance (50% data for full-dataset performance):

Our method achieves state-of-the-art performance on several datasets and multiple supervision settings. See Sec. 4.3 for details.

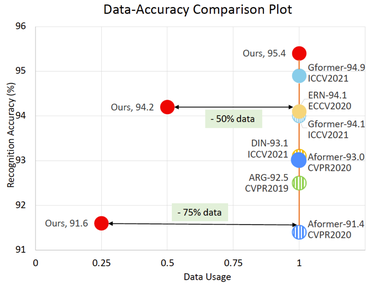


Fig.4 Accuracy comparison with data in different percentage on Volleyball dataset.

Remarkably, as shown in Fig.4, our method achieves 94.2% with 50% data, which is competitive to a number of recent approaches (ERN-ECCV 2020, GroupFormer-ICCV 2021) trained with 100% data.

Visualization:

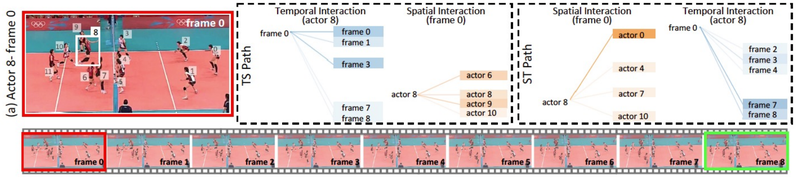


Fig.5 Actor interaction visualization for I-spike activity with connected lines. Brighter color indicates stronger relation. For actor 8 in frame 0, we visualize the temporal interaction with the same actors in different frames for ST and TS paths.

- The spiking player (actor 8) is more related with accompanying players in TS path, who are "moving" (actor 6 and 10) and "standing" (actor 9).
- Differently, in ST path, actor 8 has wider connections with other players (actor 7 and actor 10) and defending players (actor 0 and actor 4).

Contact us:

- Please visit https://mingfei.info/Dual-AI, or scan the QR Code.