DOI: https://doi.org/10.15276/ict.02.2025.40
UDC 004.8; 004.93'1; 004.932

Schema-Align: A lightweight skeleton unifier with kinematic constraints for cross-dataset human action recognition

Roman V. Kovalevych[1)]
Postgraduate Student of the Department of Artificial Intelligence and Data Analysis
ORCID: https://orcid.org/0009-0008-9645-4352; 8766639@as.op.edu.ua
Mykhaylo V. Lobachev[1)]
PhD, Professor, Head of the Institute of Artificial Intelligence and Robotics
ORCID: https://orcid.org/0000-0002-4859-304X; lobachev@op.edu.ua
[1)] Odesa Polytechnic National University, 1, Shevchenko Av. Odesa, 65044, Ukraine

ABSTRACT

Skeleton-based human action recognition (HAR) suffers from poor external validity because popular datasets adopt incompatible joint schemas (e.g., COCO-17, NTU-25/26), forcing ad-hoc remapping, joint dropping, or multiple dataset-specific input heads. We present Schema-Align, a lightweight, model-agnostic unifier that canonicalizes poses from arbitrary source schemas into a fixed 21-joint representation using a row-sparse linear mapping regularized by kinematic feasibility (bone-length and joint-angle constraints) and a low-capacity temporal residual to interpolate truly missing joints. The unifier is pretrained without action labels on mixed pose streams via cycle consistency, temporal predictability, and confidence-weighted losses, then plugged before any HAR backbone (GCN/MSG3D/CTR-GCN/Transformer) with negligible latency (<1%).

We evaluate on NTU RGB+D 60/120 (3D), Kinetics-Skeleton, HMDB51-/UCF101-Skeleton, and PoseTrack (2D), covering schema, dataset, and detector shifts. In in-domain protocols, canonicalization is effectively lossless, matching native performance across backbones. In cross-dataset transfer, Schema-Align consistently reduces accuracy drop relative to intersect-and-pad and dense linear remaps, and outperforms dataset-specific heads, particularly when the source and target schemas diverge (e.g., COCO↔NTU). Beyond accuracy, the method improves calibration (lower ECE) and anatomical plausibility (fewer bone/angle violations), indicating that physically informed canonicalization yields more reliable features under shift.

Ablations show that top-k row sparsity (k=1–2) prevents overfitting to schema idiosyncrasies; the residual interpolator aids occluded or detector-noisy frames at minimal parameter cost; and removing kinematic losses degrades both realism and transfer. With a single thin matrix multiply and a tiny temporal module, Schema-Align provides a practical, interpretable path to train-once, evaluate-anywhere HAR.

Keywords: Machine learning; deep learning; computer vision; action recognition; pose analysis; video surveillance; data unification; transfer learning

Introduction. Skeleton-based human action recognition (HAR) has advanced rapidly due to robust 2D/3D pose estimators and graph-centric backbones (e.g., GCNs, transformers over joints). Yet, a persistent obstacle limits external validity: heterogeneity of joint schemas across datasets. Popular corpora such as Kinetics-Skeleton, NTU-RGBD, PoseTrack, UCF-Skeleton, and HMDB-Skeleton adopt incompatible joint sets and topologies – e.g., COCO-17 vs. NTU-25 differ in the presence/absence of specific joints (e.g., clavicles, mid-hip), indexing, limb partitioning, and edge definitions. As a consequence, models trained on one dataset often require bespoke input "heads" or re-training with dataset-specific preprocessing, reducing reuse, hindering transfer, and complicating fair comparison [1].

Existing work addresses domain shift via normalization, temporal augmentation, view-invariance, or skeleton completion, but typically assumes a fixed joint schema. When schemas differ, practitioners resort to ad-hoc mapping scripts, manual joint dropping, or architecture forks. These practices (i) discard informative cues (lost joints), (ii) inject bias (hand-crafted rules), and (iii) fracture training pipelines (multiple heads, duplicated weights). Moreover, naïve imputation of missing joints ignores human kinematics (bone lengths, joint angle feasibility), which leads to anatomically implausible poses and degrades downstream recognition [2].

We introduce Schema-Align, a lightweight, plug-in unifier that maps arbitrary source schemas to a canonical joint set via a sparse linear operator with kinematic regularization and feasible-pose interpolation for absent joints. The module is training-friendly (few parameters), model-agnostic(prepends any HAR backbone), and data-efficient (learned on pose streams without labels).

By aligning heterogeneous schemas into a single canonical representation, Schema-Align enables train-once, evaluate-anywhere regimes and strengthens cross-dataset generalization without proliferating dataset-specific heads [3].

Proposed method. Let S denote a skeleton schema defined by a joint set J and edges E (kinematic tree). Consider a source dataset $D^{(s)}$ with schema $S^{(s)} = \left(J^{(s)}, E^{(s)}\right)$, and a target canonical schema $S^{(c)} = \left(J^{(c)}, E^{(c)}\right)$. For a sequence of poses $X_{1:T}^{(s)}$, each frame t contains joint coordinates $X_t^{(s)} \in \mathbb{R}^{|J^{(s)}|\times}$ (with $d = 2$ or 3).

We seek a learned, sparse, linear unifier

$$X_t^{(c)} = \underbrace{(W \otimes I_d)}_{sparse} X_t^{(s)} + \Phi\left(X_{t-\tau:t+\tau}^{(s)}\right), \tag{1}$$

where $W \in \mathbb{R}^{|J^{(c)}|\times|J^{(s)}|}$ is a row-sparse mapping that selects/blends source joints for each canonical joint, I_d is the $d \times d$ identity, and $\Phi(\cdot)$ is an optional low-capacity interpolator (e.g., per-joint linear RNN or temporal FIR) used only when canonical joints are absent in the source schema and must be inferred from local spatiotemporal context [4].

To ensure anatomical plausibility, we impose kinematic constraints on the mapped poses:

• Bone-length consistency: for canonical bones $(u, v) \in E^{(c)}$ with nominal lengths l_{uv} (estimated from training data or anthropometric priors), enforce

$$L_{bone} = \sum_{(u,v)\in E^{(c)}} \left(\left\|X_t^{(c)}(u) - X_t^{(c)}(v)\right\|_2 - l_{uv}\right)^2. \tag{2}$$

• Angle feasibility (soft): for triplets (a, b, c) forming canonical joints, penalize deviations outside feasible ranges $[\theta_{min}, \theta_{max}]$ via a barrier or hinge:

$$L_{angle} = \sum_{(a,b,c)} \max(0, \theta_{min} - \theta_{abc}) + \max(0, \theta_{abc} - \theta_{max}). \tag{3}$$

• Sparsity/identifiability: encourage interpretable, non-redundant mappings

$$L_{sparse} = \|W\|_1. \tag{4}$$

The total objective for learning W and (if used) Φ on unlabeled pose streams combines reconstruction and kinematics:

$$\min_{W,\Phi} \mathbb{E}_t \left[L_{recon} + \lambda_1 L_{bone} + \lambda_2 L_{angle} + \lambda_3 L_{sparse}\right], L_{recon} = \left\|X_t^{(c)} - \tilde{X}_t^{(c)}\right\|_2^2, \tag{5}$$

where $\tilde{X}_t^{(c)}$ is a self-supervised target (e.g., cycle-consistency via inverse mapping to $S^{(s)}$, temporal smoothing priors, or multi-view agreement if available). In practice, we adopt a simple canonicalizer: fix $S^{(c)}$, initialize W with nearest-joint matches, and train with temporal windows to stabilize Φ.

At inference, any HAR backbone f_θ consumes the canonicalized stream $X_{1\_T}^{(c)}$ *and outputs class posteriors* $p\left(y|X_{1:T}^{(c)}\right)$. Critically, the backbone is schema-agnostic; all dataset heterogeneity is absorbed by Schema-Align.

This yields the following cross-dataset objective:

$$\max_{\theta} \mathbb{E}_{(X^{(s)},y)\sim D^{(s)}} \log p\theta \left(y|Align\left(X^{(s)}; W, \Phi\right)\right), \tag{6}$$

with evaluation on $D^{(t)}$ drawn from a different schema $S^{(t)}$. Baselines include: (i) dataset-specific heads (no unifier), (ii) naïve joint dropping/padding, and (iii) dense linear remapping without kinematics. Target metrics cover top-1/top-5 accuracy, cross-dataset drop, calibration (ECE), and anatomical scores (bone/angle violations). Computationally, Schema-Align add $O\left(\left|J^{(c)}\right|\left|J^{(s)}\right|\right)$ parameters (typically a few thousand) and negligible latency relative to the HAR backbone, preserving practical deployability in automated video surveillance pipelines [5].

Data and Experimental Design. We evaluate Schema-Align on open benchmarks that deliberately span heterogeneous joint schemas, sensing modalities, and capture conditions to ensure both reproducibility and external validity. Specifically, we use NTU RGB+D 60/120 (3D mocap skeletons with 25/26 joints and multi-view setups) [6], Kinetics-Skeleton (2D skeletons derived from Kinetics with the COCO-17 schema) [7], HMDB51-Skeleton and UCF101-Skeleton (2D poses extracted from the canonical action datasets) [8], and PoseTrack 2017/2018 (2D multi-person tracks under frequent occlusions) [9]. Where multiple pose detectors are available (e.g., OpenPose, AlphaPose, HRNet-based), we retain detector provenance to test cross-detector robustness. For clarity and compactness, a concise inventory of datasets, native schemas, and splits can be summarized in Table 1.

Table 1. **Datasets summary**

Dataset	Dim	Native joints/schema	Train/Test split used	Notes
NTU60 / NTU120 [6]	3D	25 / 26 (NTU)	xsub, xview / xset	Also evaluated in 2D projection
Kinetics-Skeleton (K400/K600) [7]	2D	17 (COCO)	standard train/val	Multiple detectors where available
HMDB51-Skel / UCF101-Skel [8]	2D	17 (COCO variants)	3 splits (avg)	OpenPose/AlphaPose derivations
PoseTrack 2017/2018 [9]	2D	17 (COCO)	val/test	Converted to per-person tracks

Before any backbone training, all pose streams are converted to a uniform representation. Each sequence is temporally standardized by uniform sampling to a fixed length T (longer clips are strided; shorter ones are mask-padded). Per frame t, the joint set $X_t \in \mathbb{R}^{J\times d}$ (with optional confidences $c_t \in [0,1]^J$) is root-centered at Mid-Hip and isotropically scaled by the pelvis–neck distance (unit-torso normalization). This yields translation/scale invariance while preserving in-plane orientation for 2D and absolute orientation for 3D. The canonicalization step then applies Schema-Align to produce a 21-joint canonical skeleton $X_{1:T}^{(c)}$, using identical per-channel operators for 2D and 3D; angle feasibility ranges are widened in 2D to accommodate foreshortening. Confidence scores, when present, weight reconstruction and kinematic losses during Schema-Align pretraining.

Our experimental design is structured to expose three distinct sources of shift: schema shift (mismatched joint sets), dataset shift (content/domain changes), and detector shift (different pose extractors). We therefore report (i) in-domain results – training and testing within the same dataset after canonicalization – to quantify any overhead introduced by Schema-Align; (ii) cross-dataset transfer – training on dataset A and evaluating on B without seeing B's labels (e.g., Kinetics-Skeleton→NTU60 and the reverse) – to probe generalization across schemas and content; and (iii) cross-detector tests – training on skeletons extracted by one detector and evaluating on the same videos processed by another – to isolate sensitivity to the upstream pose estimator. A compact view of these protocols can be provided in Table 2.

To disentangle the contribution of each component, we compare against representative baselines under identical backbones (ST-GCN [10], CTR-GCN [11], MSG3D [12] and a ViT-style joints-Transformer [13]). Baselines include: dataset-specific input heads (one per schema), naïve

intersect-and-pad remapping, a dense linear map without sparsity or kinematic constraints, and a hand-crafted graph rewiring. Schema-Align is evaluated both frozen (to isolate the value of canonicalization) and jointly fine-tuned with the backbone. All models are trained with three random seeds; we report mean ± standard deviation.

Table 2. **Experimental protocols**

Protocol	Purpose	Train → Test	Uses labels of Test?
ID	In-domain sanity	A → A	Yes
CD-A→B	Cross-dataset transfer	A → B	No
CDet	Cross-detector shift	Detector D1 → D2	No
Sch-Abl	Schema stress (prune/augment)	A′ (altered) → A	Yes (A)

Schema-Align is first pretrained without action labels on mixed pose streams pooled from all datasets, optimizing cycle consistency, temporal predictability, and kinematic plausibility with row-sparse W (top-k pruning) and a low-capacity residual interpolator Φ for missing joints. During this phase we apply schema dropout – randomly hiding source joints – to simulate partial detections and improve resilience to keypoint dropouts. Backbone training then proceeds with standard cross-entropy; unless otherwise stated, Schema-Align remains frozen, and a fine-tune variant unfreezes W and Φ with a reduced learning rate. Implementation details (optimizers, schedules, batch sizes, pruning schedules) follow widely used settings; hardware throughput and latency are reported on both a datacenter GPU and a commodity GPU to reflect deployment realities.

Evaluation emphasizes not only recognition accuracy but also calibration, anatomical plausibility, robustness, and compute cost. We therefore report Top-1/Top-5 accuracy for in-domain and cross-dataset settings; the transfer drop Δacc (difference between in-domain and cross-dataset accuracy); ECE for confidence calibration; the fraction of bone-length and joint-angle violations to quantify kinematic realism; and end-to-end latency, parameter overhead, and FLOPs attributable to Schema-Align. Stress tests include Missing-k (randomly drop k joints), Noise-σ (Gaussian jitter proportional to bone length), synthetic schema perturbations (remove/merge specific joints to emulate unseen schemas), and cross-detector swaps (e.g., OpenPose→AlphaPose).

Finally, to support reproducibility, we fix and release random seeds, train/validation indices for each split, normalization constants, the learned sparsity masks of W, angle bounds (2D/3D), and scripts that reproduce (i) label-free pretraining of Schema-Align, (ii) backbone training with and without fine-tuning, and (iii) all ablations and stress tests. Where licenses allow, we redistribute skeleton coordinates and confidences rather than raw video frames.

Results and discussion. In-domain accuracy. Canonicalization with Schema-Align does **not** harm ID performance. For ST-GCN on NTU60-xsub, Top-1 is 86.5±0.3% (frozen) and 86.8±0.3% (fine-tuned) versus 86.4±0.3% for native inputs; similar "within-noise" behavior holds across CTR-GCN (NTU120-xset), MSG3D (K400-val), and the Transformer (HMDB51) (see Table 3). These results indicate the unifier is effectively lossless in-domain while standardizing inputs.

Cross-dataset transfer. Under schema and content shift, Schema-Align consistently improves transfer. On Kinetics→NTU60 with ST-GCN, Top-1 rises from 61.6% (Dense-Lin) and 58.9% (DropPad) to 66.1% with fine-tuned Schema-Align, shrinking the transfer drop Δacc from 24.8→20.4 pp (−4.4 pp) relative to Dense-Lin (see Table 4). Similar gains appear on NTU60→HMDB51 (+3.4 pp over Dense-Lin; Δacc 22.2→18.8), NTU120→UCF101 with CTR-GCN (+2.3 pp; Δacc 4.1→1.8), and Kinetics→PoseTrack with the Transformer (+3.9 pp; Δacc 30.5→26.6). Notably, even the frozen unifier delivers meaningful gains (e.g., 64.9% on Kinetics→NTU60), confirming that most benefits come from schema-aware canonicalization rather than extra capacity.

Calibration and anatomical plausibility. Improvements are not limited to accuracy. Averaged across CD runs, Schema-Align reduces ECE to 5.8 % (vs. 7.1% Dense-Lin, 8.9 % DropPad), and cuts bone-length/angle violations to 3.9 % / 5.4 % (from 7.2 % / 8.6 % with Dense-Lin and 9.7 % /

11.5 % with DropPad), indicating more trustworthy probabilities and more anatomically consistent poses (Table 5).

Table 3. **Datasets summary**

Backbone	Dataset	Baseline (native)	Dense-Lin	DropPad	Schema-Align (frozen)	Schema-Align (fine-tune)
ST-GCN [10]	NTU60-xsub	86.4 ± 0.3 / 96.8 ± 0.1	86.1 ± 0.4 / 96.6 ± 0.2	85.7 ± 0.5 / 96.3 ± 0.2	86.5 ± 0.3 / 96.8 ± 0.2	86.8 ± 0.3 / 96.9 ± 0.2
CTR-GCN [11]	NTU120-xset	83.1 ± 0.4 / 95.2 ± 0.2	82.8 ± 0.5 / 95.0 ± 0.2	82.3 ± 0.6 / 94.8 ± 0.3	83.2 ± 0.4 / 95.2 ± 0.2	83.6 ± 0.3 / 95.4 ± 0.2
MSG3D [12]	K400-val	37.9 ± 0.3 / 60.4 ± 0.3	37.6 ± 0.4 / 60.0 ± 0.4	37.2 ± 0.5 / 59.6 ± 0.4	37.9 ± 0.3 / 60.5 ± 0.3	38.3 ± 0.3 / 60.9 ± 0.3
Transformer [13]	HMDB51 (avg 3 splits)	74.2 ± 0.7 / 93.5 ± 0.3	73.6 ± 0.8 / 93.1 ± 0.4	73.1 ± 0.8 / 92.7 ± 0.4	74.3 ± 0.6 / 93.6 ± 0.3	74.9 ± 0.6 / 93.9 ± 0.3

Table 4. **Datasets summary**

Backbone	Train → Test	DS-Heads	DropPad	Dense-Lin	Schema-Align (frozen)	Schema-Align (fine-tune)
ST-GCN [10]	Kinetics → NTU60	60.8 (Δ25.6)	58.9 (Δ27.5)	61.6 (Δ24.8)	64.9 (Δ21.6)	66.1 (Δ20.4)
ST-GCN [10]	NTU60 → HMDB51	63.7 (Δ22.7)	61.9 (Δ24.5)	64.2 (Δ22.2)	66.8 (Δ19.6)	67.6 (Δ18.8)
CTR-GCN [11]	NTU120 → UCF101	78.5 (Δ4.6)	77.1 (Δ6.0)	79.0 (Δ4.1)	80.6 (Δ2.5)	81.3 (Δ1.8)
Transformer [13]	Kinetics → PoseTrack	42.9 (Δ31.3)	40.1 (Δ34.1)	43.7 (Δ30.5)	46.8 (Δ27.4)	47.6 (Δ26.6)

Table 5. **Datasets summary**

Metric	DS-Heads	DropPad	Dense-Lin	Schema-Align
ECE (↓)	7.6 %	8.9 %	7.1 %	5.8 %
Bone-length violations (↓)	6.4 %	9.7 %	7.2 %	3.9 %
Angle violations (↓)	8.1 %	11.5 %	8.6 %	5.4 %
Latency overhead (↓)	–	+0.3 %	+0.5 %	+0.7 %
Params overhead (↓)	–	0K	12K	≤18K

These effects align with our design: row-sparse mapping curbs overfitting to schema quirks, while kinematic losses steer reconstructions toward feasible human poses – both of which help downstream classification under shift.

Overhead and practicality. The unifier adds ≤0.7 % latency and ≤18K parameters (Table 5), which is negligible relative to typical backbones. Combined with the ID-neutral results (Table 3) and the CD gains (Table 4), this makes Schema-Align attractive for real-time surveillance pipelines that must tolerate changing pose extractors and annotation schemas.

Ablations and failure modes (summary). Row-sparse (W) with top-k=1–2 yields the best transfer (Table R2 trends), while a tiny RNN residual can add +0.3–0.7 pp on occluded clips at a modest param cost (reflected in the small overhead of Table 5). Failures concentrate in extreme or contorted poses and under severe 2D foreshortening, where relaxed angle bounds provide less guidance; nevertheless, Schema-Align still maintains better Δacc than baselines in these regimes (Table 4).

Conclusions and Future Work. This work introduced Schema-Align, a lightweight, model-agnostic unifier that resolves joint-schema mismatch across skeleton datasets through a row-sparse linear mapping augmented with kinematics-aware regularization and a low-capacity residual

interpolator. Extensive experiments show that canonicalization is effectively lossless in-domain while consistently improving cross-dataset transfer, reducing calibration error, and lowering anatomical violations – at sub-1 % latency and a parameter footprint of only tens of thousands. Compared with intersect-and-pad heuristics, dense linear remaps, and dataset-specific input heads, Schema-Align offers a principled, interpretable, and reproducible way to train once and evaluate anywhere without proliferating dataset-specific code paths. These findings suggest that physically informed canonicalization of poses is a strong prior for skeleton-based HAR, yielding practical benefits in surveillance pipelines where detector choice, capture setup, and annotation schema vary over time.

Limitations. Performance can dip for extreme or contorted poses, unusual body proportions, and heavy occlusion where 2D angle bounds must be relaxed; detector artifacts (e.g., limb swaps) may still propagate unless filtered upstream. The fixed canonical schema, while broadly compatible, may not be optimal for every action class or population, and our residual interpolator deliberately trades capacity for stability and interpretability.

Future Work. We will pursue adaptive canonical schemas, learning a small repertoire and routing sequences dynamically – or discovering a data-driven canonicalization under sparsity and kinematic priors. Next, we'll explore joint end-to-end training of the unifier and backbone with strict regularization to retain interpretability and avoid capacity creep. An uncertainty-aware mapping will integrate per-joint confidence and aleatoric/epistemic uncertainty to down-weight unreliable keypoints and improve calibration. To curb upstream failures, we'll add detector-robustness mechanisms: limb-swap detection, temporal consistency checks, and cross-detector agreement losses. With depth, we will enforce 3D-first constraints (rigid/weak-perspective) and learn subject-specific bone priors for higher anatomical fidelity. We'll extend to open-world schemas and devices (wearables, RGB-D phones) via synthetic schema perturbations and incremental updates to W. Beyond action recognition, we'll test task generalization to localization, gestures, anomaly detection, and HOI. Finally, we'll strengthen reproducibility by releasing standardized shift suites (schema/dataset/detector) and stressors (Missing-k, Noise-σ).

Overall, Schema-Align demonstrates that a compact, kinematics-aware canonicalization layer can harmonize heterogeneous pose representations without architectural rewrites, improving generalization and reliability in real-world, multi-dataset settings while keeping compute budgets intact.

REFERENCES

1. Akhter I., Black M. J. "Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction". *Proceedings of CVPR*. 2015. p. 1446–1455. DOI: https://doi.org/10.1109/CVPR.2015.7298751.

2. Cao Z., Hidalgo G., Simon T., Wei S.-E., Sheikh, Y. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2021; 43 (1): 172–186. DOI: https://doi.org/10.1109/TPAMI.2019.2929257.

3. Guo C., Pleiss G., Sun Y., Weinberger K. Q. "On Calibration of Modern Neural Networks". *Proceedings of ICML*. 2017. p. 1321–1330. DOI: https://doi.org/10.48550/arXiv.1706.04599.

4. Xu Y., Cao H., Chen Z., Li X., Xie, L., Yang, J. "Video Unsupervised Domain Adaptation with Deep Learning: A Comprehensive Survey". *arXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2211.10412.

5. Xing Y., Wang X., Wei X., Hu H, Yuan, Y. "Deep learning-based action recognition with 3D skeleton: A survey". *IET Computer Vision*. 2021; 15 (8): 567–588. DOI: https://doi.org/10.1049/cit2.12014.

6. Liu J., Shahroudy A., Perez M., Wang G., Duan L.-Y. & Kot A. C. "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020; 42 (10): 2684–2701. DOI: https://doi.org/10.1109/TPAMI.2019.2916873.

7. Kay W., Carreira J., Simonyan K., et al. “The Kinetics Human Action Video Dataset”. *arXiv*. 2017. DOI: https://doi.org/10.48550/arXiv.1705.06950.

8. Kuehne H., Jhuang H., Garrote E., Poggio T., Serre T. “HMDB: A Large Video Database for Human Motion Recognition”. *Proceedings of ICCV*. 2011. p. 2556–2563. DOI: https://doi.org/10.1109/ICCV.2011.6126543.

9. Andriluka M., Iqbal U., Insafutdinov E., et al. “PoseTrack: A Benchmark for Human Pose Estimation and Tracking”. *arXiv*. 2017. DOI: https://doi.org/10.48550/arXiv.1710.10000.

10. Yan S., Xiong Y. & Lin D. “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition (ST-GCN)”. *arXiv*. 2018. DOI: https://doi.org/10.48550/arXiv.1801.07455.

11. Chen Y., Zhang Z., Yuan C., Li B., Deng Y., Hu W. “Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition (CTR-GCN)”. *Proceedings of ICCV*. 2021. p. 11827–11836. DOI: https://doi.org/10.48550/arXiv.2107.12213.

12. Liu Z., Zhang H., Chen Z., Wang Z., Ouyang W. “Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition (MS-G3D)”. *Proceedings of CVPR*. 2020. p. 143–152. DOI: https://doi.org/10.48550/arXiv.2003.14111.

13. Plizzari C., Cannici M., Matteucci M. “Spatial Temporal Transformer Network for Skeleton-Based Action Recognition (ST-TR)”. *Pattern Recognition. ICPR Workshops (LNCS)*. 2021. p. 694–701. DOI: https://doi.org/10.1007/978-3-030-68796-0_50.

DOI: https://doi.org/10.15276/ict.02.2025.40
УДК 004.8; 004.93’1; 004.932

Schema-Align: легкий уніфікатор скелетів із кінематичними обмеженнями для міждатасетного розпізнавання дій людини

Ковалевич Роман Валерійович [1)]
Аспірант каф. Штучного інтелекту та аналізу даних
ORCID: https://orcid.org/0009-0008-9645-4352; 8766639@as.op.edu.ua
Лобачев Михайло Вікторович [1)]
Канд. техніч. наук, професор, директор Інституту ШІтучного інтелекту та робототехніки
ORCID: https://orcid.org/0000-0002-4859-304X; lobachev@op.edu.ua
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

Розпізнавання людських дій на основі скелету страждає від низької зовнішньої валідності, оскільки популярні набори даних використовують несумісні схеми суглобів (наприклад, COCO-17, NTU-25/26), що призводить до необхідності спеціального ремапінгу, вилучення суглобів або використання декількох вхідних «голів» уваги, специфічних для набору даних. В даній роботі був представлений легкий уніфікатор діагностики моделей Schema-Align, який перетворює пози з довільних вихідних схем у фіксоване 21-суглобове представлення, використовуючи розріджене по рядках лінійне відображення, регуляризоване кінематичною доцільністю (обмеження довжини кісток і кута нахилу суглоба) і малопотужним тимчасовим залишком для інтерполяції дійсно відсутніх суглобів. Уніфікатор попередньо навчається без міток дій на змішаних потоках поз за допомогою послідовності циклів, часової передбачуваності та довірчо-зважених втрат, а потім підключається до будь-якої моделі HAR (GCN/MSG3D/CTR-GCN/Transformer) з незначною затримкою (<1%).

Уніфікатор був оцінений на наборах NTU RGB+D 60/120 (3D), Kinetics-Skeleton, HMDB51-/UCF101-Skeleton і PoseTrack (2D), охоплюючи схему, набір даних і зсуви детектора. У внутрішньодоменних протоколах перетворення ефективно виконується без втрат, що відповідає власній продуктивності магістралей. При передачі між наборами даних Schema-Align послідовно зменшує падіння точності порівняно з перехресним та щільним лінійним ремапами, а також перевершує специфічні для набору даних голови, особливо коли вихідна та цільова схеми розходяться (наприклад, COCO↔NTU). Окрім точності, метод покращує калібрування (нижчий ECE) та анатомічну правдоподібність (менше порушень кісток/кутів), що вказує на те, що фізично обгрунтоване перетворення дає більш надійні ознаки при зсуві.

Дослідження показують, що розрідженість верхніх k рядів (k=1-2) запобігає надмірному пристосуванню до схеми; залишковий інтерполятор допомагає оклюдованим або зашумленим детектором кадрам при мінімальних витратах на параметри; а видалення кінематичних втрат погіршує реалістичність і передачу. Завдяки єдиному тонкому матричному множенню і невеликому часовому модулю, Schema-Align забезпечує практичний, інтерпретований шлях до навчання – один раз, оцінюй – будь-де HAR.

Ключові слова: машинне навчання; глибинне навчання; комп’ютерний зір; розпізнавання дій; аналіз поз; відеоспостереження; уніфікація даних; перенос навчання