Deep Learning-based Group Activity Recognition: Survey Summary

Prepared by Omid Naeej Nejad

November 2025

Outline

What is Group Activity Recognition (GAR)?

- Task: Identify the activity performed by a group of people in a video.
- More complex than single-person action recognition.
- Requires understanding:
  - Individual actions
  - Inter-person interactions
  - Temporal evolution
  - Group-level dynamics

Why is GAR important?

- Autonomous driving (predict pedestrian behaviour)
- Sports analytics (volleyball, basketball, hockey)
- Surveillance (crowd behaviour, anomaly detection)
- Smart homes & healthcare

Definition

- Group activity = coordinated behavior of 2+ individuals.
- Requires human detection + temporal modeling + relation reasoning.
- Output:
  - Individual action labels
  - Group activity label

Major Challenges

- Complex spatial–temporal relationships.
- Multiple subgroups; key actors matter.
- Real-time joint detection–tracking–recognition.
- Visual privacy & weak generalization.
- Non-standardized, imbalanced datasets.

Overview of Survey Taxonomy

Four main taxonomies:

1. Supervision Types
2. Network Types
3. Modeling Mechanisms
4. Input Types

Supervision Types

- Fully-supervised
- Weakly-supervised
- Self-supervised
- Semi-supervised & Reinforcement Learning

Fully-supervised Methods

Characteristics:

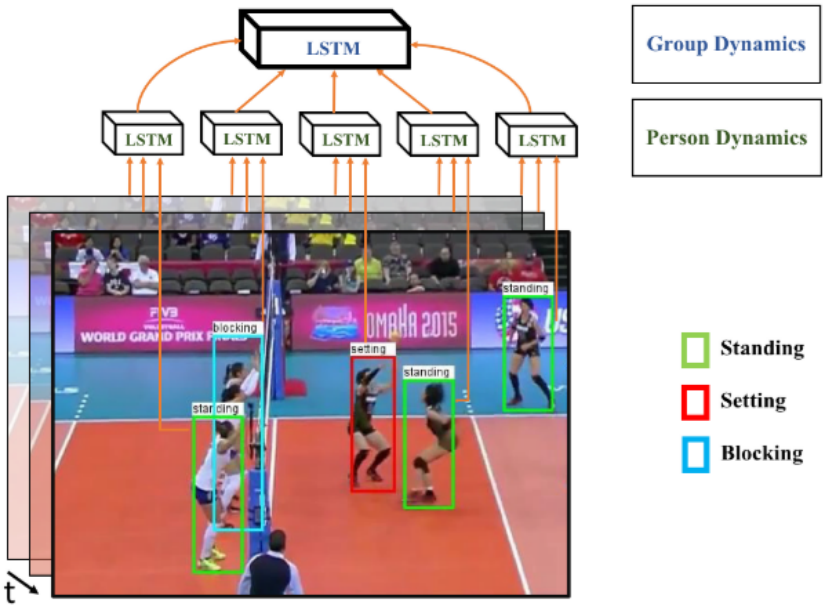
- Require bounding boxes + individual actions + group labels.
- Highest accuracy.

Examples:

- HDTM (CNN+LSTM)
- ARG (GCN)
- GroupFormer (Transformer)
- MLST-Former (Transformer)

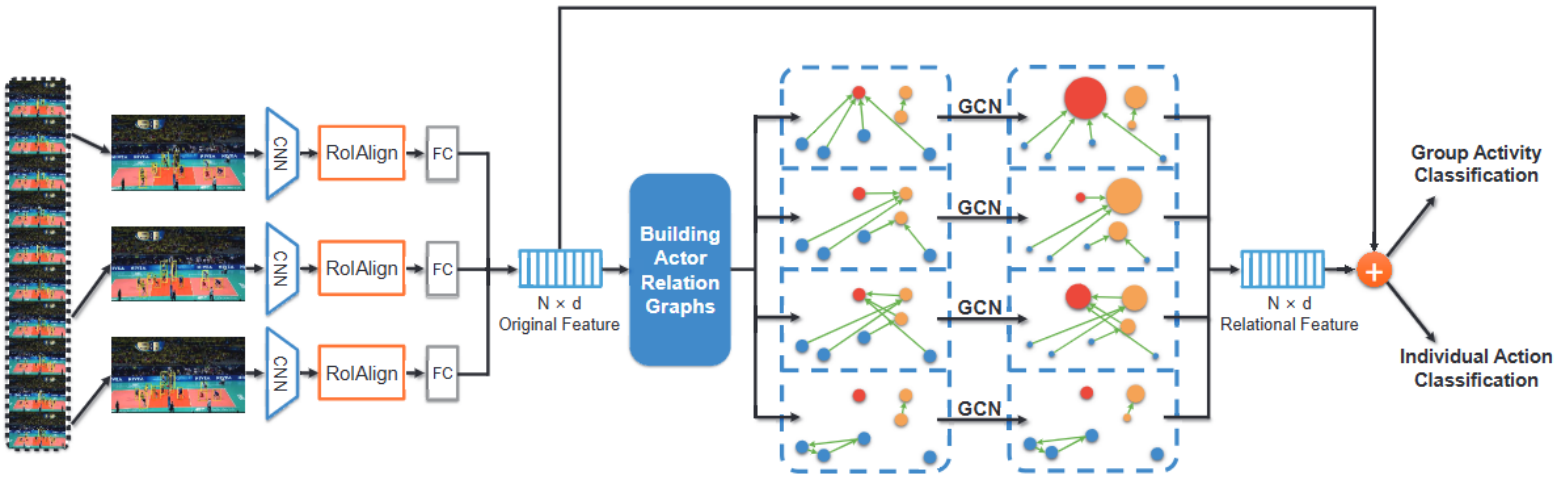
HDTM

HDTM (Hierarchical Deep Temporal Model) uses a two-level temporal architecture. First, CNN features of each person are processed by individual-level LSTMs to capture their temporal actions. Then a group-level LSTM aggregates all individual sequences to infer the global activity. This hierarchical design captures both person-level evolution and group-level dynamics, forming a structured interpretation pipeline.



ARG (Actor Relation Graph)

ARG builds a graph where nodes represent actors and edges represent pairwise relations derived from visual features. A GCN then propagates information across this graph to learn interaction-aware representations. By explicitly modeling relationships between individuals, ARG improves recognition of complex group activities and long-range dependencies.



GroupFormer

GroupFormer treats each person as a token in a transformer.
Self-attention layers let the network learn long-range interactions, identify important actors, and model both spatial and temporal structure jointly.
Its transformer design provides stronger global reasoning than traditional LSTMs or GCNs.

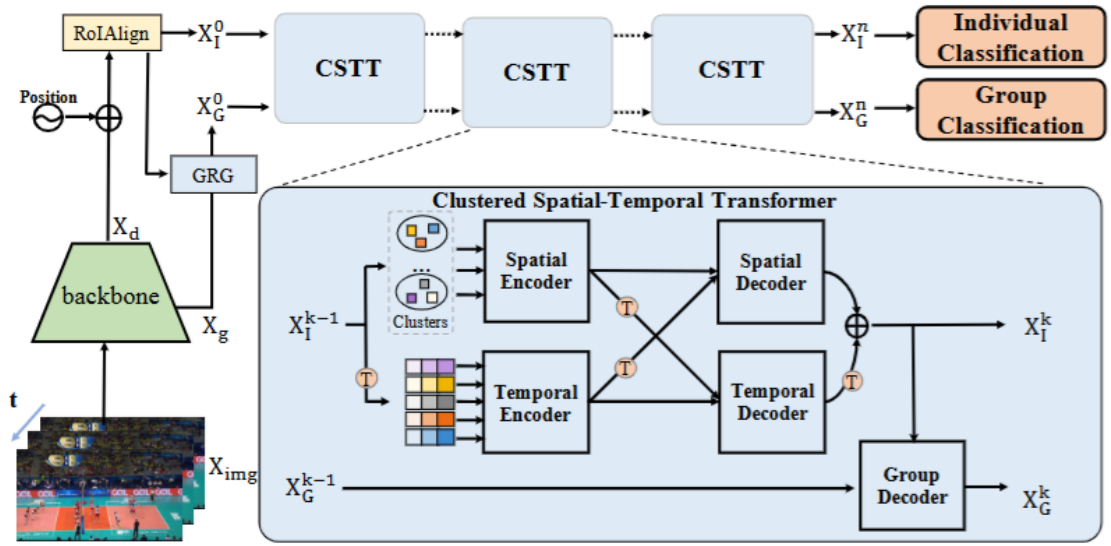


Figure 2. Illustration of our proposed GroupFomer. It contains three main components: 1) a CNN backbone that extracts feature representation of video clips. 2) a Group Representation Generator that initializes the group representation from individual and scene features. 3) a Clustered Spatial-Temporal Transformer that models the spatial-temporal relations and refines the group representation and individual representation.

MLST-Former

MLST-Former (Multi-Level Spatial-Temporal Transformer) extracts actor features at multiple hierarchical scales. Local subgroups, global formations, and temporal sequences are jointly processed through stacked spatial-temporal attention blocks. This multi-level transformer captures fine-grained relations and global group structure more effectively than single-scale models.

Fully-Supervised Methods Glance

Model	Volleyball (VD)	Collective (CAD)	CAED
HDTM	yes	yes	no
ARG	yes	yes	no
GroupFormer	yes	yes	no
MLST-Former	yes	yes	yes

Dataset	Group activities (scene-level)	Individual actions (person-level)
Volleyball (VD)	right set, right spike, right pass, right winpoint, left set, left spike, left pass, left winpoint	waiting, setting, digging, falling, spiking, blocking, jumping, moving, standing
Collective (CAD)	crossing, waiting, queueing, walking, talking	NA, crossing, waiting, queueing, walking, talking
CAED	crossing, waiting, queueing, talking, dancing, jogging	same as CAD (individual-level labels)

Model	VD group acc.	VD indiv. acc.	CAD group acc.	CAED group acc.
HDTM	51.1%	–	81.5%	–
ARG	92.5%	83.0%	91.0%	–
GroupFormer	95.7%	85.6%	96.3%	–
MLST-Former	94.5%	84.5%	96.8%	95.9%

Weakly-supervised Methods

Characteristics:

- Use only video-level labels.
- No bounding boxes required.

Examples:

- LRMM (Local Relative Motion Module)
- Social Adaptive Module (SAM)
- Detector-free WSGAR
- ZSTGroupCLIP

LRMM

LRMM (Local Relative Motion Module) models group activities by explicitly capturing the relative motion between individuals, rather than relying on bounding-box annotations or detailed actor labels. The method extracts local motion features for each person and then computes pairwise relative-motion descriptors that encode how individuals move with respect to each other. These relational motion cues are aggregated over time to form a group-level representation.

SAM (Social Adaptive Module)

SAM learns to assign soft attention weights to individuals without any actor-level supervision. The module implicitly identifies which people are most relevant for the group activity by learning social importance scores. This attention mechanism improves performance when only video-level annotations are available.

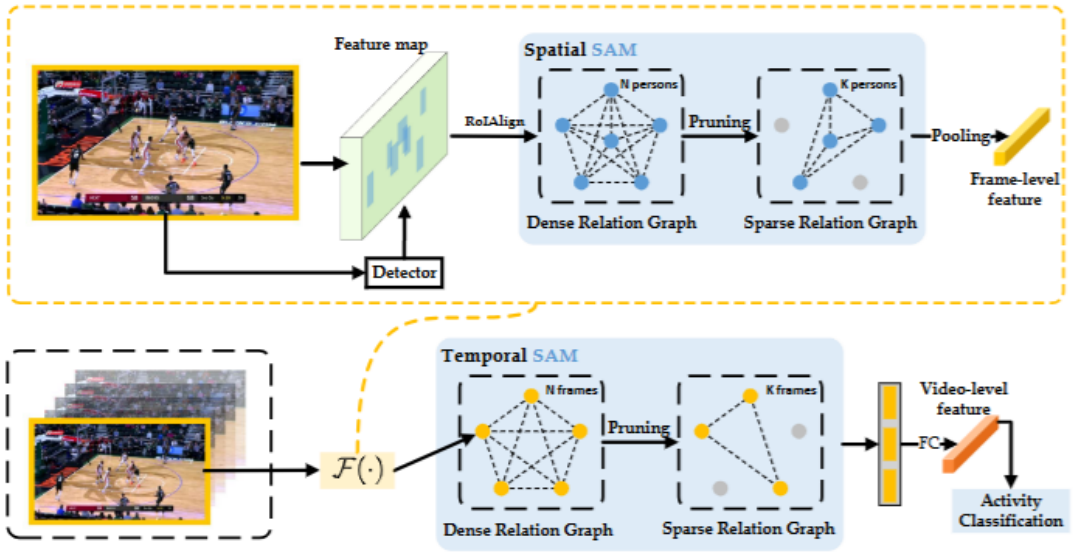


Fig. 2. Overview of our approach for weakly-supervised GAR. The inputs are a set of frames and the associating pre-detected bounding boxes for people. We apply SAM to concurrently select discriminative person-level feature in spatial domain and effective frame-level representations in temporal domain (Best viewed in color)

Detector-free WSGAR

This model removes the need for any person detection and instead uses 3D CNNs or spatio-temporal transformers to learn implicit representations of interactions. The network discovers actor groups and their relations directly from raw video, bypassing detector errors and annotation cost.

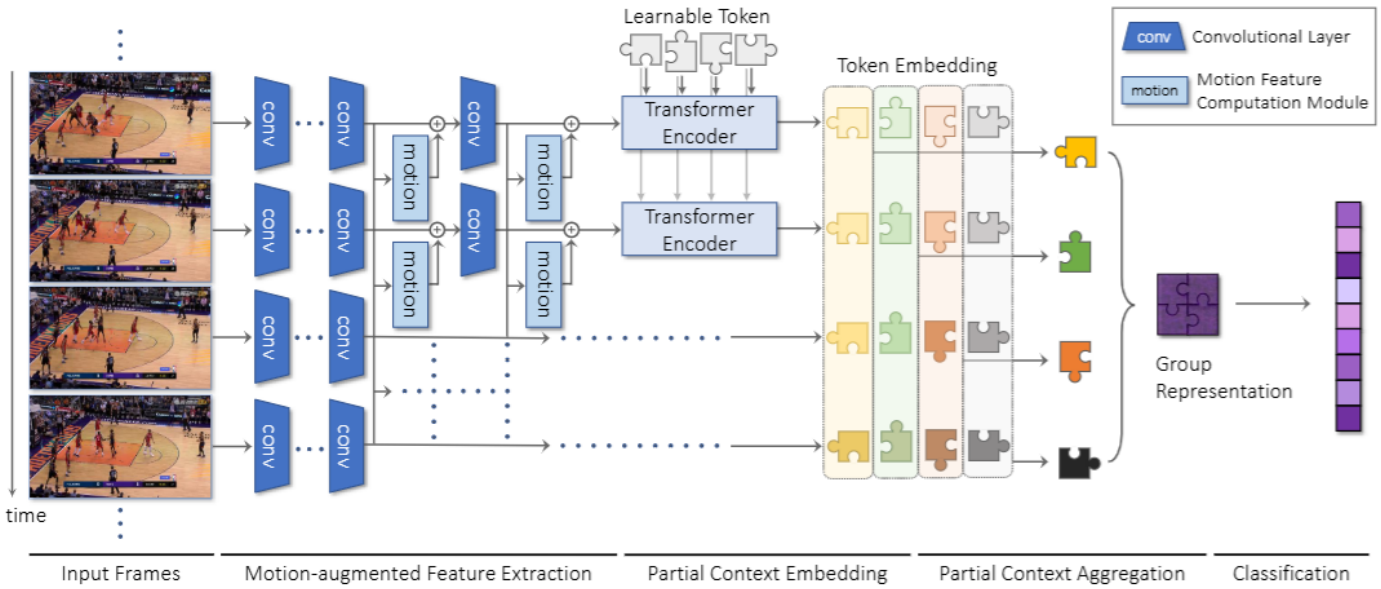


Figure 2. Overall architecture of our model. A CNN incorporating motion feature computation modules extracts a motion-augmented feature map per frame. At each frame, a set of learnable tokens (unpainted pieces of Jigsaw puzzles) are embedded to localize clues useful for group activity recognition through the attention mechanism of the Transformer encoder. The token embeddings (painted pieces of Jigsaw puzzles) are then fused to form a group representation in two steps: First aggregate embeddings of the same token (pieces with the same shape) across time, and then aggregate the results of different tokens (pieces with different shapes and colors). Finally, the group representation is fed into the classifier which predicts group activity class scores.

ZSTGroupCLIP

ZSTGroupCLIP adapts CLIP's vision–language alignment to GAR. It describes group activities using text prompts and compares them with video embeddings. Because it relies on cross-modal alignment rather than annotated GAR datasets, it performs zero-shot recognition for unseen activities.

Weakly-supervised Methods Glance

Model	Volleyball (VD)	NBA	Collective (CAD)
SAM	yes	yes	no
DFWSGAR	yes	yes	no
LRMM+GCM	yes	yes	yes
ZSTGroupCLIP	yes	yes	yes

Dataset	Group activities (scene-level)	Individual actions / notes
Volleyball (VD)	right set, right spike, right pass, right winpoint, left set, left spike, left pass, left winpoint	9 individual volleyball actions exist, but weakly-supervised methods use only video-level team labels (no boxes).
NBA	basketball event types: 2-point and 3-point attempts (success or miss) and layups with offensive/defensive rebounds	only video-level event labels; no individual player annotations.
Collective (CAD)	crossing, walking, waiting, talking, queueing	in weakly-supervised GAR, only the clip-level majority group label is used (no explicit person labels).

Model	VD MCA	VD merged MCA	NBA MCA	NBA MPCA
SAM (ResNet-18)	86.3%	93.1%	54.3%	51.5%
DFWSGAR	90.5%	94.4%	75.8%	71.2%
LRMM+GCM	92.8%	95.6%	77.8%	73.2%

Self-supervised Methods

Characteristics:

- No human labels.
- Use contrastive or predictive pretext tasks.

Examples:

- AAGCM (Skeleton SSL)
- SPARTAN (SSL Transformer)
- SoGAR (Social SSL)
- GSTCo (Contrastive SSL)

AAGCM

AAGCM (Adversarial Adaptive Graph Contrastive Model) learns skeleton-based GAR without labels by applying contrastive learning on graph-structured data. It perturbs nodes/edges adversarially and forces consistency in representation. This improves robustness to noise, occlusion, and viewpoint changes while capturing social interactions through the graph structure.

SPARTAN

SPARTAN uses a transformer encoder trained with self-supervised objectives such as masked prediction, temporal ordering, and contrastive alignment. These tasks force the transformer to learn contextual and relational information from unlabeled video, enabling strong downstream GAR performance.

SoGAR

SoGAR (Social Self-Supervised GAR) designs pretext tasks based on social consistency—e.g., predicting spatial formations or relative proximities. By exploiting social dynamics, it learns group-aware features without requiring manually annotated labels.

GSTCo

GSTCo (Graph Spatio-Temporal Contrastive) trains on skeleton graphs using multi-level contrastive learning (node, person, and group). This enforces consistent representations across augmentations and time windows, enabling strong self-supervised GAR performance.

Self-supervised Methods Glance 1

Model	Volleyball (VD)	Collective (CAD)	NBA	JRDB-PAR
AAGCM (skeleton SSL)	yes	yes	no	no
SPARTAN (ViT SSL)	yes	no	yes	no
SoGAR (social SSL)	yes	no	yes	yes
GSTCo (contrastive)	yes	yes	no	no

Dataset	Group activities (scene-level)	Individual / social labels and notes
Volleyball (VD)	8 volleyball group activities: right set, right spike, right pass, right winpoint, left set, left spike, left pass, left winpoint.	9 player actions such as waiting, setting, digging, falling, spiking, blocking, jumping, moving, standing; bounding boxes for each player.
Collective (CAD)	5 collective activities: crossing, walking, waiting, queueing, talking.	6 individual actions: NA, crossing, walking, waiting, queueing, talking; person boxes for each pedestrian.
JRDB-PAR	11 social group activities in panoramic robot videos (e.g. standing closely, chatting, walking together, sitting closely, group working).	27 individual actions (e.g. walking, talking) and 7 global scene activities; dense bounding boxes for all people in each frame.

Self-supervised Methods Glance 2

Model	VD MCA / MPCA	CAD MCA / MPCA	NBA MCA / MPCA	JRDB-PAR (P_g / R_g / F_g)
AAGCM	94.0% / 94.4%	97.1% / 96.1%	–	–
SPARTAN	92.9% / –	–	82.1% / 72.8%	–
SoGAR	–	–	83.3% / 73.5%	49.3% / 47.1% / 48.7%
GSTCo	–	96.8% / –	–	–

Semi-supervised & Reinforcement Learning

- Semi-supervised: use pseudo-labels.
- RL: learn interaction reasoning policies.

Examples:

- MLS-GAN (semi-supervised)
- PRL (reinforcement learning)

MLS-GAN

MLS-GAN combines CNN features with a GAN structure. The generator predicts group activity using both labeled and unlabeled data, while the discriminator evaluates prediction credibility. The multi-level design captures both individual and scene-level context, improving GAR when labeled data is limited.

PRL (Policy Reinforcement Learning)

PRL formulates GAR as a reinforcement learning problem where the model learns a policy to select key individuals and relations. The agent receives rewards for identifying informative interactions, gradually improving relational reasoning and group activity predictions.

Semi-supervised & Reinforcement Learning Glance

Model	Volleyball (VD)	Collective (CAD)	Other
MLS-GAN (semi-supervised)	yes	yes	–
PRL (reinforcement learning)	yes	yes	–

Dataset	Group activities (scene-level)	Individual actions / notes
Volleyball (VD)	8 team activities: right set, right spike, right pass, right winpoint, left set, left spike, left pass, left winpoint.	9 player actions: waiting, setting, digging, falling, spiking, blocking, jumping, moving, standing.
Collective (CAD)	5 collective activities: crossing, walking, waiting, talking, queueing.	6 individual actions: NA, crossing, walking, waiting, talking, queueing; group label is majority of individual actions.

Model	VD MCA	VD MPCA	CAD MCA	CAD MPCA
MLS-GAN	93.0%	–	91.7%	–
PRL	91.4%	91.8%	–	93.8%

Network Types

- CNN / 3D CNN
- RNN / LSTM
- Graph Convolutional Networks (GCN)
- Transformers

CNN / 3D CNN

Strengths:

- Strong spatial feature extraction.
- 3D CNNs capture short-term temporal info.

Example Models:

- CRM
- I3D variants

CRM

CRM (Contextual Relation Model) uses 3D CNNs to extract spatio-temporal features and a relation reasoning module to identify interactions among individuals. Local motion patterns and contextual cues are integrated to infer higher-level group behaviors.

RNN / LSTM

Strengths:

- Sequential temporal modeling.

Representative:

- HDTM
- SBGAR
- CERN

SBGAR

SBGAR (Semantics-Based Group Activity Recognition) models group activities by building a semantic graph where individuals and contextual cues form nodes linked through meaningful relations such as cooperation or spatial alignment. A hierarchical reasoning module aggregates these semantic relations into a coherent group-level representation. By emphasizing relational semantics rather than only motion, SBGAR improves recognition of complex coordinated behaviors.

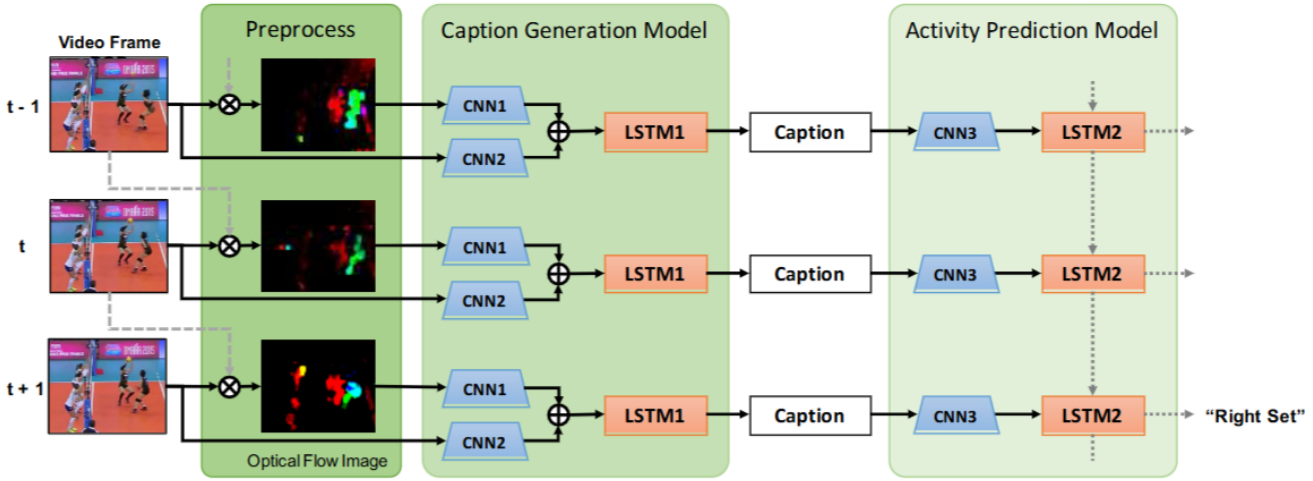


Figure 1: The architecture of the proposed Scheme. Caption Generation Model generates a caption to describe the corresponding frame. Activity Prediction Model is used to predict the group activity based on generated captions of a continuous sequence of frames. Symbol $\otimes$ indicates the operation of computing the dense optical flow image using two continuous frames, while symbol $\oplus$ indicates the operation of concatenating two CNN feature vectors into one single vector. In order to simplify the figure, the details of models are not shown here. Please refer to Figure 2 for more details of the Caption Generation Model, and Figure 3 for the Activity Prediction Model.

CERN

CERN (Confidence-Energy Recurrent Network) integrates LSTMs with an energy-based model that enforces consistency across individual and group predictions. The energy function penalizes inconsistent outputs, resulting in more stable and robust group activity classification.

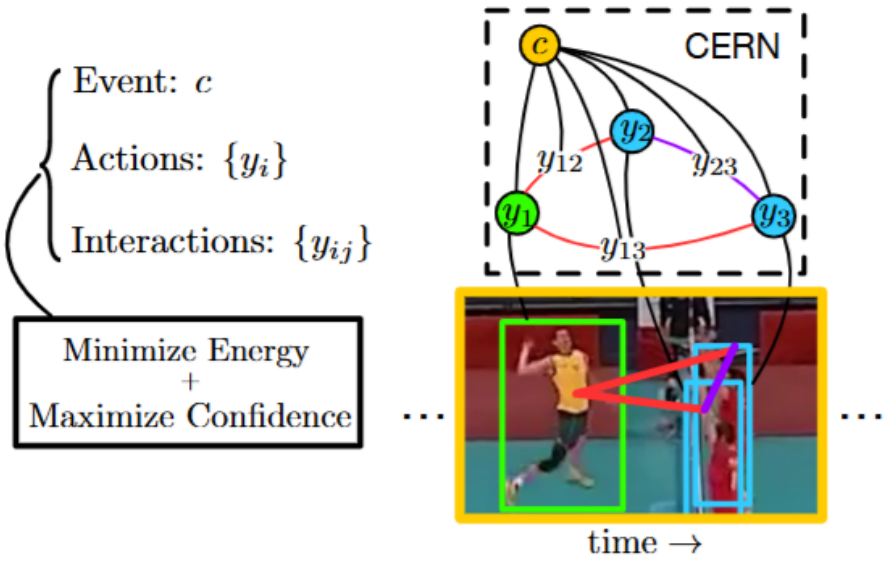


Figure 1: Our CERN represents a two-level hierarchy of LSTMs grounded onto human trajectories, where the LSTMs predict individual actions $\{y_i\}$, human interactions $\{y_{ij}\}$, or the event class c in a given video. CERN outputs an optimal configuration of LSTM predictions which jointly minimizes the energy of the predictions and maximizes their confidence, for addressing the brittleness of cascaded predictions under uncertainty. This is realized by extending the two-level hierarchy with an additional energy layer, which

Graph Convolutional Networks (GCN)

Strengths:

- Explicit relational reasoning.
- Actor relation graphs.

Examples:

- ARG
- DIN
- GAIM

DIN

DIN (Dynamic Inference Network) builds a time-varying interaction graph that adapts at each frame based on the appearance, motion, and spatial context of individuals. Instead of using a fixed relation structure, the model dynamically predicts which actor pairs should interact, enabling flexible reasoning over evolving group formations. This adaptive graph inference improves recognition of complex and rapidly changing group activities.

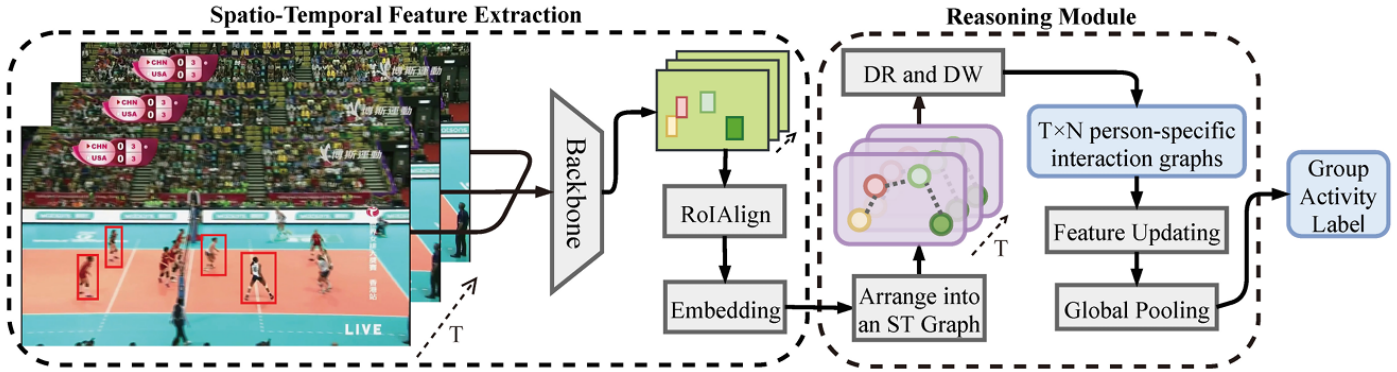


Figure 3. **The overall pipeline of Dynamic Inference Network**. Generally, it consists of two stages: i) Spatio-temporal feature extraction, ii) Reasoning module. Note that there will be $T \times N$ unique interaction graphs for updating. In our codebase, the first stage is shared with previous methods. The main variations are in the Reasoning Module. We only illustrate 4 bounding boxes in the image for clarity.

GAIM

GAIM (Group Activity Interaction Model) uses hierarchical graphs to represent subgroup-level and global interactions. GCN layers capture multi-level dependencies, making it effective for scenes with structured formations such as team sports.

Transformers

Strengths:

- Long-range interactions.
- Joint spatial–temporal attention.

Examples:

- Actor-Transformer (AT)
- GroupFormer
- MLST-Former
- KRGFormer

Actor Transformer

Actor Transformer embeds each person as a token and applies self-attention across actors and time. This enables long-range relational modeling and key-actor identification, outperforming traditional sequence models.

KRGFormer

KRGFormer (Key-Relation Guided Transformer) enhances the transformer by estimating important actor–actor relations and guiding attention toward them. This targeted relational focus improves accuracy in crowded scenes with many distractors.

DynamicFormer

DynamicFormer predicts dynamic attention masks that change over time, allowing the model to adapt to shifting group formations. This helps recognize evolving activities in sports and irregular crowd behavior.

Network-based Methods Glance 1

Model	Volleyball (VD)	Collective (CAD)	NBA / other
CRM	yes	yes	–
SBGAR	yes	yes	–
CERN)	yes	yes	–
DIN	yes	yes	–
GAIM	yes	yes	CollectiveSports
Actor-Transformer	yes	yes (re-eval.)	–
KRGFormer	yes	–	NBA, VolleyTactic
DynamicFormer	yes	yes	keypoint-only (VD, CAD)

l >p0.40 >p0.40

Dataset	Group activities (scene-level)	Individual / person-level actions or notes
Volleyball (VD)	right set, right spike, right pass, right winpoint, left set, left spike, left pass, left winpoint	waiting, setting, digging, falling, spiking, blocking, jumping, moving, standing
Collective Activity (CAD)	crossing, walking, waiting, talking, queueing	NA, crossing, walking, waiting, talking (individual action labels)
NBA	nine basketball events: 2-point and 3-point attempts (made / missed) and layups with offensive / defensive rebound variants	only group-activity labels in this setting (no individual actions)

Network-based Methods Glance 2

Model	VD MCA	VD MPCA	Comment
SBGAR	66.9	67.6	early semantic context baseline
CERN-2	83.3	83.6	confidence–energy recurrent model
CRM	92.1	–	convolutional relational machine
Actor-Transformer	90.0	90.2	transformer over actor tokens
DIN (VGG-16 backbone)	93.6	93.8	dynamic inference relations
KRGFormer	94.0	94.4	key-role guided transformer
DynamicFormer (keypoint-only)	95.3	–	dynamic composition + interaction
GAIM	–	–	uses VD/CAD; numeric values omitted here

Modeling Mechanisms

- Hierarchical temporal modeling
- Interaction relationship modeling
- Attention modeling

Hierarchical Temporal Modeling

Two-stage LSTM:

- Individual-level LSTM
- Group-level LSTM

Examples:

- HDTM
- SBGAR
- GLIL

Interaction Relationship Modeling

- Graph-based relational reasoning
- Hierarchical relation networks
- Hypergraphs

Examples:

- HRN
- HiGCIN
- Hypergraph GAR

Attention Modeling

Three types:

1. Key actor extraction
2. Multi-level attention
3. Self-attention

Examples:

- PCTDM
- HAN-HCN
- Actor-Transformer
- DynamicFormer

Input Types

- Single-modality (RGB, Skeleton)
- Multimodal Fusion
- Multimodal Input (visual + non-visual)

RGB-based Methods

RGB methods use appearance and motion cues from video frames. While they capture rich visual information, they can be sensitive to lighting, occlusion, and background clutter. Transformers and GCNs have improved relational reasoning in RGB-based GAR.

- Most common.
- Sensitive to lighting and occlusion.

Examples:

- GroupFormer
- ARG
- MLST-Former

Skeleton-based Methods

Skeleton-based models use pose keypoints, offering better privacy and robustness to appearance variation. Methods like POGARS, Composer, and AAGCM use GCNs or transformers on skeletal graphs to learn human–human interactions effectively.

- Better privacy.
- Strong structural representation.

Examples:

- POGARS
- Composer
- AAGCM

Multimodal Fusion

Multimodal fusion combines RGB, flow, pose, or other modalities. Early fusion merges features at input, while late fusion merges predictions. Hybrid approaches integrate interactions at multiple network stages. Techniques like MLST-Former and AT leverage multimodal cues effectively.

- Early fusion: feature-level
- Late fusion: score-level
- Hybrid fusion

Examples:

- Multi-stream CNN
- AT
- MLST-Former

Surveillance Datasets

- CAD, CAED, NCAD
- UCLA Courtyard
- Nursing Home Dataset

Sports Datasets

- Volleyball Dataset (VD)
- NBA Dataset
- NCAA Basketball
- Soccer Dataset
- HARD (Hockey)

Evaluation Metrics

- Multi-class Classification Accuracy (MCA)
- Mean Per-Class Accuracy (MPCA)
- Confusion Matrix

Future Research Directions

- End-to-end real-time GAR
- Few-shot / zero-shot GAR
- Larger standardized datasets
- Robust low-frequency activity recognition
- Multimodal GAR (text, audio, sensors)
- Extensions: GAP, PAR, SGAR, EAR

Conclusion

- GAR is complex and evolving.
- Transformers and GCNs dominate SOTA.
- Multimodal learning and SSL are promising.

References I

[1] L. Shao, Z. Gao, M. Zhang, X. Chen, Y. He, and D.-Y. Yeung, "Deep learning-based group activity recognition in videos: A survey," Information Fusion, 2024.
[2] I. A. Papadopoulos et al., "Group activity recognition using deep models," IEEE Transactions on Pattern Analysis and Machine Intelligence.
[3] Q. Li, Z. Han, and X. Wu, "GroupFormer: Group Activity Recognition with Temporal Transformers," CVPR.
[4] X. Wang, H. Xu, and M. Chen, "Actor-Relation Graph for Group Activity Recognition," CVPR.
[5] W. Wu et al., "MLST-Former: Multi-Level Spatio-Temporal Transformer for Group Activity Recognition," AAAI.
[6] L. Zhang et al., "Detector-Free Weakly Supervised GAR," ICCV.
[7] Y. Yan et al., "SoGAR: Social Self-Supervised Group Activity Representation," ECCV.

References II

[8] G. Chen et al., "GSTCo: Graph Spatio-Temporal Contrastive Learning for Group Detection," ICCV.