FOUNDATION MODEL FOR SKELETON-BASED HUMAN ACTION UNDERSTANDING

Hongsong Wang, Wanjiang Weng, Junbo Wang, Fang Zhao, Guo-Sen Xie, Xin Geng, Senior Member, IEEE, and Liang Wang, Fellow, IEEE

ĐẶT VẤN ĐỀ

Ứng dụng hiểu hành động con người dựa trên khung xương

- Robot
- Tương tác người - máy
- Môi trường ảo nhập vai
- Công nghệ hỗ trợ, phục hồi chức năng
- Phân tích thể thao
- ...

➔ Dữ liệu khung xương gọn nhẹ, tinh giản, mang hiệu quả về mặt tính toán,lợi thế về bảo mật quyền riêng tư.

ĐẶT VẤN ĐỀ

Thách thức

- Tổng quát hóa cho các hành vi mới (chưa có nhãn).
- Đòi hỏi dữ liệu huấn luyện lớn được gán nhãn.
- Rất dễ quá khớp trên dữ liệu nhỏ, thiếu tính đa dạng.

➔ Mô hình hóa chuỗi bị che khuất và học tương phản.

ĐỀ XUẤT

Unified Skeleton-based Dense Representation Learning (USDRL)

- ✓ Multi-Grained Feature Decorrelation (MG-FD)
- ✓ Dense Spatio-Temporal Encoder (DSTE):
  - Convolutional Attention (CA)
  - Dense Shift Attention (DSA)
- ✓ Multi-Perspective Consistency Training (MPCT)
  - Multi-View Training
  - Multi-Modal Training

KIẾN TRÚC TỔNG QUÁT

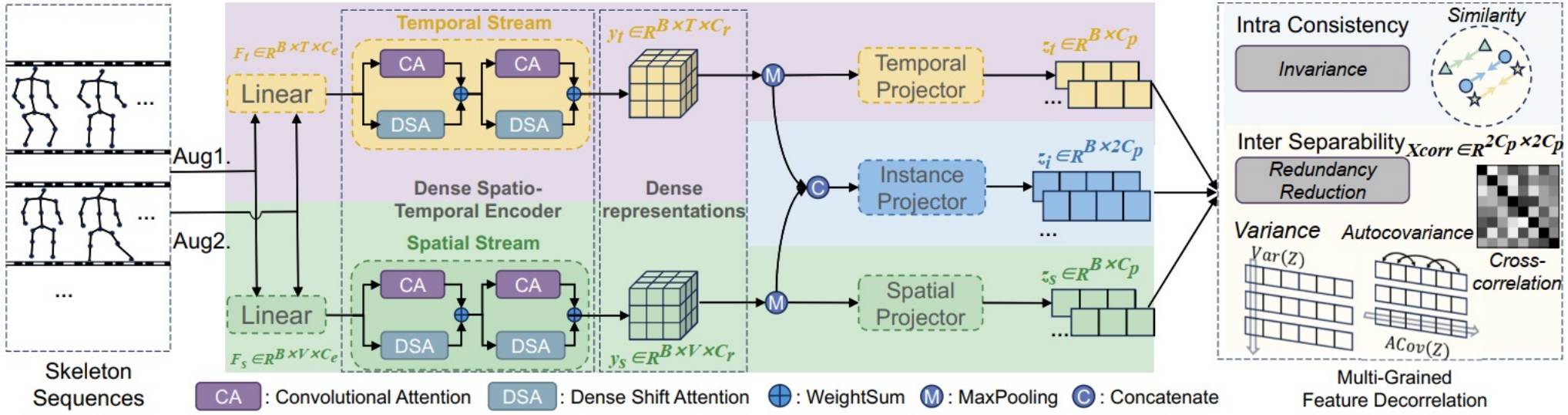


Fig. 3: The proposed Unified Skeleton-based Dense Representation Learning (USDRL) framework. USDRL incorporates a two-stream architecture with the Dense Spatio-Temporal Encoder (DSTE). The DSTE processes skeleton sequences to derive dense representations, which are further refined through MaxPooling and concatenation to generate condensed vectors. The Multi-Grained Feature Decorrelation is devised to mitigate model collapse and guarantee both intra-sample consistency and inter-sample separability. To further enhance robustness across different viewpoints and facilitate multimodal learning, a multi-perspective consistency training strategy is used during training.

KIẾN TRÚC DSTE

$\mathbf{y} = \alpha\ \mathrm{CA}\ (\mathbf{F}) + \beta\ \mathrm{DSA}\ (\mathbf{F}\ )$

trong đó:

$\mathbf{F}_h = \mathrm{ReLU}\ (W_1\mathbf{F}_1)W_2 + \mathbf{F}_1$

$\mathbf{F}_m = \mathbf{Mask} \odot \mathbf{F}_h + \mathbf{Mask} \odot \mathbf{F}$

$\mathbf{F}_d = \mathrm{FFN}(\mathrm{SA}(\mathbf{F}m)) + \mathrm{FFN}(\mathrm{SA}(\mathbf{F}))$

$\mathbf{F}_g = \mathrm{FFN}(\mathrm{SA}(\mathrm{Conv}(\mathbf{F}) + \mathbf{F}))$

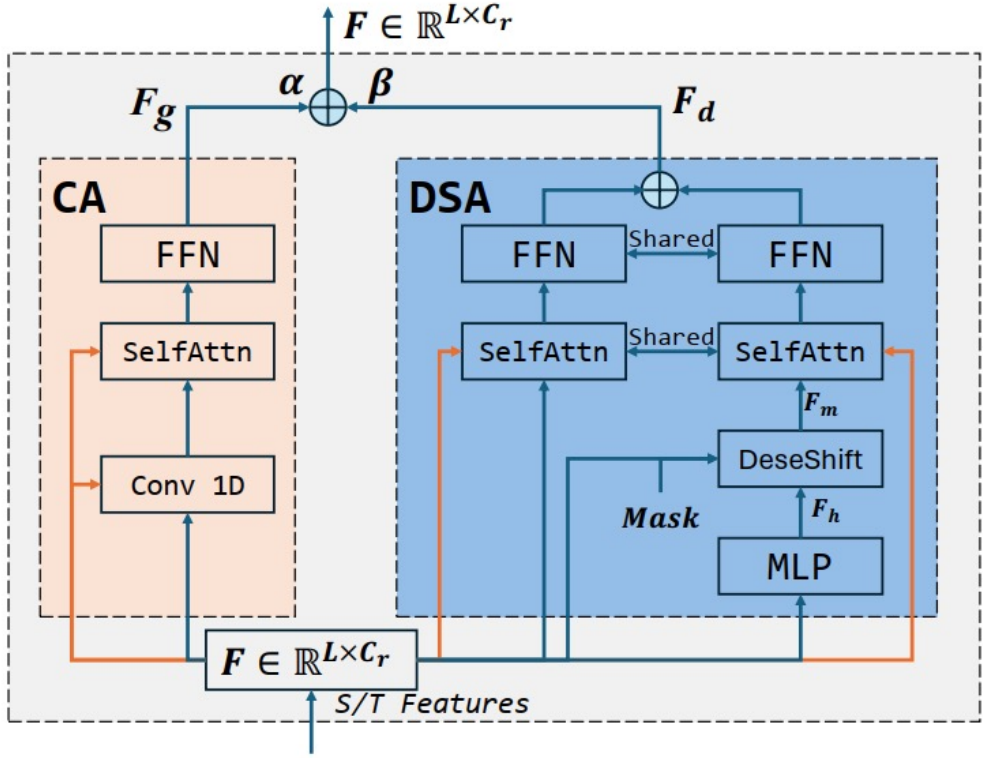


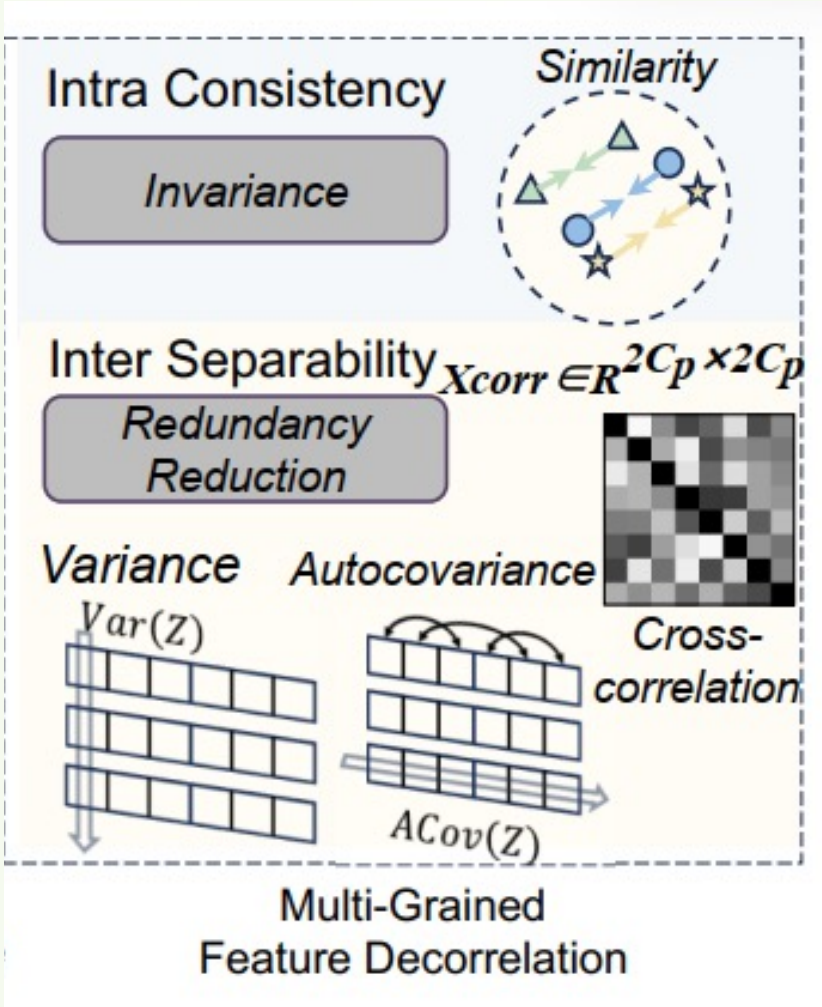
Fig. 4: The fundamental layer of the Dense Spatio-Temporal Encoder. It consists of the Convolutional Attention (CA) and Dense Shift Attention (DSA) modules, where $\oplus$ represents a weighted sum operation.

KIẾN TRÚC MG-FD

$$L = L_{fd}(\boldsymbol{Z}) + \tau(L_{fd}(\boldsymbol{Z}_s) + L_{fd}(\boldsymbol{Z}_t))$$

trong đó:

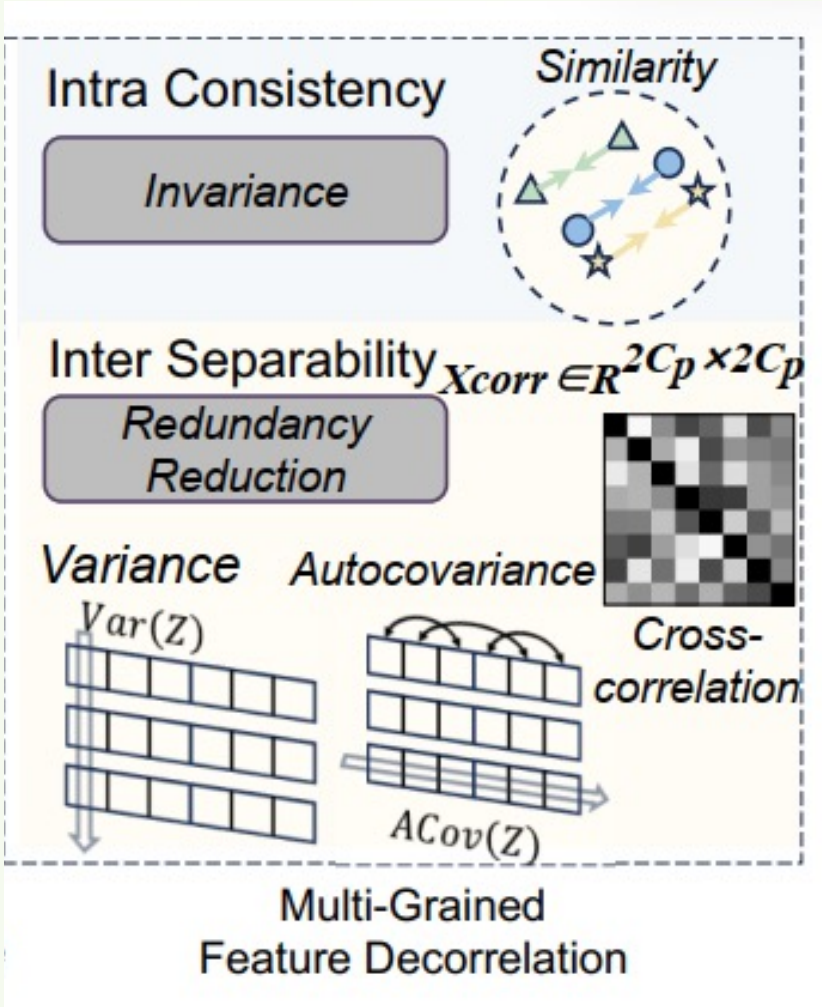
$$L_{fd}(\boldsymbol{Z}) = L_{con}(\boldsymbol{Z}) + L_{sep}(\boldsymbol{Z})$$



KIẾN TRÚC MG-FD

L_{con} hàm mất mát nội mẫu (inter-sample consistency) đảm bảo các phiên bản tăng cường dữ liệu khác nhau của cùng một mẫu vẫn giữ được ý nghĩa ngữ nghĩa giống nhau.

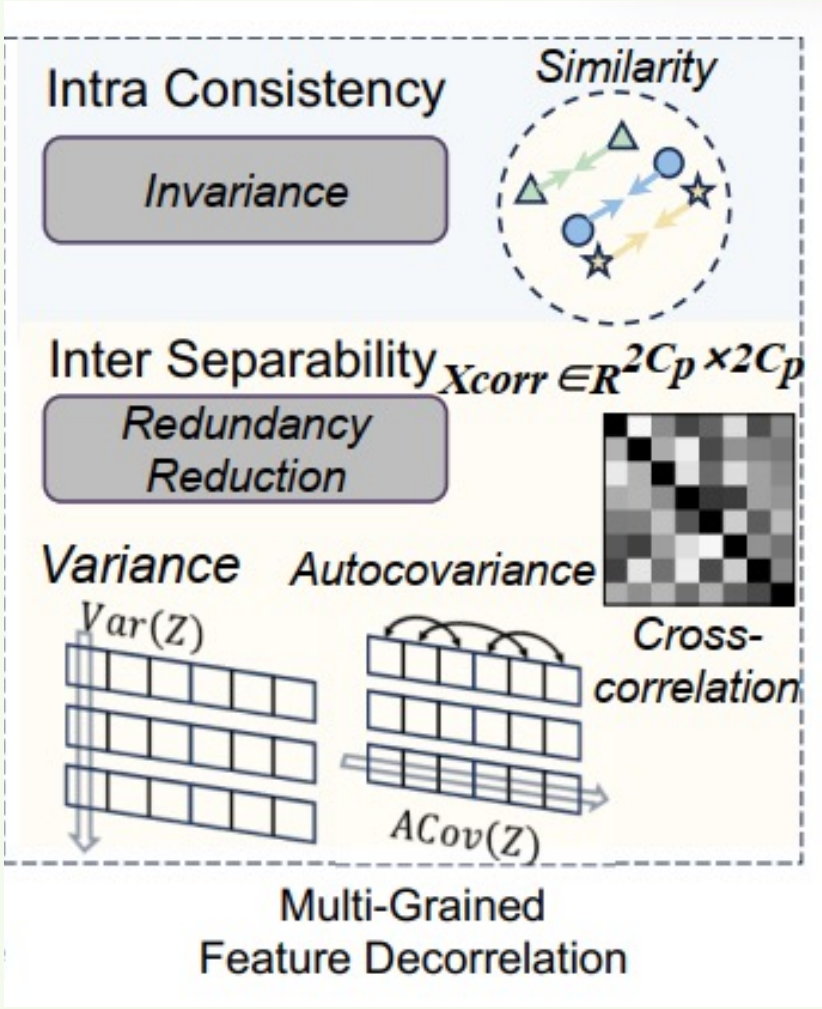
$$\mathcal{L}_{con} = \frac{1}{K}\sum_{a=1}^{K}\left(\kappa\|z_a - \bar{z}\|_2 + \eta \sum_{b=1|b\neq a}^{K} tr(I - \hat{z}_a{}^T \hat{z}_b)\right)$$



KIẾN TRÚC MG-FD

L_{sep} là hàm mất mát phân tách giữa các mẫu (inter-sample separability) để tránh hiện tượng sụp đổ mô hình và giảm dư thừa, đảm bảo các biểu diễn đặc trưng không bị tương quan quá mức.

$$\mathcal{L}_{sep} = \sum_{a=1}^{K} (\mu V(Z_a) + AC(Z_a) + \lambda \sum_{b=a+1}^{K} XC(Z_a, Z_b))$$



CHIẾN LƯỢC MPCT

- Multi-Modal Training: mô hình khai thác đồng thời các phương thức (modalities) khác nhau của khung xương như **khớp (joints)**, **xương (bones)** và **chuyển động (motion),** sử dụng cơ chế dung hợp sớm bằng trung bình cộng trước khi đưa vào DSTE.

$$y = DSTE(Fusion(F_{\text{joint}}, F_{\text{bone}}, F_{\text{motion}}))$$

- Multi-Vew Training: can thiệp vào cách mô hình nhìn nhận tính nhất quán của hành động thông qua các góc nhìn camera, ghi lại từ **các góc nhìn (viewpoints) khác nhau** để làm cặp mẫu tương đồng và sau đó qua các phép biến đổi (augmentation) riêng biệt.

THỰC NGHIỆM

- Datasets: NTU-60, NTU-120, PKUMMD, UAV-Human
- gap = 4
- α, β, τ đều là 0.5
- $\kappa = 5$, $\eta = 0{,}0005$, $\mu = 1$, $\lambda = 0{,}001$
- Sử dụng Adam optimizer
- Batch size: 324
- Epochs: 450 (NTU) và 1200 (PKU-MMD II)

KẾT QUẢ DỰ ĐOÁN THÔ (COARSE PREDICTION)

TABLE 1: Comparison of unsupervised action recognition results. J: Joint, M: Motion, B: Bone.

Method	Publisher	Modality	NTU-60		NTU-120		PKU-MMD II
			x-sub	x-view	x-sub	x-set	x-sub
Masked Sequence Modeling							
MS^{2}L [55]	ACM MM'20	J	52.6	-	-	-	27.6
3s-Colorization [56]	ICCV'21	J	75.2	83.1	-	-	-
GL-Transformer [57]	ECCV'22	J	76.3	83.8	66	68.7	-
Masked Colorization [13]	TPAMI'23	J	79.1	87.2	69.2	70.8	49.8
PCM^3 [27]	ACM MM'23	J	83.9	90.4	76.5	77.5	51.5
SkeletonMAE [58]	ICMEW'23	J	74.8	77.7	72.5	73.5	36.1
MAMP [12]	ICCV'23	J	84.9	89.1	78.6	79.1	53.8
MacDiff [59]	ECCV'24	J	86.4	91.0	79.4	80.2	-
MMFR [60]	TCSVT'24	J	84.2	89.5	77.1	78.8	54.4
Contrastive Learning							
AimCLR [15]	AAAI'22	J	74.3	79.7	63.4	63.4	38,5
CMD [22]	ECCV'22	J	79.8	86.9	70.3	71.5	43.0
CPM [14]	ECCV'22	J	78.7	84.9	68.7	69.6	48.3
PSTL [61]	AAAI'23	J	77.3	81.8	66.2	67.7	49.3
HaLP [62]	CVPR'23	J	79.7	86.8	71.1	72.2	43.5
HiCo [63]	AAAI'23	J	81.1	88.6	72.8	74.1	49.4
2s-DMMG [64]	TIP'23	J+M	84.2	89.3	72.7	72.4	58.8
Skeleton-logoCLR [65]	TCSVT'24	J	82.4	87.2	72.8	73.5	54.7
KTCL [66]	TMM'24	J	82.4	89.4	74.4	74.5	55.5
SCD-Net [67]	AAAI'24	J	86.6	91.7	76.9	80.1	54.0
ViA [68]	IJCV'24	J+M	78.1	85.8	69.2	66.9	-
IGM [69]	ECCV'24	J	86.2	91.2	80.0	81.4	-
Feature Decorrelation							
HYSP [70]	ICLR'23	J	78.2	82.6	61.8	64.6	-
UmURL [38]	ACM MM'23	J	82.3	89.8	73.5	74.3	52.1
UmURL [38]	ACM MM'23	J+M+B	84.2	90.9	75.2	76.3	52.6
Heter-Skeleton [41]	CVPR'25	J	80.2	88.0	70.7	73.5	47.7
USDRL (STTR)	Preliminary work	J	84.2	90.8	76.0	76.9	51.8
USDRL (DSTE)	Preliminary work	J	85.2	91.7	76.6	78.1	54.4
USDRL (STTR)	This work	J+M+B	85.8	91.8	77.5	78.8	54.7
3s-ensemble							
3s-CMD [22]	ECCV'22	J+M+B	84.1	90.9	74.7	76.1	52.6
3s-CPM [14]	ECCV'22	J+M+B	83.2	87.0	73.0	74.0	51.5
3s-HiCLR [23]	AAAI'23	J+M+B	80.4	85.5	70.0	70.4	-
3s-SkeAttnCLR [71]	IJCAI'23	J+M+B	82.0	86.5	77.1	80.0	55.5
3s-SSRL [72]	TCSVT'23	J+M+B	81.6	85.1	69.2	71.5	50.2
3s-PCM^3 [27]	ACM MM'23	J+M+B	87.4	93.1	80.0	81.2	58.2
3s-ActCLR [73]	CVPR'23	J+M+B	84.3	88.8	74.3	75.7	-
3s-RVTCLR+ [74]	ICCV'23	J+M+B	79.7	84.6	68.0	68.9	-
3s-PSTL [61]	AAAI'23	J+M+B	79.1	82.6	69.2	70.3	52.3
3s-CSTCN [75]	TMM'23	J+M+B	85.8	92.0	77.5	78.5	53.9
3s-UmURL [38]	ACM MM'23	J+M+B	84.4	91.4	75.8	77.2	54.3
3s-USDRL (DSTE)	This work	J+M+B	**87.1**	**93.2**	**79.3**	**80.6**	**59.7**

TABLE 2: Comparison of performance under semi-supervised evaluation protocol on the NTU60 dataset.

Method	x-sub		x-view	
	1 % data	10 % data	1 % data	10 % data
MS^{2}L [55]	33.1	65.2	-	-
ISC [77]	35.7	65.9	38.1	72.5
HiCLR [23]	51.1	74.6	50.9	79.6
CMD [22]	50.6	75.4	53	80.2
PCM^3 [27]	53.8	77.1	53.1	82.8
HiCo [63]	54.4	73.0	54.8	78.3
Heter-Skeleton [41]	55.0	76.3	55.0	79.1
USDRL (STTR)	55.0	76.1	59.1	82.0
USDRL (DSTE)	**57.3**	**80.2**	**60.7**	**84.0**

TABLE 3: Results of 2D skeleton-based action recognition on the UAV-Human dataset. S and U denote supervised and unsupervised training, respectively.

Method	Modality	Training	CS-v1	CS-v2
ST-GCN [6]	J	S	30.2	56.1
2s-AGCN [7]	J+B	S	34.8	66.7
HARD-Net [78]	J	S	37.0	-
Shift-GCN [8]	J	S	38.0	67.0
LLM-AR [40]	J	S	46.3	-
USDRL (STTR)	J	U	31.7	50.2
USDRL (DSTE)	J	U	36.3	60.8

KẾT QUẢ DỰ ĐOÁN DÀY ĐẶC (DENSE PREDICTION)

TABLE 5: Comparsion of action detection results on PKU-MMD I xsub benchmark with an overlap ratio of 0.5.

Method	mAP_a (%)	mAP_v (%)
MS^2L [55]	50.9	49.1
CRRL [79]	52.8	50.5
ISC [77]	55.1	54.2
CMD [22]	59.4	59.2
PCM^3 [27]	61.8	61.3
USDRL (STTR)	66.1	65.9
USDRL (DSTE)	**75.7**	**74.9**

TABLE 6: Comparsion of action prediction results on the NTU-60 dataset.

Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DeepSCN [80]	16.8	21.5	30.5	39.9	48.7	54.6	58.2	60.2	60.0	58.6
MSRNN [81]	15.2	20.3	29.5	41.4	51.6	59.2	63.9	67.4	68.9	69.2
P-TSL [36]	27.8	35.8	46.3	58.5	67.4	73.9	77.6	80.1	81.5	82.0
USDRL (STTR)	24.7	36.5	54	65.7	72.8	76.9	80.3	82.4	83.7	84.2
USDRL (DSTE)	25.5	36.8	54.8	66.7	73.6	77.9	81.4	83.6	84.5	85.2

TABLE 7: Comparison of action segmentation results on the PKU-MMD II dataset.

Method	Acc	Edit	F10	F25	F50
ST-GCN [6]	64.9	-	-	-	15.5
MS-TCN [82]	65.5	-	-	-	46.3
ETSN [83]	68.4	67.1	70.4	65.5	52.0
CTC [29]	69.2	-	69.9	66.4	53.8
DeST [50]	67.6	66.3	71.7	68	55.5
USDRL (STTR)	68.7	67.5	70.9	67.3	56.2

KẾT QUẢ DỰ ĐOÁN CHUYỂN ĐỔI (TRANSFERRED PREDICTION)

TABLE 8: Comparisons of transferred action recognition results on the PKU-MMD II.

Method	Transfer to PKU-MMD II	
	NTU-60	NTU-120
MS^{2}L [55]	45.8	-
ISC [77]	45.9	-
SCD-Net [67]	56.3	-
HiCo [63]	56.3	55.4
CMD [22]	56.0	57.0
UmURL [38]	58.2	57.6
USDRL	57.2	**58.3**

TABLE 9: Comparisons of transferred action retrieval results on the PKU-MMD II.

Method	Transfer to PKU-MMD II			
	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-set
UmURL [38]	43.4	42.8	42.5	43.0
USDRL	44.4	44.7	44.0	43.8

CẢM ƠN THẦY & CÁC ANH CHỊ ĐÃ LẮNG NGHE!