Idempotent Unsupervised Representation Learning for Skeleton-Based Action Recognition

Lilang Lin, Lehong Wu, Jiahang Zhang, Jiaying Liu | Wangxuan Institute of Computer Technology, Peking University

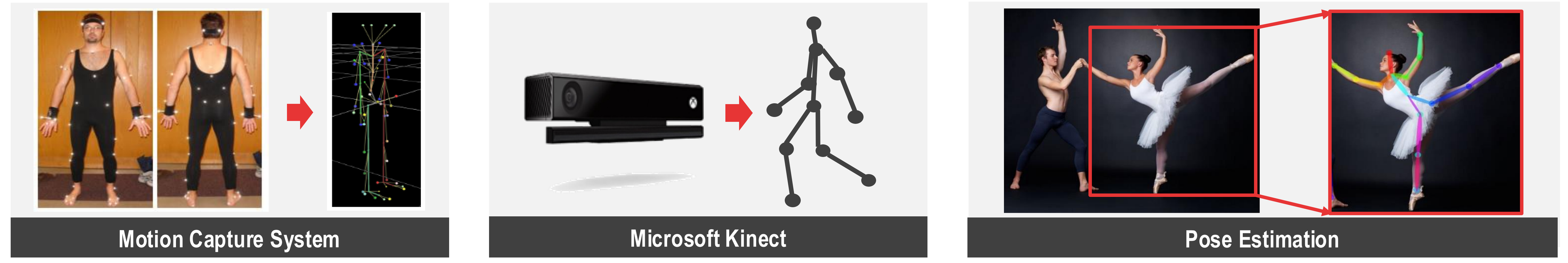
EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO

Introduction

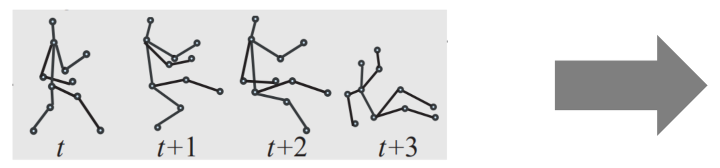
Background: unsupervised skeleton-based action recognition

- Skeletons represent human joints using 3D coordinate locations



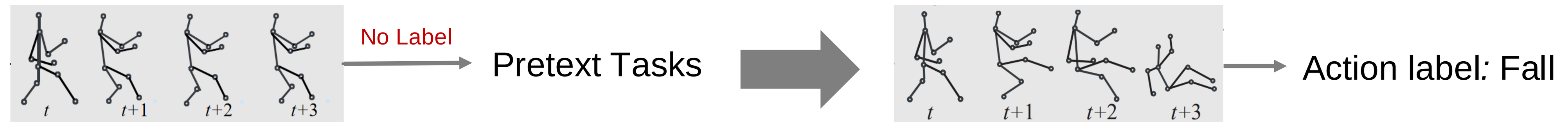
- Supervised learning → self-supervised learning

■ Supervised learning



Action label: Fall

■ Self-supervised learning



Self-Supervised Pretrain

Supervised Finetune

Challenges: gaps between generative models and contrastive learning

- Generative models preserve too much appearance information
- Contrastive learning result in a significant detail information loss

Self-Conditional Generative Models as Maximum Entropy Coding

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\mathcal{D}(g(\mathbf{z}),\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}|\mathbf{x}}}[-\log p(\mathbf{x}|\mathbf{z})]] = H(\mathbf{x}|\mathbf{z})$$

$$I(\mathbf{z};\mathbf{x}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{z}) = H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x})$$

$$L = \left(\frac{m+d}{2}\right)\log\det\left(\mathbf{I} + \frac{d}{m\varepsilon^2}\mathbf{Z}^T\mathbf{Z}\right)$$

$$L = \mathrm{Tr}\left(\mu\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}(\lambda\mathbf{Z}^T\mathbf{Z})^n\right)$$

Method

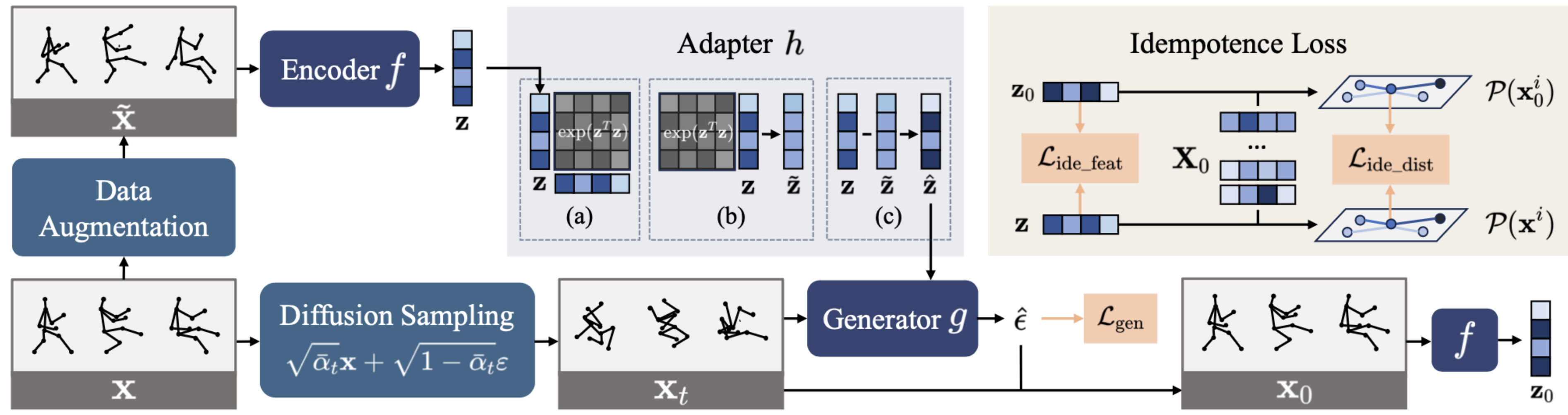
Idempotent Generative Models as Spectral Contrastive Learning

- Idempotent loss & Spectral contrastive learning

$$\mathcal{L}_{\text{ide}} = \|f(\hat{\mathbf{x}}) - \mathbf{z}\|^2 = -2f(\hat{\mathbf{x}})^T f(\mathbf{x})$$

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{ide}} - L &= -2\sum_{\mathbf{x},\hat{\mathbf{x}}} p(\mathbf{x},\hat{\mathbf{x}}) f(\hat{\mathbf{x}}_i)^T f(\mathbf{x}_i) + \sum_{\mathbf{x},\mathbf{x}'} p(\mathbf{x})p(\mathbf{x}')\left(f(\mathbf{x})^T f(\mathbf{x}')\right)^2 + \mathbf{R} \\ &= -2\mathbb{E}_{(\mathbf{x},\hat{\mathbf{x}})\sim p(\mathbf{x},\hat{\mathbf{x}})}\left[f(\hat{\mathbf{x}})^T f(\mathbf{x})\right] + \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim p(\mathbf{x})p(\mathbf{x}')}\left[\left(f(\mathbf{x})^T f(\mathbf{x}')\right)^2\right] + \mathbf{R} \\ &= -2\mathrm{Tr}\left(\mathbf{F}\mathbf{A}\mathbf{F}^T\right) + \mathrm{Tr}\left(\left(\mathbf{F}^T\mathbf{F}\right)^2\right) + \mathbf{R} = 2\mathrm{Tr}\left(\mathbf{F}\mathbf{L}\mathbf{F}^T\right) + \mathrm{Tr}\left(\left(\mathbf{F}^T\mathbf{F}\right)^2\right) + \mathbf{R} + \text{const} \\ &= \|\mathbf{A} - \mathbf{F}^T\mathbf{F}\|_F^2 + \mathbf{R} + \text{const},\end{aligned}$$

- Idempotent Diffusion Generation Model



- Noise prediction loss

$$\mathcal{L}_{\text{gen}} = \|g(\mathbf{x}_t, h(\mathbf{z}), t) - \varepsilon\|^2,$$
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x} + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$$

- Feature idempotency constraint

$$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}}g(\mathbf{x}_t, h(\mathbf{z}), t)\right)$$
$$\mathcal{L}_{\text{ide\_feat}} = -f(\mathbf{x})^T f(\mathbf{x}_0, \mathbf{z}_{t'}, t, t'),$$
$$\mathbf{z}_{t'} = \sqrt{\bar{\alpha}_{t'}}\mathbf{z} + \sqrt{1-\bar{\alpha}_{t'}}\varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$$

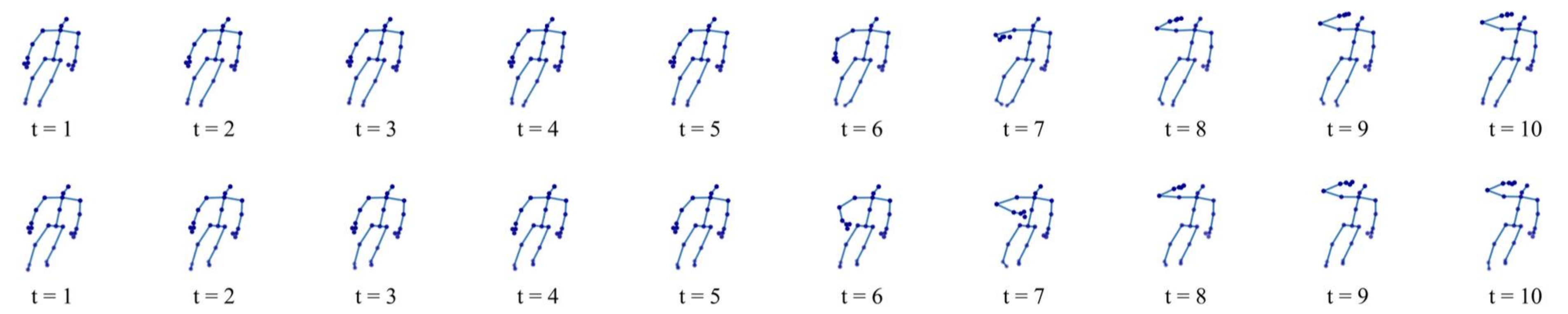
Experiments

Evaluation and comparison

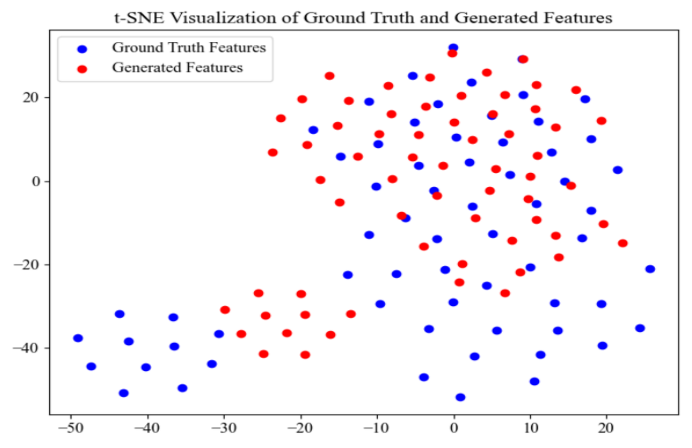
Models	Architecture	NTU 60		NTU 120	
		xview	xsub	xset	xsub
Contrastive Learning:					
3s-AimCLR [11]	GCN	83.4	77.8	66.7	67.9
3s-CPM [64]	GCN	84.9	78.7	69.6	68.7
3s-CMD [29]	GRU	90.9	84.1	76.1	74.7
GL-Transformer [18]	Transformer	83.8	76.3	68.7	66.0
3s-ActCLR [21]	GCN	88.8	84.3	75.7	74.3
Generative Learning:					
3s-Colorization [61]	DGCNN	87.2	79.1	70.8	69.2
SkeletonMAE [58]	GCN	77.7	74.8	73.5	72.5
MAMP [28]	Transformer	89.1	84.9	79.1	78.6
Contrative Learning & Generative Learning:					
CRRL [53]	GRU	73.8	67.6	57.0	56.2
PCM^3 [65]	GRU	90.4	83.9	77.5	76.3
IGM (Ours)	Transformer	**91.2**	**86.2**	**81.4**	**80.0**

Ablation Study

- Visualization of ground truth data and generated data



- Visualization of features



Project Website
For more resources and codes:
https://langlandslin.github.io/projects/IGM/

Team Website
Interested in our team STRUCT? Navigate to:
http://struct.wict.pku.edu.cn

If you have any questions, please contact: linlilang@pku.edu.cn