FROM SPARSE TO DENSE: SPATIO-TEMPORAL FUSION FOR MULTI-VIEW 3D HUMAN POSE ESTIMATION WITH DENSEWARPER

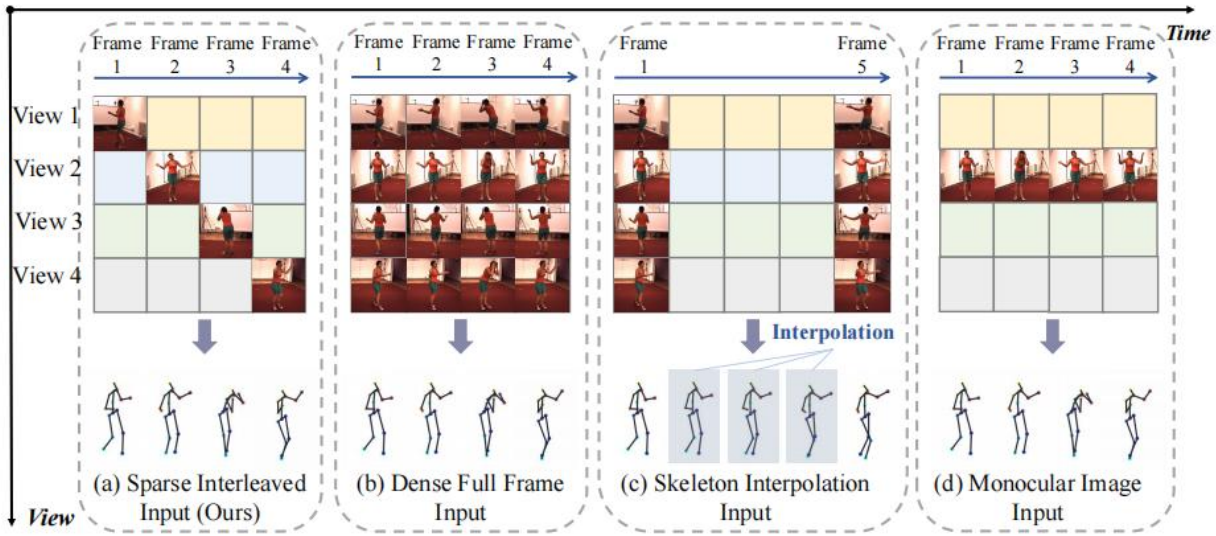
Ling Li*, Changjie Chen†, Yuyan Wang^, Jiaqing Lyu*, Kenglun Chang, Yiyun Chen††, and Zhidong Deng***

* Department of Computer Science, THUAI, BNRist, Tsinghua University, Beijing, China
†Dalian University of Technology, Dalian, China
**Apple, Beijing, China
††Hong Kong University of Science and Technology (Guang Zhou), Guang Zhou, China
^University of Manchester, Manchester, UK

Repoter: Ling Li

1. Introduction

➢ Sparse Interleaved Input



Common approaches for 3D multi-view pose estimation

➢ Motivation

a) Our proposed sparse interleaved input, where each view selects asingle temporally interleaved image as input to leverage spatio-temporal information across views fully;

b) illustration of dense, full-framemulti-view input;

c) keypoint interpolation input, which enhances the output frame rate;

d) illustration of single-view image input

1. Introduction

➢ Motivation

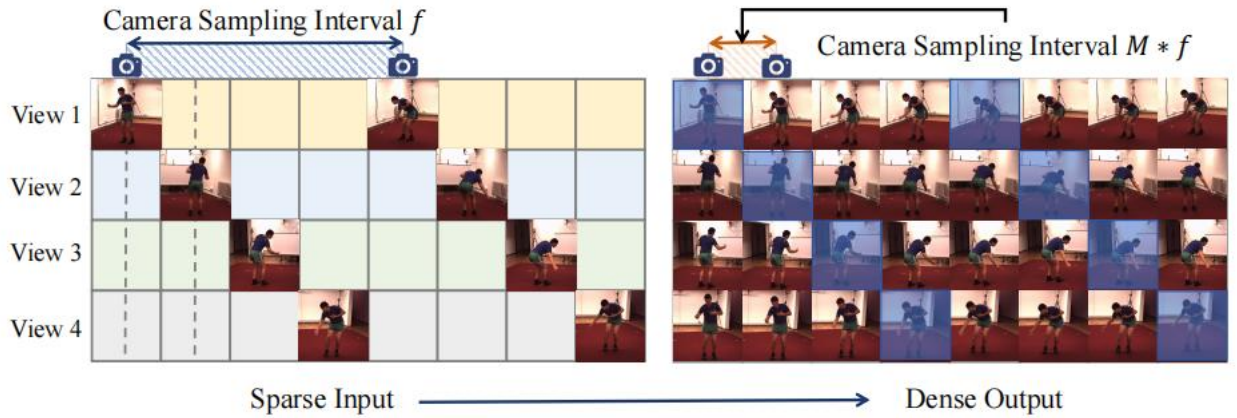


Illustration of frame rate enhancement through interleaved multi-view input.

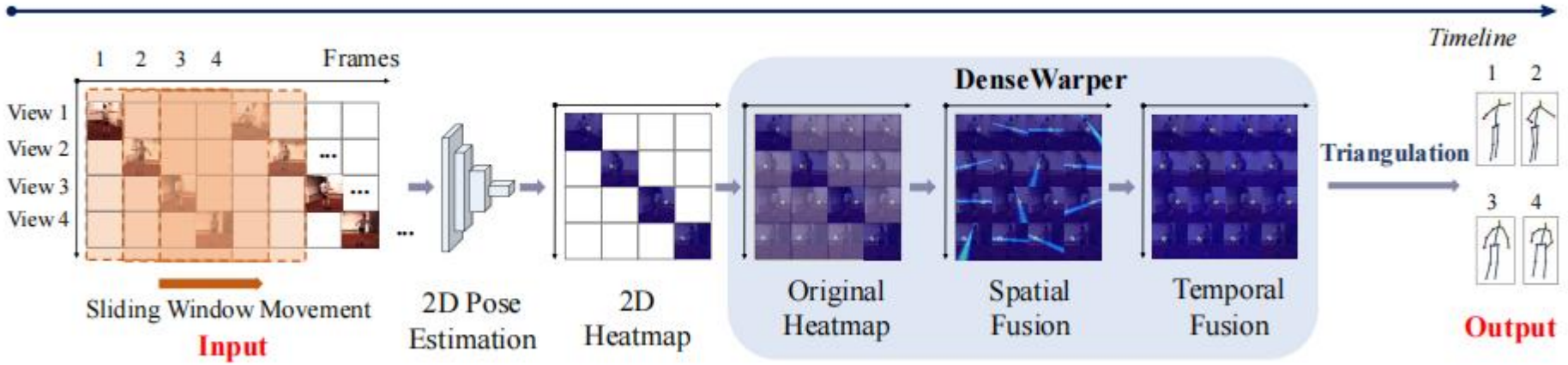
a) Synchronous Sampling Bottleneck——Traditional multi-view systems rely on synchronous frame capture, which is limited by the hardware frame rate (F) of individual cameras, failing to capture high-frequency motion.

b) Spatio-Temporal Information Waste—— Existing paradigms treat multi-view inputs as redundant spatial snapshots, overlooking the potential to use temporal phase differences (interleaved sampling) to reconstruct higher-resolution signals.

c) Real-Time Latency Trade-off—— Current 3D pose methods often require waiting for a full set of multi-view frames to be captured, creating a computational "dead time" that hinders low-latency, real-time applications.

➢ 1. DenseWarper

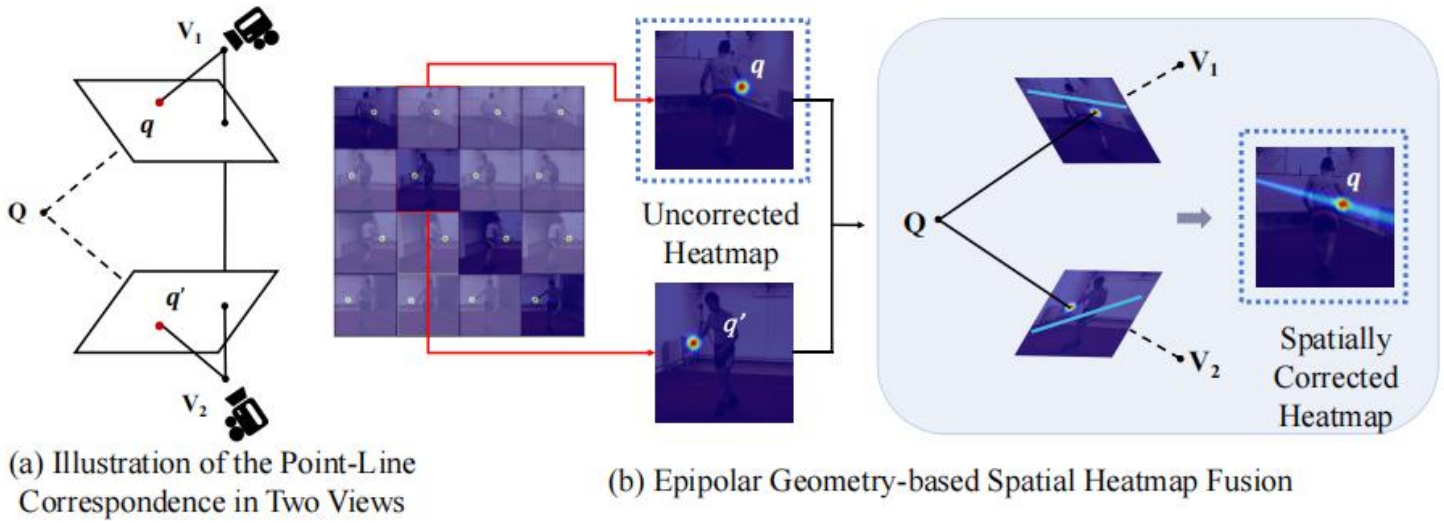
a) Model Overview



- A sliding window is used to sample sparse interleaved images, with a 2D pose estimation model generating initial heatmaps for each view.
- Missing information is filled to create uncorrected heatmaps. These are then spatially fused and corrected using an epipolar geometry-based method, yielding a spatially fused heatmap.
- Deformable convolutions are then applied for temporal fusion.
- Finally, the resulting spatiotemporally enriched heatmap is processed via triangulation to obtain accurate 3D keypoints.

➢ 1. DenseWarper

a) Epipolar geometry-based spatial heatmap fusion architecture



(a) Illustration of the Point-Line Correspondence in Two Views

(b) Epipolar Geometry-based Spatial Heatmap Fusion

- (a) Geometric interpretation of the point-line relationship for keypoints across different views;

- (b) the pipeline for spatial heatmap fusion based on epipolar geometry. For an inaccurate heatmap point q, we use accurate points q' from other views to correct it. First, we compute the corresponding epipolar lines in the other two heatmaps. Then, we identify the maximum response along the line associated with q and add these values to the original response at q in its heatmap. This process yields a spatially corrected heatmap. In the figure, non-diagonal heatmaps with masking represent the target heatmaps for correction, all processed according to this method

➢ 1. DenseWarper

a) The structure of the temporal fusion module

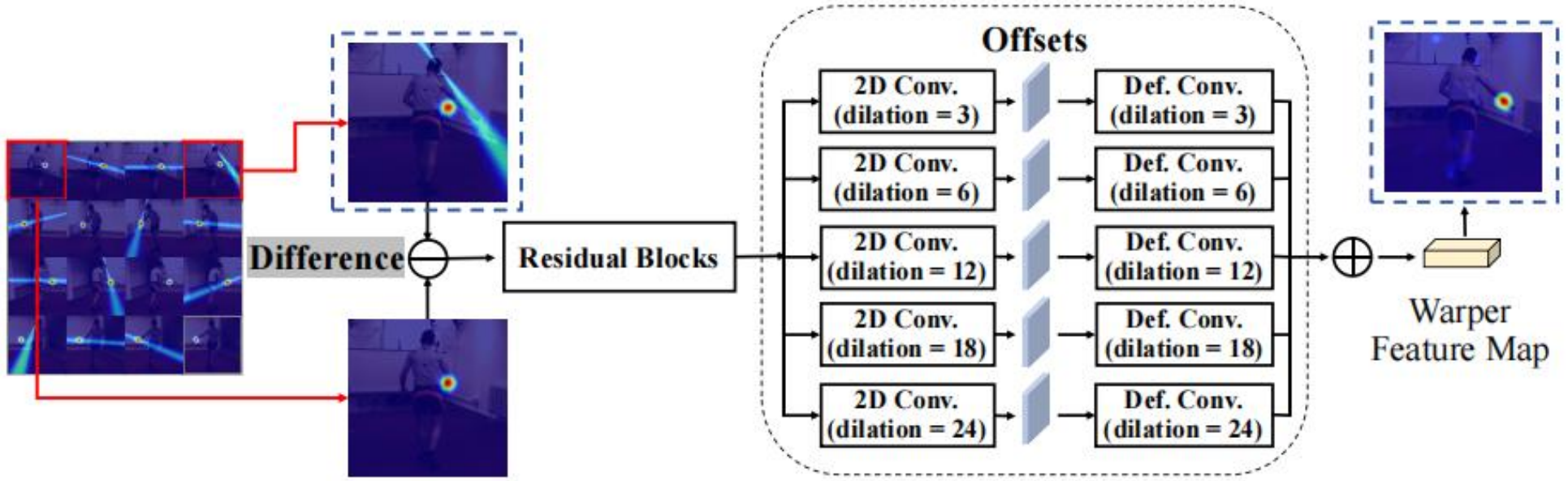


Figure 4: The structure of the temporal fusion module (Warper). We perform temporal correction based on the initial corrected heatmaps obtained from multi-view spatial fusion. For each heatmap in a target time frame (i.e., non-diagonal heatmaps in the figure), we compute its difference with the corresponding accurate heatmap in the same view (the diagonal heatmap) and apply a temporal pose feature learning module to correct the heatmaps along the temporal dimension further. We feed the computed differences into a stack of 3×3 residual blocks, followed by five 3×3 convolutional layers with dilation rates $d \in \{3, 6, 12, 18, 24\}$. Each convolutional layer predicts a set of five offsets $o^{(d)}(p_n)$ for each pixel location p_n, which are used to rewarp pose heatmap B. The five rewarped heatmaps are then summed, and the resulting tensor is used to predict the target heatmap.

➢ 2. Heatmap Fusion

a) Heatmaps

$$\mathbf{H} = \{\{\mathbf{H}^1_{V_1}(x), \mathbf{H}^1_{V_1}(x), \ldots, \mathbf{H}^1_{V_1}(x)\}, \{\mathbf{H}^2_{V_2}(x), \mathbf{H}^2_{V_2}(x), \ldots, \mathbf{H}^2_{V_2}(x)\}, \ldots, \{\mathbf{H}^M_{V_M}(x), \mathbf{H}^M_{V_M}(x), \ldots, \mathbf{H}^M_{V_M}(x)\}\}$$

b) Heatmap Process for Spatial Fusion

$$\hat{\mathbf{H}}^n_v(x) = \lambda \mathbf{H}^n_v(x) + \frac{(1-\lambda)}{M} \sum_{u=1}^{M} \max_{x' \in \mathbf{p}^u(x)} \mathbf{H}^n_u(x'),$$

c) Temporal Fusion

$$\Phi^n_{V_j}(x) = \hat{\mathbf{H}}^n_{V_j}(x) - \mathbf{H}^{M \cdot i + j}_{V_j}.$$

$$\tilde{\mathbf{H}}^n_{V_j} = \sum_{d=1}^{5} \mathbf{Warper}(\Phi^n_{V_j}, o^{(d)}_{V_j}(x)),$$

➢ 1. Dataset and Evaluation Metric

- **Human3.6M.** Human3.6M is a large-scale benchmark dataset widely used for 3D human pose estimation in controlled indoor environments. It consists of 3.6 million frames recorded from four synchronized high-resolution cameras capturing 11 professional actors (6 male, 5 female) performing 15 distinct activities, including walking, sitting, and object interactions
- **MPI-INF-3DHP.** The MPI-INF-3DHP dataset is a comprehensive resource for multi-view 3D human pose estimation, featuring annotated frames from indoor and everyday settings. It includes 8 actors (4 male, 4 female), each performing 8 activity sets, such as walking, sitting, complex exercises, and dynamic actions. With diverse scenarios and multi-view recordings, the dataset enables robust environment.

The **MPJPE** measures 3D pose accuracy via mean Euclidean distance between predicted ($\hat{P} = \{\hat{p}_1, \ldots, \hat{p}_J\}$) and ground truth ($P = \{p_1, \ldots, p_J\}$) joints:

$$\text{MPJPE} = \frac{1}{J} \sum_{i=1}^{J} \|\hat{p}_i - p_i\|_2 \tag{11}$$

where $\| \cdot \|_2$ is Euclidean norm.

➢ 2. Main Results in Human3.6M

METHOD	INPUT	ACTIONS															AVG
		Dir.	Disc.	Eat.	Greet.	Phone.	Photo.	Pose.	Pur.	Sit.	SitD.	Smoke.	Wait.	Walk.	WalkD.	WalkT.	
2D—Ground Truth (GT)																	
GLA-GCN (T=243) (Yu et al., 2023b)	Single	26.6	27.2	29.2	25.4	28.2	31.7	29.5	27.0	37.8	40.0	29.9	27.0	20.5	27.3	20.8	28.5
KTP-Former (T=243) (Peng et al., 2024)	Single	**22.7**	23.4	21.8	22.5	24.2	29.9	25.7	22.9	30.3	36.9	24.4	23.3	**17.3**	24.3	**18.2**	24.5
Adafuse (Zhang et al., 2021)	Full	26.3	25.4	22.4	23.9	22.9	22.6	24.1	24.4	23.7	21.6	24.0	23.9	23.1	23.9	22.8	23.7
Adafuse + MCC (Su et al., 2021)	Interp	26.0	25.6	22.2	23.6	22.3	23.9	23.8	24.3	24.4	23.3	24.4	23.9	22.3	24.9	22.4	23.8
Adafuse + SLERP (Chen et al., 2022)	Interp	26.1	25.2	22.3	23.6	22.7	22.4	23.8	24.3	23.6	21.4	23.8	23.8	22.9	23.7	22.5	23.5
Adafuse	Sparse	27.3	27.4	23.6	26.3	24.3	24.1	25.0	27.3	24.4	22.7	25.0	25.3	26.6	26.4	26.0	25.4
PPT (Ma et al., 2022)	Full	23.2	26.3	22.0	22.9	25.2	23.1	23.8	28.5	31.2	25.2	27.4	23.6	26.5	23.4	25.0	25.2
PPT + MCC (Ma et al., 2022; Su et al., 2021)	Interp	23.5	27.5	21.5	21.9	24.7	29.0	23.0	23.7	30.6	34.1	26.6	22.2	22.4	27.6	24.0	25.5
PPT + SLERP (Ma et al., 2022; Chen et al., 2022)	Interp	23.0	26.0	21.6	**21.6**	24.9	27.1	22.8	23.3	28.2	32.4	24.9	22.1	23.1	26.2	24.7	24.8
PPT	Sparse	24.4	27.1	22.8	24.6	25.8	24.2	25.8	28.2	31.1	25.8	28.2	25.2	28.4	27.4	28.2	26.4
Ours	Sparse	23.2	**22.5**	**21.0**	21.9	**20.5**	**21.2**	**20.5**	**22.0**	**21.2**	**19.7**	**21.4**	**20.5**	22.1	**21.8**	20.7	**21.3**
2D—CPN																	
GLA-GCN (T=243)	Single	41.4	44.4	40.8	41.8	46.0	54.1	42.1	41.5	57.9	62.9	45.1	42.8	29.3	45.9	29.9	44.4
KTP-Former (T=243)	Single	37.7	39.7	35.9	37.7	42.1	48.0	38.7	39.2	52.5	56.2	41.3	40.0	26.8	39.6	27.6	40.2
FinePose (T=243) (Xu et al., 2024a)	Single	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.2
Adafuse	Full	35.0	37.1	32.2	34.9	35.2	36.6	33.0	34.6	40.5	41.3	37.2	35.3	33.8	37.4	33.1	35.8
Adafuse + MCC	Interp	**31.9**	36.9	30.1	**32.8**	33.4	**32.0**	32.0	32.4	**37.4**	48.8	**33.9**	33.7	32.4	36.5	31.6	34.4
Adafuse + SLERP	Interp	34.4	36.8	31.9	34.0	34.6	35.7	32.3	34.1	40.2	41.0	36.7	34.8	33.4	36.8	32.4	35.3
Adafuse	Sparse	35.9	37.2	33.3	36.2	36.7	37.2	33.1	36.9	41.8	41.0	37.6	35.8	37.0	38.2	35.1	36.9
Sgraformer (Zhang et al., 2024)	Full	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	35.4
Ours	Sparse	32.0	**35.2**	**30.0**	32.4	**33.0**	34.4	**30.2**	**32.3**	38.9	**40.1**	35.5	**32.9**	**31.4**	**36.0**	**30.4**	**33.6**
2D—SimpleBaseline																	
GLA-GCN (T=243)	Single	41.1	42.9	40.5	39.3	44.2	52.8	42.5	40.9	54.1	60.6	44.5	40.4	32.2	44.8	35.2	43.7
KTP-Former (T=243)	Single	35.4	36.9	34.3	34.7	39.6	43.4	36.3	35.0	47.5	57.4	39.4	35.4	27.5	38.7	29.4	38.1
FinePose (T=243)	Single	31.7	32.3	28.7	29.7	33.6	34.9	29.1	28.7	40.9	40.9	32.9	31.4	23.1	31.6	22.4	31.4
Adafuse	Full	28.3	29.9	25.3	29.5	26.9	26.4	27.0	28.1	28.7	32.1	27.8	28.8	26.7	29.6	25.5	28.1
Adafuse + MCC	Interp	27.6	30.0	25.1	29.4	26.5	26.9	26.4	27.9	29.2	32.3	27.9	28.5	26.4	29.9	25.4	28.0
Adafuse + SLERP	Interp	28.3	30.0	25.3	30.5	26.9	26.4	26.9	28.1	28.7	32.1	27.7	28.8	26.8	29.6	25.5	28.1
Adafuse	Sparse	29.9	30.6	26.1	31.2	27.9	27.4	27.7	30.2	29.7	32.3	28.5	29.8	30.8	31.0	29.0	29.5
Algebraic (Iskakov et al., 2019a)	Full	19.8	23.0	20.3	49.9	21.9	21.7	**18.3**	20.5	23.4	58.6	22.1	48.4	23.0	22.5	24.1	27.5
Volumetric (Iskakov et al., 2019a)	Full	**18.8**	**21.7**	**19.6**	50.1	21.2	**21.0**	18.4	**20.2**	**21.8**	57.1	**21.5**	48.4	22.5	**21.7**	22.5	26.7
Sgraformer	Full	24.3	25.1	21.1	24.7	24.5	24.9	21.6	22.1	26.5	32.1	25.1	24.0	**21.3**	25.4	21.6	24.3
Ours	Sparse	21.2	24.7	19.7	**23.0**	**19.8**	21.6	19.0	21.6	22.9	**31.2**	21.6	**23.2**	21.7	23.4	**19.8**	**22.3**

MPJPE Comparison with state-of-art pose estimation methods on Human3.6M (mm) using ground-truth and detected 2Dposes. Best in bold

➢ 2. Main Results in Human3.6M

METHOD	INPUT	ACTIONS															AVG
		Dir.	Disc.	Eat.	Greet.	Phone.	Photo.	Pose.	Pur.	Sit.	SitD.	Smoke.	Wait.	Walk.	WalkD.	WalkT.	
2D—SimpleBaseline-P-MPJPE																	
GLA-GCN (T=243) (Yu et al., 2023b)	Single	32.1	35.1	33.2	32.0	35.4	40.9	33.1	33.4	43.5	50.0	36.5	32.5	25.5	37.1	27.1	35.2
KTP-Former (T=243) (Peng et al., 2024)	Single	28.6	31.1	28.3	28.9	32.7	34.9	29.0	29.2	39.9	47.3	33.5	29.0	22.4	32.5	23.9	31.4
FinePose (T=243) (Xu et al., 2024a)	Single	24.8	26.3	24.7	24.0	27.0	28.1	22.4	23.8	33.7	33.3	27.4	24.6	19.2	25.8	18.5	25.6
Adafuse (Zhang et al., 2021)	Full	21.0	22.4	19.1	20.9	20.8	20.2	19.4	20.0	22.2	23.1	21.3	20.1	19.4	22.1	18.1	20.7
Adafuse + MCC (Zhang et al., 2021; Su et al., 2021)	Interp	20.5	22.6	18.8	21.0	20.5	21.0	19.1	19.4	22.6	23.3	21.6	20.1	19.4	22.5	18.4	20.7
Adafuse + SLERP (Zhang et al., 2021; Chen et al., 2022)	Interp	20.9	22.4	19.0	22.2	20.8	20.2	19.3	19.9	22.2	**23.0**	21.2	20.1	19.3	22.2	18.1	20.7
Adafuse (Zhang et al., 2021)	Sparse	22.5	22.6	19.7	22.5	21.5	20.9	20.0	20.5	22.5	23.4	21.7	20.6	23.3	23.4	20.9	21.7
Sgraformer (Zhang et al., 2024)	Full	**19.9**	**20.1**	18.1	**18.2**	20.8	20.3	17.1	17.7	23.0	26.2	21.9	18.6	**17.7**	20.7	17.9	19.9
Ours	Sparse	20.3	22.9	**17.8**	**18.2**	**17.9**	**19.0**	**15.6**	**17.4**	**21.3**	27.3	**18.9**	**18.2**	19.0	**20.5**	**16.6**	**19.4**

Note: Complete version with all baseline comparisons. Gray rows highlight our method. Action abbreviations: Directions (Dir), Discussion (Disc), Sitting Down (SitD), Walking Dog (WalkD), Walking Together (WalkT). Time frames (T=243) are shown where applicable. For the 2D pose estimation, we utilize ground truth, CPN (Cascaded Pyramid Network), and SimpleBaseline to obtain the corresponding 2D pose sequences. T represents the number of input time frames. MCC (Motion Consistency and Continuity) and SLERP (Spherical Linear Interpolation) are keypoint interpolation methods. MCC is a neural network-based interpolation method while SLERP is a traditional interpolation technique.

P-MPJPE Comparison on with state-of-art pose estimation methods on Human3.6M (mm) using the ground-truth and detected 2D poses. Input types: Single-view (Single), Multi-view full-frame (Full), Multi-view interpolated (Interp). Best in bold.

➢ 3. Main Results in MPI-3D-INF

MPJPE (SimpleBaseline(2D))	Input Method	MPJPE ↓
GLA-GCN (T=243)	Single	75.00
KTP-Former (T=243)	Single	67.59
Adafuse	Full	78.57
Adafuse + MCC	Interpolation	-
Adafuse + SLERP	Interpolation	83.37
PPT	Full	106.30
PPT + MCC	Interpolation	-
PPT + SLERP	Interpolation	110.34
Ours	sparse Interleaved	**65.89**

Reconstruction Error (MPJPE in mm) on the MPI-INF-3DHP Dataset. Input 2D pose sequences are obtained using a SimpleBaseline detector. T denotes the number of input frames. Best results are high lighted in bold.

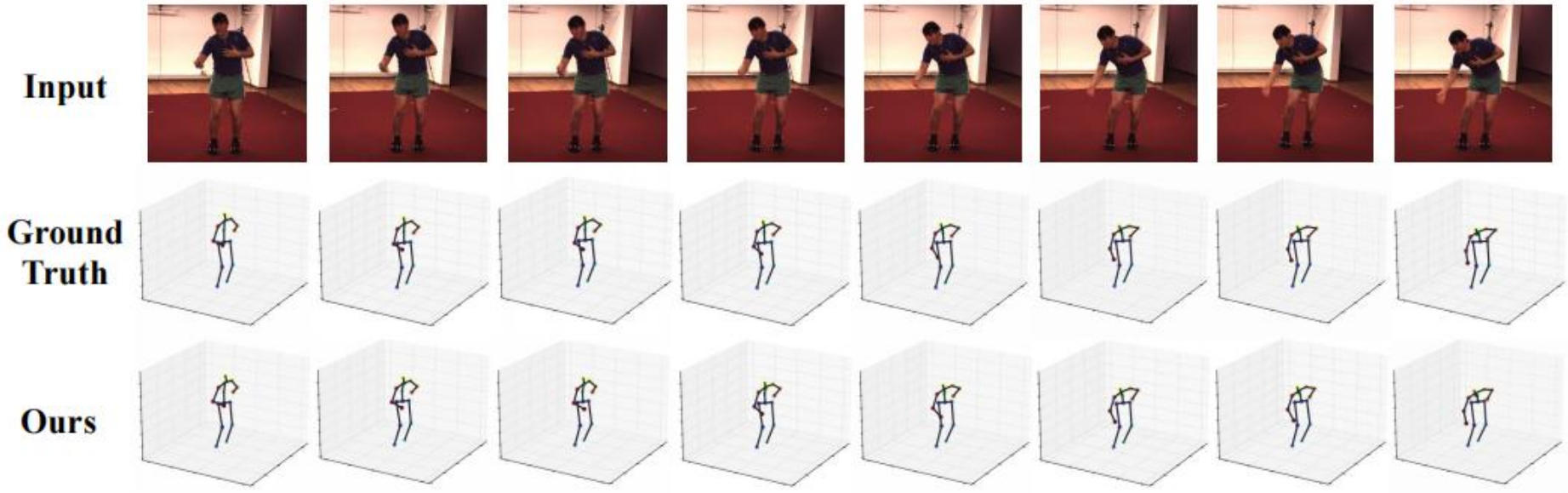
Table 4: Model Parameter Count and Performance Efficiency. Performance Efficiency (MPJPE/mm per MB) is calculated as the ratio of MPJPE (in mm) to model size (in MB). Smaller values of this metric indicate better trade-offs between performance (MPJPE) and model size (MB), with more efficient models achieving lower MPJPE while maintaining smaller parameter sizes. The average latency specifically refers to the computational time of a single model inference (in milliseconds).

Method	Para.(M) ↓	Flops.(GFLOPs)↓	Average Latency. (ms)↓	Performance per MB (MPJPE/mm per MB) ↓
GLA-GCN (T=243)	69.99	**51.13**	**24.10**	0.624
KTP-Former (T=243)	103.85	51.64	24.11	0.367
FinePose (T=243)	269.23	287.32	82.24	**0.117**
Adafuse (T=1)	**69.66**	204.26	96.028	0.403
Adafuse + SLERP	**69.66**	204.26	96.03	0.403
Adafuse + MCC	72.25	204.26	96.028	0.388
Sgraformer + Full	81.23	204.28	99.19	0.299
Ours	76.51	**111 .36**	**44.51**	**0.291**

➢ 4. Ablation Study

Method	Spatial Heatmap Fusion	Warper	Avg. ↓
Ours (Human3.6M)	✗	✗	36.06
	✓	✗	31.54
	✓	✓	**22.28**
Ours (MPI-INF-3DHP)	✗	✗	94.46
	✓	✗	88.63
	✓	✓	**65.89**

Ablation study results. We conducted ablation studies on the Human3.6M and MPI-INF-3DHP datasets to validate the effectiveness of the proposed space fusion module based on epipolar geometry and the temporal fusion module Warper. We use SimpleBaseline as 2D baseline model. We have bolded the best results



4. Conclusion

➢ 1. Conclusion

a) Paradigm Innovation: Pioneered a sparse interleaved input method that breaks the traditional reliance on dense, synchronized multi-view data.

b) High-Frequency Output: Demonstrated that high-resolution 3D pose signals can be accurately reconstructed from temporally staggered, low-frame-rate inputs.

c) DenseWarper Framework: Validated an efficient end-to-end module that achieves state-of-the-art performance on Human3.6M and MPI-INF-3DHP benchmarks.

d) Impact: Provided a compelling proof-of-concept for resource-efficient, real-time 3D perception in robotics and VR/AR.

➢ 2. Limation

a) Non-Uniform Intervals: The current model has not been fully explored under irregular or extremely sparse camera sampling.

b) Temporal Density Dependency: Effectiveness may decrease when inter-camera intervals are excessively large, making it harder to recover fine-grained spatio-temporal information.

➢ 3. Future Work

a) Generalizability: Extending the interleaved paradigm to other 3D tasks, such as multi-view object detection.

b) Theoretical Depth: Investigating the underlying interpretability and mathematical foundations of the novel input method.

THANK YOU

FROM SPARSE TO DENSE: SPATIO-TEMPORAL FUSION FOR MULTI-VIEW 3D HUMAN POSE ESTIMATION WITH DENSEWARPER

Ling Li*, Changjie Chen†, Yuyan Wang^, Jiaqing Lyu*, Kenglun Chang**, Yiyun Chen††, and Zhidong Deng*

Repoter: Ling Li