COMPOSER: Compositional Reasoning of Group Activity in Videos with Keypoint-Only Modality

Honglu Zhou[1], Asim Kadav[2], Aviv Shamsian[3], Shijie Geng[1], Farley Lai[2], Long Zhao[4], Ting Liu[4], Mubbasir Kapadia[1], and Hans Peter Graf[2]

[1] Rutgers University & [2] NEC Laboratories America, Inc. & [4] Google Research & [3] Bar-Ilan University

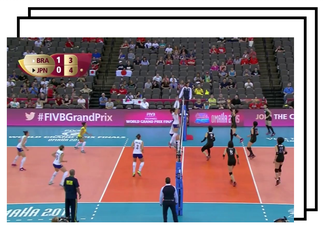
GitHub: https://github.com/hongluzhou/composer

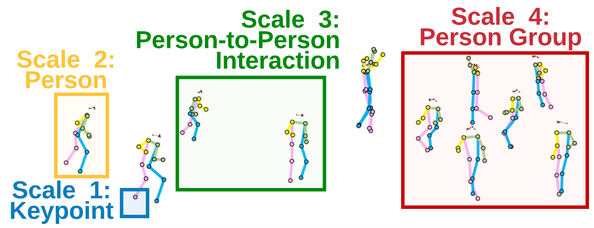
Introduction

Group Activity Recognition (GAR) detects the activity collectively performed by a group of actors in a short video.

Motivations:
1. GAR requires compositional reasoning of actors and objects.
2. Prior works suffer from scene biases with privacy and ethical concerns.

Our Solutions:
1. Model the video as tokens that represent the multi-scale semantic concepts.
2. Use only the keypoint modality.



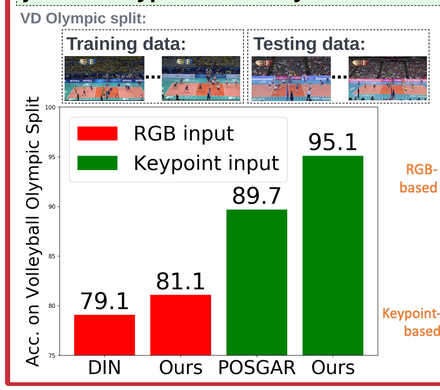


Results

Dataset: VD (Volleyball) & CAD (Collective Activity)
Metric: Accuracy

Best, Second, Third

The keypoint-only setup generalizes better for GAR. Ours achieves +5.4% improvement with just the keypoint modality on the VD Olympic split.



Model	VD	
	Original	Olympic
I3D (Carreiraet et al., 2017)	84.6	73.9
VGG16 +RoIAlign (Simonyanet et al., 2015)	91.6	76.4
PCTDM (Yan et al., 2018)	91.7	75.2
SACRF (Pramonoet et al., 2020)	91.8	71.1
Actor-Transformer (Gavrilyuket et al., 2020)	93.0	76.9
ARG (Wu et al., 2019)	93.3	77.8
TCE-STBiP(Yuan et al., 2021)	93.5	78.5
DIN (Yuan et al., 2021)	93.6	79.1
POSGAR – no ball track (Thilakarathneet et al., 2021)	93.2	89.7
COMPOSER – no ball track	93.7	95.1
Improvement	+0.1%	+5.4%

RGB-based / Keypoint-based

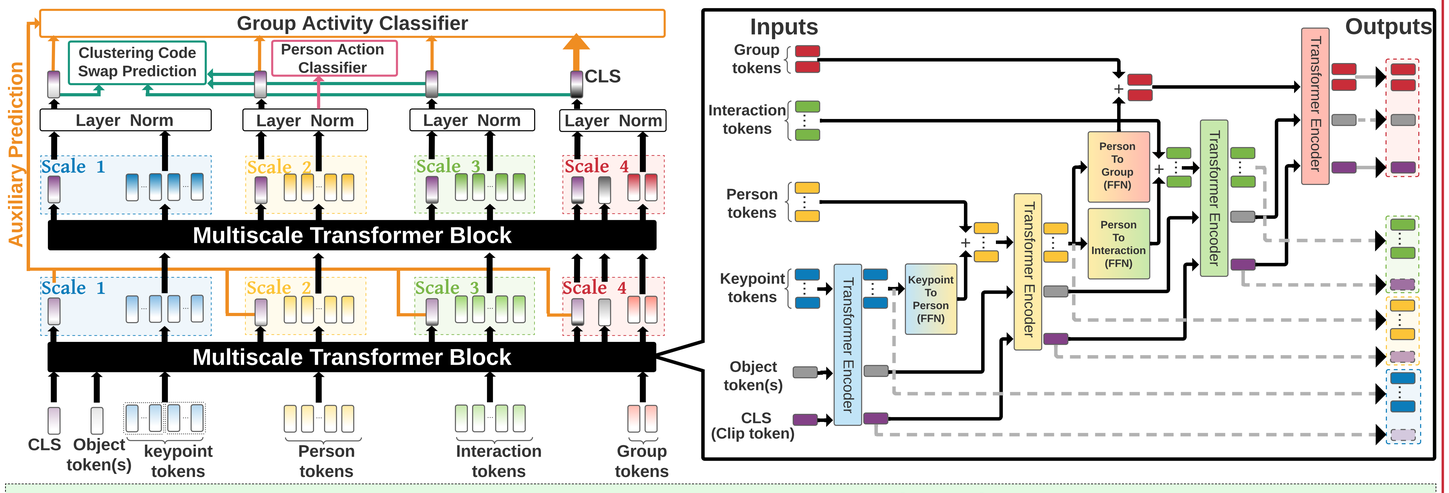
COMPOSER outperforms the GAR methods that use a single modality, and performs favorably compared with methods that exploit multiple expensive modalities (COMPOSER use less computations!).

Model	Pose	RGB	Flow	Scene	VD	CAD
HDTM		✓			81.9	81.5
CERN		✓			83.3	87.2
stagNet		✓			89.3	89.1
CRM		✓	✓		93.0	85.8
ARG		✓			92.5	91.0
HiGCIN		✓			91.5	93.4
Ehsanpouret et al.		✓		✓	93.1	89.4
DIN		✓			93.6	N/A
Actor-Transformer		✓	✓		93.0	92.8
SACRF	✓	✓	✓	✓	95.0	95.2
GroupFormer	✓	✓	✓	✓	95.7	96.3
COMPOSER (ours)	✓				94.6	96.2

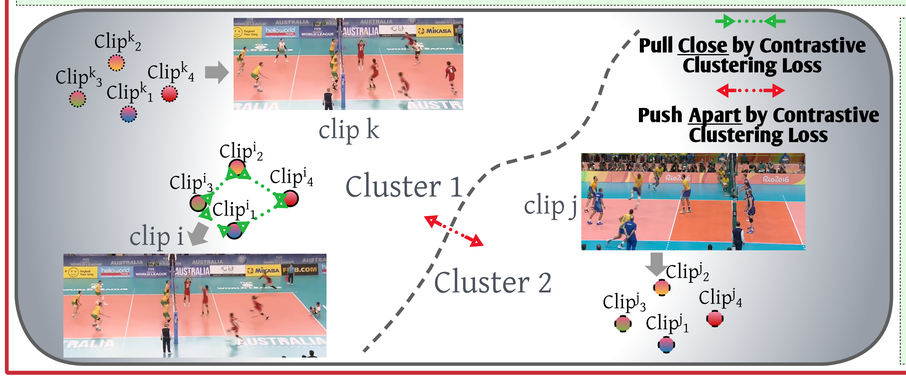
Method

Research Question: How to learn multi-scale representations to aid GAR using the keypoint-only modality?

Novelty 1: COMPOSER is proposed, a Multiscale Transformer based architecture that performs attention-based reasoning over tokens at each scale and learns group activity compositionally.



Novelty 2: Contrastive Multiscale Representation Learning, achieved by clustering the intermediate scale representations, while *maintaining consistent cluster assignments between scales of the same video clip*.



Novelty 3: Auxiliary Prediction & Data Augmentations, tailored to the keypoint signals.

- Actor Dropout
- Horizontal Flip
- Horizontal Move
- Vertical Move

} Data Augmentations

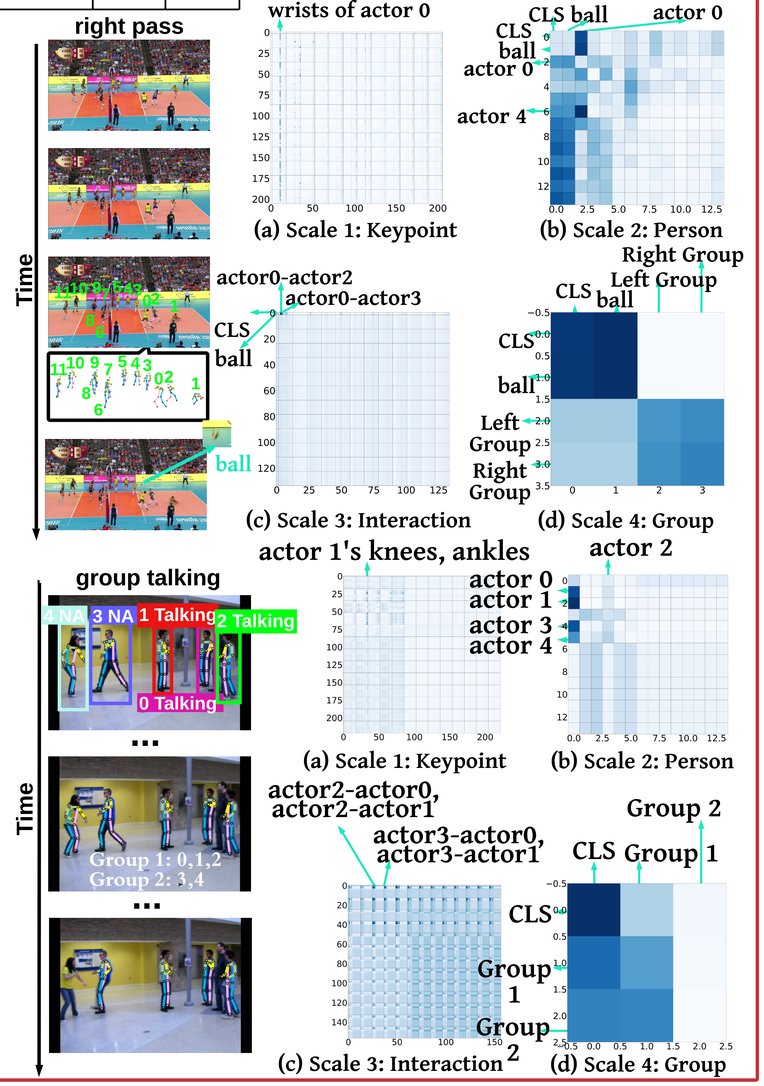
$$\mathcal{L}_{\text{total}} = \sum_{m=1}^{M-1} \mathcal{L}_{\text{groupAux}} + \lambda\left(\mathcal{L}_{\text{groupLast}} + \mathcal{L}_{\text{person}} + \mathcal{L}_{\text{cluster}}\right)$$

Auxiliary Multi-task Learning

m is the index of the Multiscale Transformer block

Qualitative analysis – showcasing attention matrices. Tokens that the model has mostly attended to at each scale are highlighted.

- COMPOSER is able to attend to relevant information across different scales.
- COMPOSER can produce interpretable results.



Conclusion

Contributions:
1. Compositional and relational reasoning at different scales for GAR.
2. Contrastive clustering, auxiliary prediction, data augmentation techniques to improve the intermediate representations.
3. Extraordinary scene generalization capability using only the keypoint modality.

Future work:
- More complex scenarios, e.g., crowd understanding.
- Use additional modalities like RGB while maintaining scene generalization.

Interested to know more?

Check out the supplementary video!

SCAN ME