JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Decompose More and Aggregate Better: Two Closer Looks at Frequency Representation Learning for Human Motion Prediction

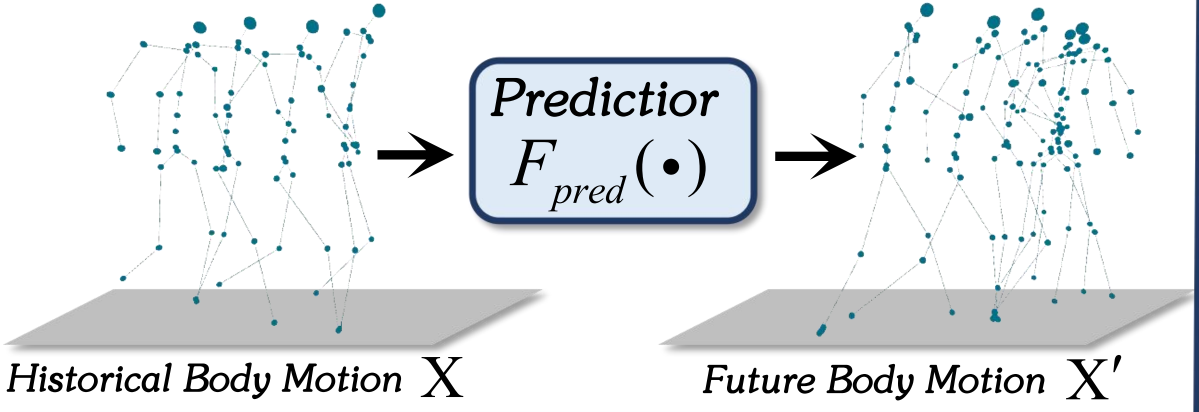
Xuehao Gao[1], Shaoyi Du[1], Yang Wu[2], Yang Yang[1]

[1]Xi'an Jiaotong University, [2]Tencent AI Lab

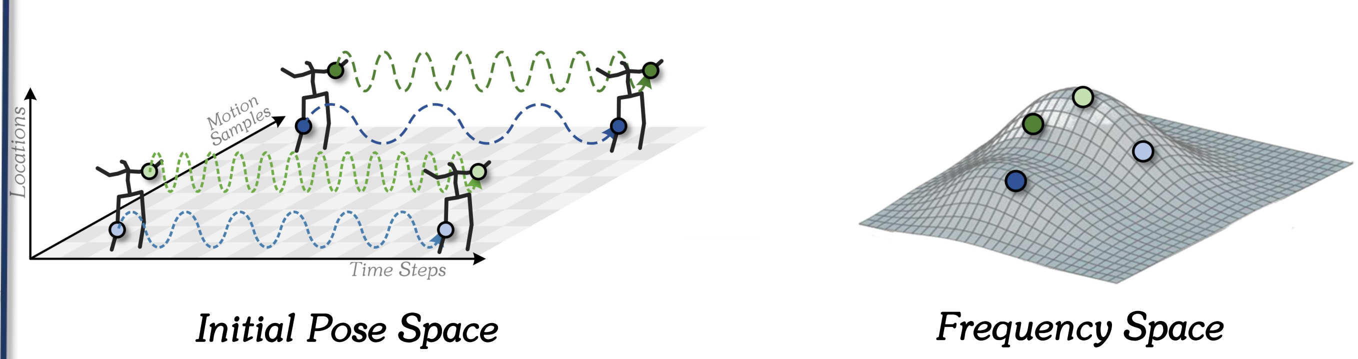
Part 1

Quick Preview

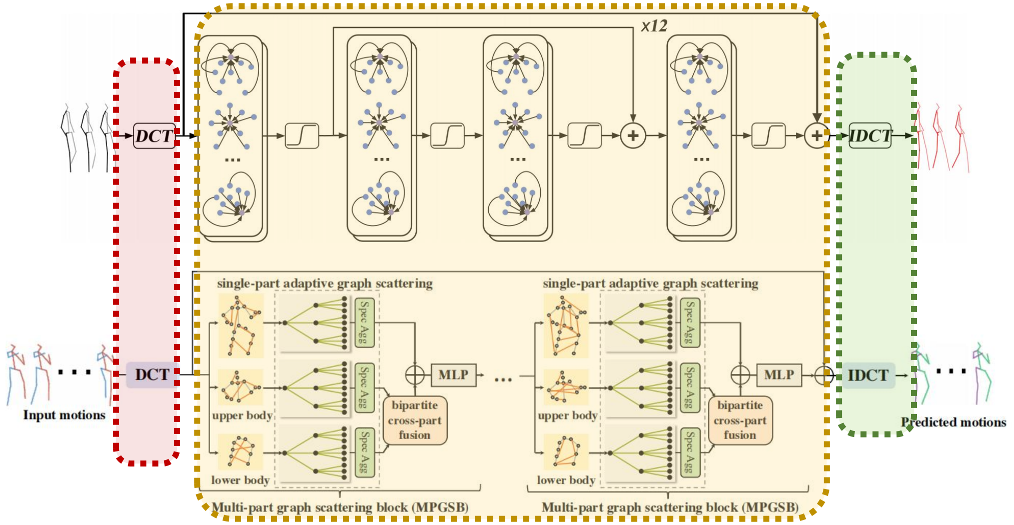
① Human Motion Prediction Task



② Frequency Representation Learning



③ Previous Predictive Methods



④ Paradigm Review

- *Others adopt:*

$$F_{pred}(\mathrm{X}) = F_{IDCT}(F_{enc}(F_{DCT}(\mathrm{X})))$$

- *We propose:*

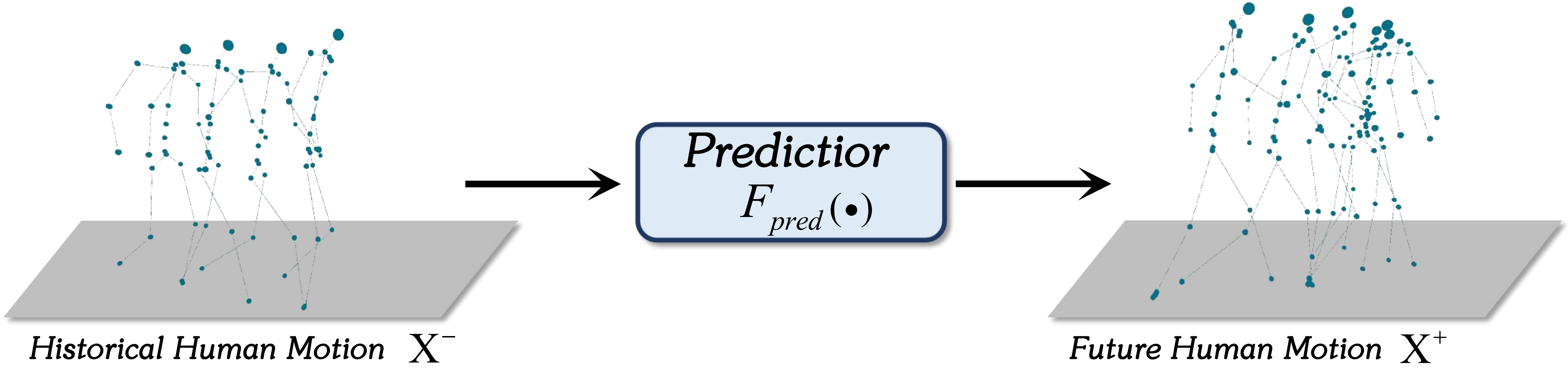
$$F_{pred}(\mathrm{X}) = F_{IDCT}\left(F_{enc}(\bar{\mathrm{X}}_1, \bar{\mathrm{X}}_2, \cdots, \bar{\mathrm{X}}_K)\right)$$

where $\bar{\mathrm{X}}_k = F_{filt}^{k}\left(F_{DCT}(\mathrm{X})\right)$

Part 2

Detailed Introduction

• Task



• Formulation

$$\mathrm{X}^{+} = F_{pred}(\mathrm{X}^{-})$$

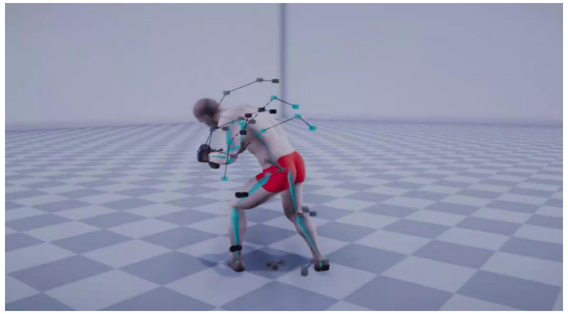
where

$\mathrm{X}^{-} \in \mathbb{R}^{T^{-} \times V \times 3}$: Past T^{-}–step 3D human poses.

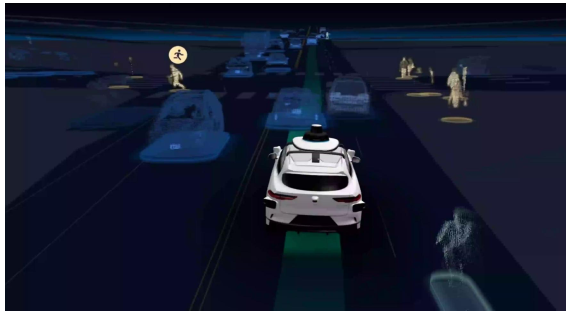
$\mathrm{X}^{+} \in \mathbb{R}^{T^{+} \times V \times 3}$: Next T^{+}–step 3D human poses.

V: Number of body joints for each skeleton.

• Application



Sports Analysis



Autonomous Driving

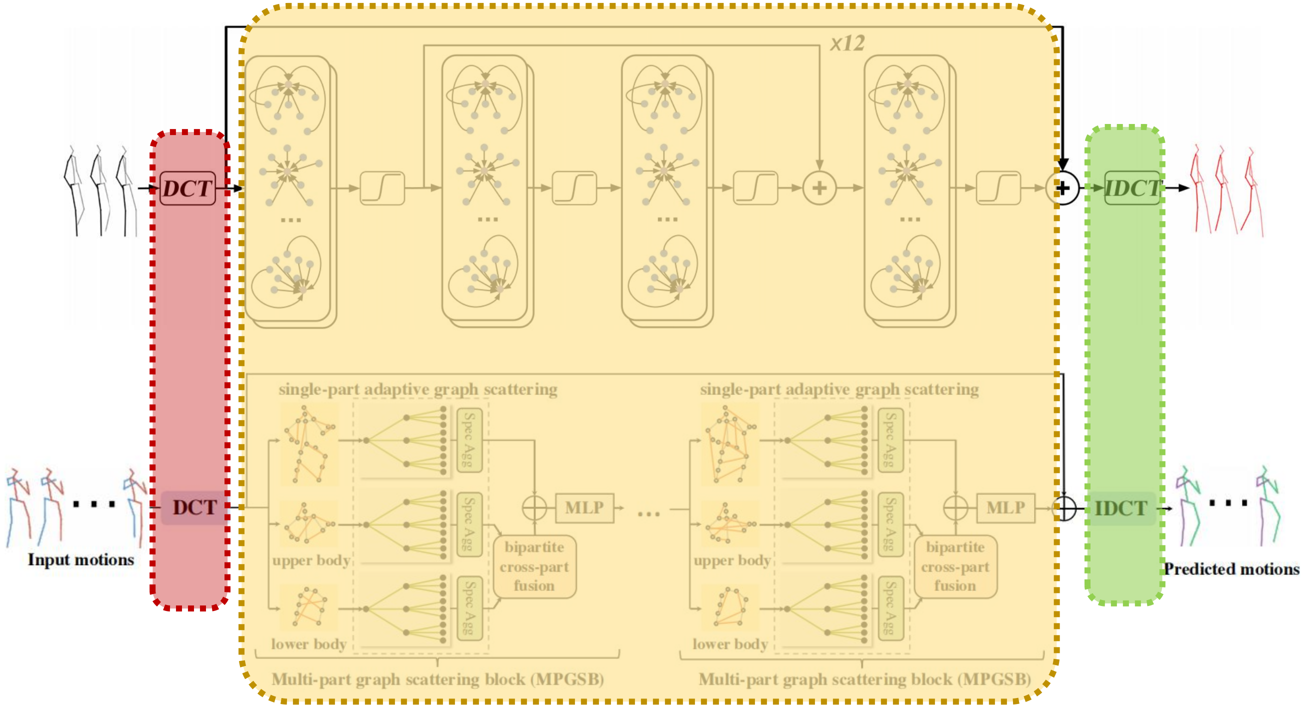
- *Previous Methods*

DCT: *Discrete Cosine Transform*

IDCT: *Inverse DCT*

➯*ICCV–19*[1]

➯*ECCV–22*[2]



. . .

- *Paradigm Review*

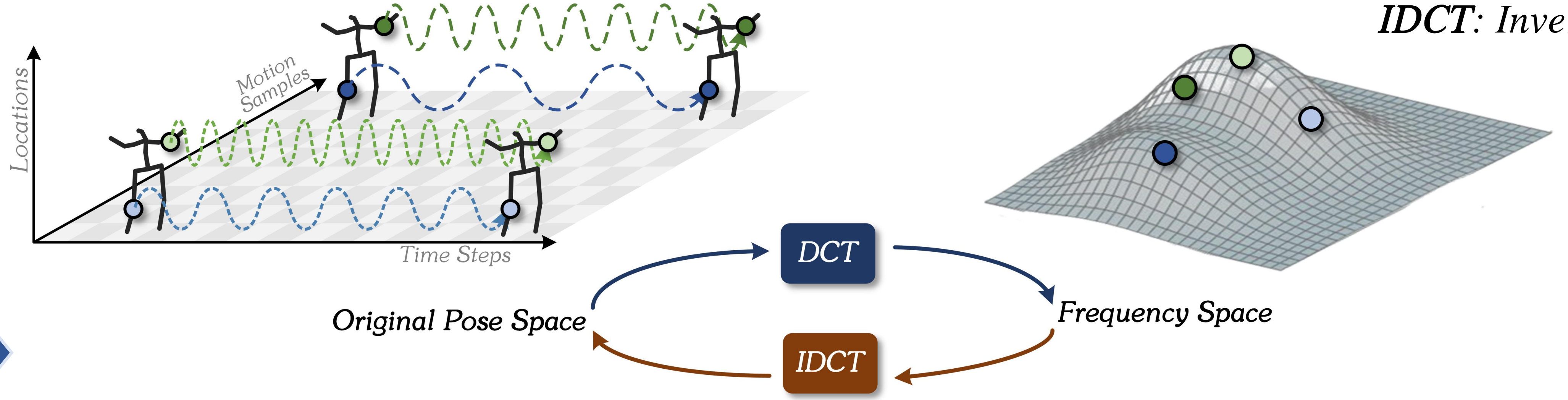
$$F_{pred}\left(X^{-}\right)=F_{IDCT}\left(F_{Enc}\left(F_{DCT}(X^{-})\right)\right)$$

[1] Mao et al., Learning Trajectory Dependencies for Human Motion Prediction. ICCV-2019.
[2] Li et al., Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. ECCV-2022.

I Introduction
II Previous Works
III Limitations
IV Our Method
V Results
VI Summary

DCT*: *Discrete Cosine Transform

IDCT*: *Inverse DCT



- ***Advantage:***

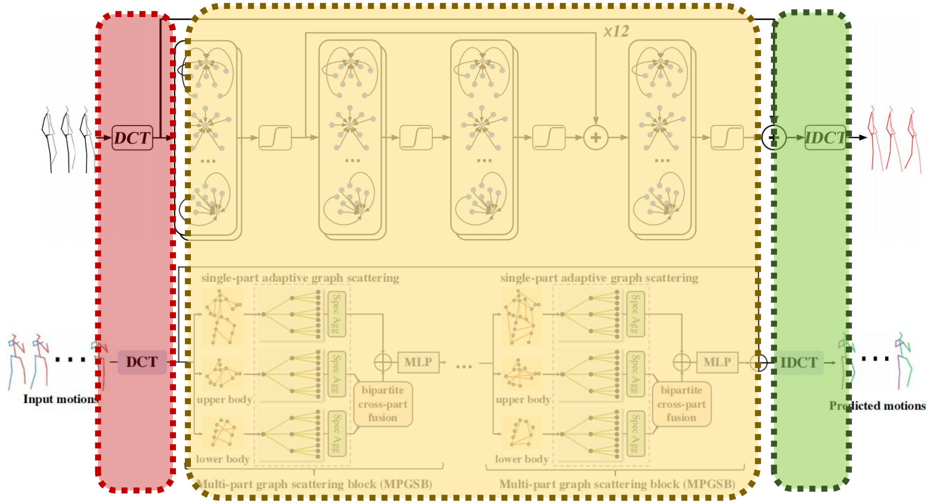
***Frequency space** encourages human motion presidction systems to focus on **trajectory−related cues** (e.g. temporal smoothness).*

» *robust to body shape perturbation*

» *robust to coordinate system shift*

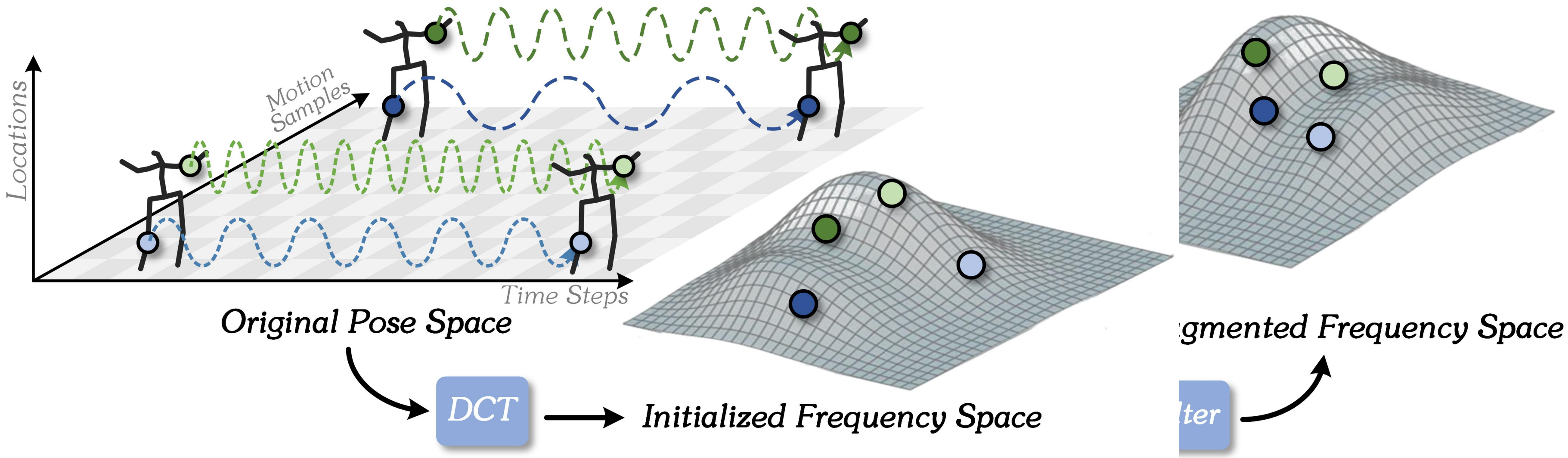
…

- ***Scheme:***



$$F_{pred}\left(X^{-}\right)=F_{IDCT}\left(F_{Enc}\left(F_{DCT}\left(X^{-}\right)\right)\right)$$

- *Diverse frequency distributions* bring *challenges* to *robust* human motion prediction



➪ *intra–sample difference*

Different body joints exhibit different frequency appearances

➪ *inter–sample difference*

Different personal motion styles in the same activity brings subtle intra–class bias to different data samples

- *Multi–view augmentation learning* can be developed into *a promising solution* for *robust* human motion prediction.

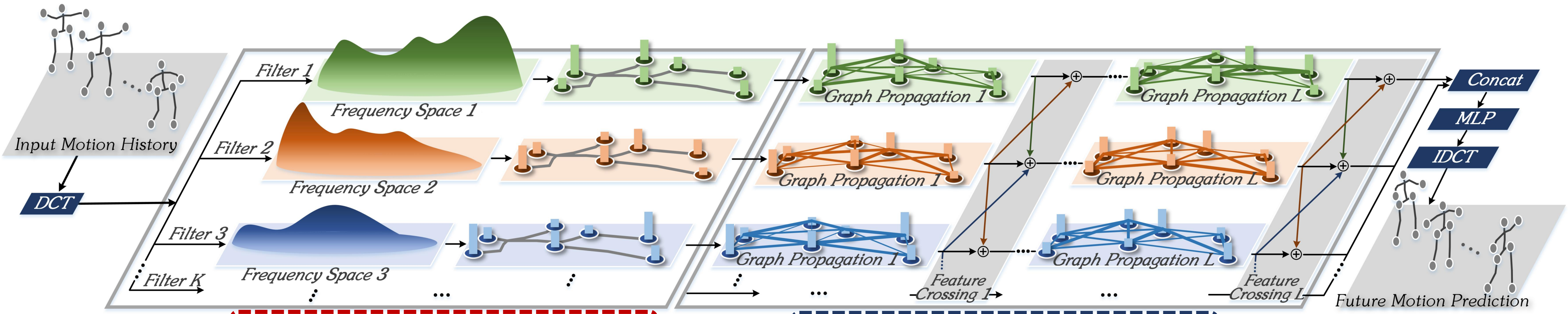
- *Toward Effective Frequency Representation Learning*

Two Closer Looks

- ◎ Decompose More
- ◎ Aggregate Better

Two Key Components

- ◎ Frequency Decomposition Unit
- ◎ Feature Aggregation Unit



- *Paradigm Review*

$$F_{pred}(X) = F_{IDCT}(F_{enc}(F_{DCT}(X)))$$

Conventional Scheme

$$F_{pred}(X) = F_{IDCT}\left(F_{enc}(\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_K)\right)$$

where $\bar{X}_k = F_{filt}^{k}(F_{DCT}(X))$

Proposed Scheme

Short-term Prediction on Human3.6M Dataset

walking, eating, smoking, discussion, direction, greeting, phoning, posing, purchases, sitting, sitting down, taking photo, posing, walking dog, walking together

scenarios	walking				eating				smoking				discussion			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [1]	17.3	30.7	54.6	65.2	11.0	21.4	36.2	43.9	9.0	17.6	32.1	40.3	17.3	34.8	61.0	69.8
MSR-GCN [2]	12.2	22.7	38.6	45.2	8.4	17.1	33.0	40.4	8.0	16.3	31.3	38.2	12.0	26.8	57.1	69.7
PGBIG [3]	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
SPGSN [4]	10.1	19.4	34.8	41.5	7.1	14.9	30.5	37.9	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1
Ours	**8.8**	**16.9**	**31.5**	**37.0**	**6.3**	**13.7**	**29.1**	**36.3**	**5.1**	**9.1**	**21.3**	**29.9**	**7.4**	**17.1**	**42.9**	**50.4**

scenarios	directions				greeting				phoning				posing			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [1]	13.1	24.6	64.7	81.9	23.3	50.3	107.3	132.1	12.5	25.8	48.1	58.3	15.3	29.3	71.5	96.7
MSR-GCN [2]	8.6	19.7	43.3	53.8	16.5	37.0	77.3	93.4	10.1	20.7	41.5	51.3	12.8	29.4	67.0	85.0
PGBIG [3]	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
SPGSN [4]	7.4	17.1	39.8	50.3	14.6	32.6	70.6	86.4	8.7	18.3	38.7	48.5	10.7	25.3	59.9	76.5
Ours	**6.6**	**16.4**	**39.6**	**50.1**	**13.0**	**30.7**	**63.1**	**78.2**	**7.8**	**17.2**	**37.5**	**47.3**	**7.5**	**19.3**	**47.1**	**62.0**

scenarios	purchases				sitting				sittingdown				takingphoto			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [1]	21.4	38.7	75.7	92.7	11.9	25.1	44.6	50.2	15.0	32.9	77.1	93.0	13.6	29.0	46.0	58.8
MSR-GCN [2]	14.8	32.4	66.1	79.6	10.5	22.0	46.3	57.8	16.1	31.6	62.5	76.8	9.9	21.0	44.6	56.3
PGBIG [3]	12.5	28.7	60.1	73.3	8.8	19.2	42.4	53.8	**13.9**	27.9	57.4	71.5	8.4	18.9	42.0	53.3
SPGSN [4]	12.8	28.6	61.0	74.4	9.3	19.4	42.3	53.6	14.2	27.7	56.8	70.7	8.7	18.9	41.5	52.7
Ours	**11.8**	**27.2**	**56.4**	**63.9**	**8.7**	**18.9**	**42.1**	**53.2**	**13.9**	**25.6**	**54.2**	**67.2**	**8.1**	**18.0**	**39.2**	**50.6**

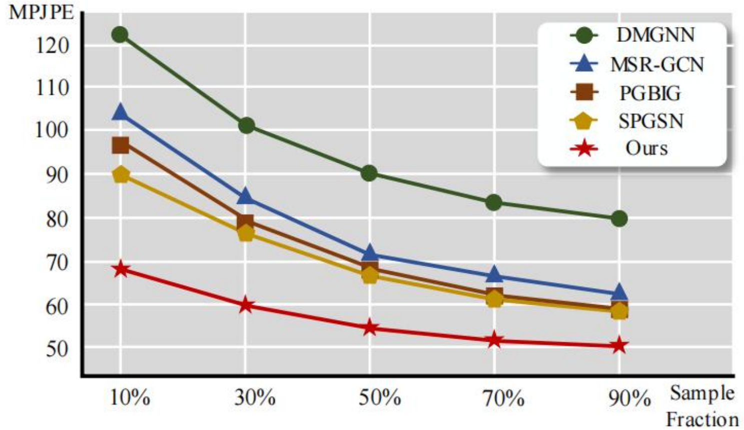
scenarios	waiting				walkingdog				walkingtogether				average			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [1]	12.2	24.2	59.6	77.5	47.1	93.3	160.1	171.2	14.3	26.7	50.1	63.2	17.0	33.6	65.9	79.7
MSR-GCN [2]	10.7	23.1	48.3	59.2	20.7	42.9	80.4	93.3	10.6	20.9	37.4	43.9	12.1	25.6	51.6	62.9
PGBIG [3]	8.9	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
SPGSN [4]	9.2	19.8	43.1	54.1	18.2	37.3	71.3	84.2	8.9	18.2	33.8	40.9	10.4	22.3	47.1	58.3
Ours	**8.2**	**18.4**	**41.3**	**52.1**	**14.5**	**32.7**	**63.8**	**76.0**	**7.4**	**15.2**	**30.0**	**36.4**	**9.3**	**19.7**	**41.0**	**51.1**

[1] Maosen Li, et al. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In CVPR, 2020.
[2] Lingwei Dang, et al. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In ICCV, 2021.
[3] Tiezheng Ma, et al. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In CVPR, 2022.
[4] Maosen Li, et al. Skeleton-parted graph scattering networks for 3d human motion prediction. In ECCV, 2022.

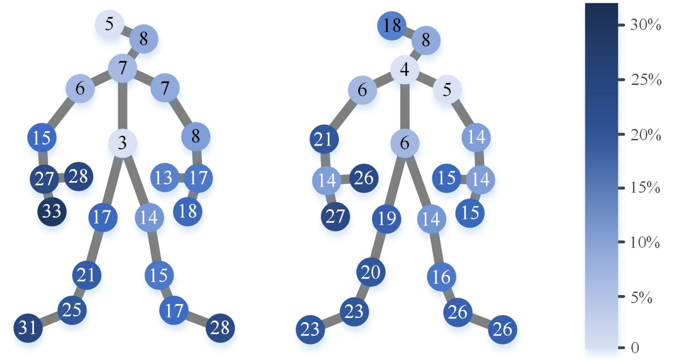
Long-term Prediction on Human3.6M Dataset

scenarios	walking		eating		smoking		discussion		directions		greeting		phoning		posing	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
DMGNN [1]	73.4	95.8	58.1	86.7	50.9	72.2	81.9	138.3	110.1	115.8	152.5	157.7	78.9	98.6	163.9	310.1
MSR-GCN [2]	52.7	63.0	52.5	77.1	49.5	71.6	88.6	117.6	71.2	100.6	116.3	147.2	68.3	104.4	116.3	174.3
PGBIG [3]	48.1	56.4	51.1	76.0	46.5	69.5	87.1	118.2	69.3	100.4	110.2	143.5	65.9	102.7	106.1	164.8
SPGSN [4]	46.9	53.6	49.8	73.4	46.7	68.6	89.7	118.6	70.1	100.5	111.0	143.2	66.7	102.5	110.3	165.4
Ours	**45.2**	**50.3**	**49.0**	**71.1**	**40.6**	**59.3**	**59.5**	**92.3**	**68.1**	**97.2**	**109.4**	**141.8**	**65.1**	**96.7**	**93.3**	**149.5**
scenarios	purchases		sitting		sittingdown		takingphoto		waiting		walkingdog		walkingtogether		average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
DMGNN [1]	118.6	153.8	60.1	104.9	122.1	168.8	91.6	120.7	106.0	136.7	194.0	182.3	83.4	115.9	103.0	137.2
MSR-GCN [2]	101.6	139.2	78.2	120.0	102.8	155.5	77.9	121.9	76.3	106.3	111.9	148.2	52.9	65.9	81.1	114.2
PGBIG [3]	95.3	133.3	74.4	116.1	96.7	147.8	74.3	118.6	72.2	103.4	104.7	139.8	51.9	64.3	76.9	110.3
SPGSN [4]	96.5	133.9	75.0	116.2	98.9	149.9	75.6	118.2	73.5	103.6	102.4	138.0	49.8	60.9	77.4	109.6
Ours	**94.8**	**130.7**	**72.3**	**114.5**	**94.3**	**145.3**	**72.2**	**116.1**	**70.0**	**101.2**	**94.6**	**123.1**	**47.9**	**58.7**	**67.2**	**100.3**

Performance Gains on Few-sample Prediction



Performance Gains on Each Joint

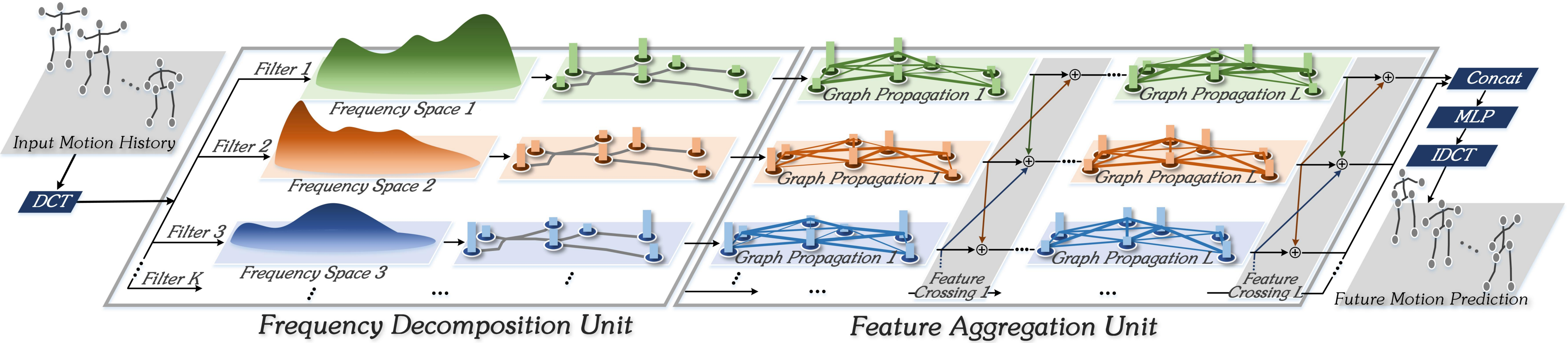


Ours vs PGBIG

Ours vs SPGSN

[1] Maosen Li, et al. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In CVPR, 2020.
[2] Lingwei Dang, et al. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In ICCV, 2021.
[3] Tiezheng Ma, et al. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In CVPR, 2022.
[4] Maosen Li, et al. Skeleton-parted graph scattering networks for 3d human motion prediction. In ECCV, 2022.

Decomposition –Aggregation Scheme



Contributions:

- *We propose a frequency decomposition unit (FDU) that develops multiple versatile filters to embed each body joint trajectory into multiple frequency spaces, enriching its encodings in the spectral domain.*

- *We design a feature aggregation unit (FAU) that deploys a series of intra–space and inter–space feature aggregation layers to extract comprehensive representations from multiple frequency spaces, collecting richer multi–view body features for robust motion prediction.*

Thanks For Watching!

Our Team:

Xuehao Gao

Shaoyi Du

Yang Wu

Yang Yang

西安交通大學
XI'AN JIAOTONG UNIVERSITY

Tencent AI Lab