HandDAGT : A Denoising Adaptive Graph Transformer for 3D Hand Pose Estimation

Wencan Cheng, **Eunji Kim**, Jong Hwan Ko

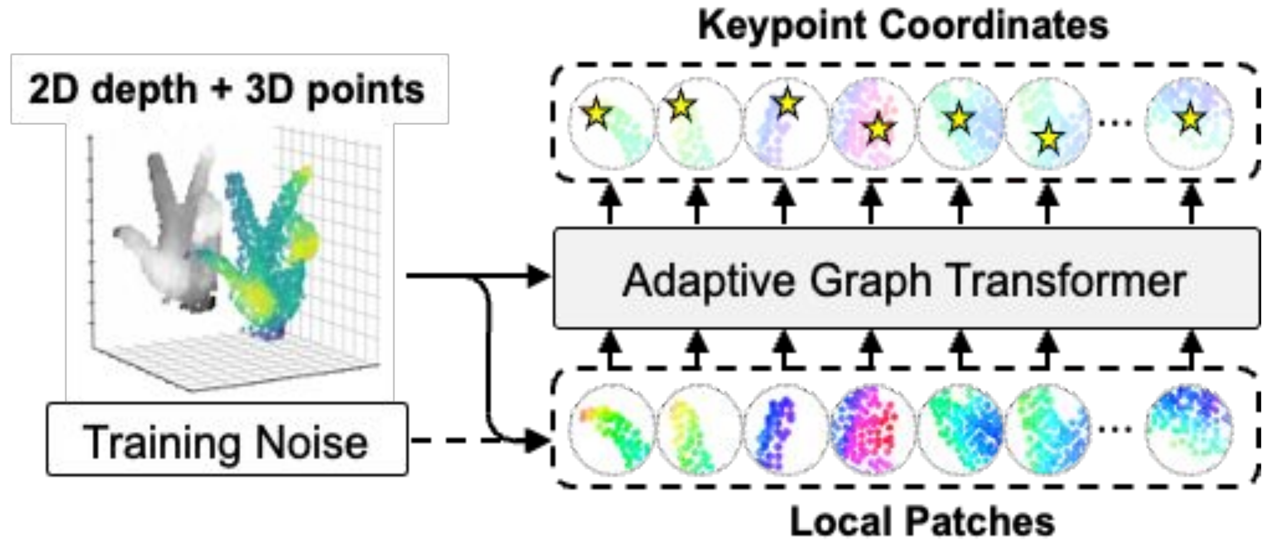
Sungkyunkwan University

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Motivation and Overview

• Overview



• Motivation

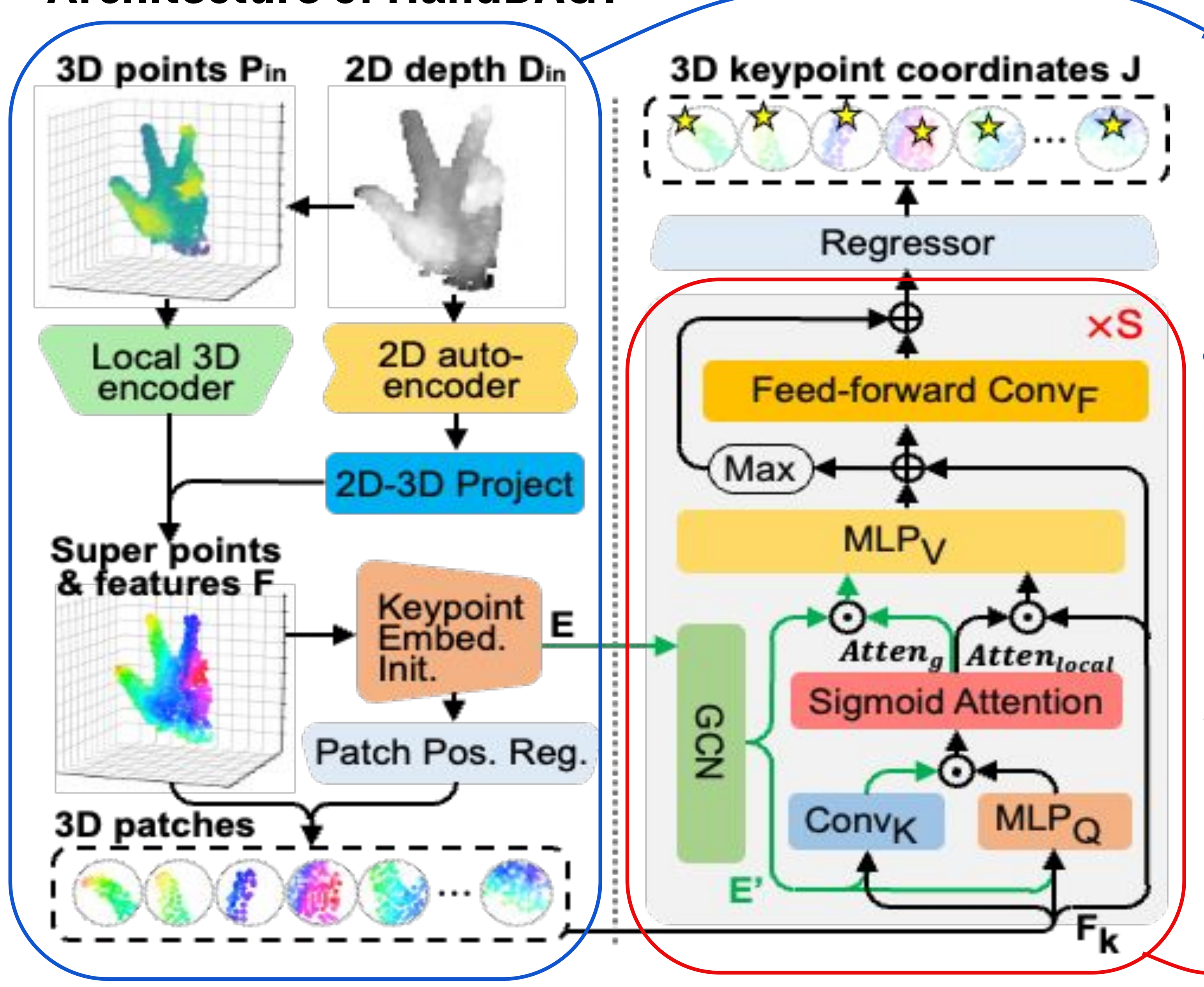
- This task proposes a denoising adaptive graph transformer (HandDAGT) to accurately estimate 3D hand poses by adapting to various occlusion scenarios.

• Contribution

- **Novel Transformer Architecture:** Integrates both 2D depth images and 3D point clouds as multi-modal inputs.
- **Adaptive Attention Mechanism:** Dynamically adjusts focus between local geometric details and kinematic correspondences.
- **Denoising Training Strategy:** Enhances robustness and accuracy by training the model to correct noisy input estimations.

Method and Approach

• Architecture of HandDAGT



- **Embedding Initialization**
  Captures the kinematic topology between keypoints
- **Patch position Initialization**
  Projects embeddings into 3D space and gathers 3D patches using K-nearest super points
- **Adaptive Graph Transformer**
  Augments keypoint embeddings with GCN, applies novel attention, and aggregates embeddings for keypoint regression

• Training Loss → Introduces noise only during the training phase

$$\mathcal{L} = \sum \mathrm{L1}_{\mathrm{smooth}}(\mathbf{D}_{T_1}(\mathbf{J}_0 + \mathcal{N}) - \mathbf{J}^*) + \sum_{2}^{S} \sum \mathrm{L1}_{\mathrm{smooth}}(\mathbf{D}_{T_s}(\mathbf{J}_{s-1}) - \mathbf{J}^*)$$

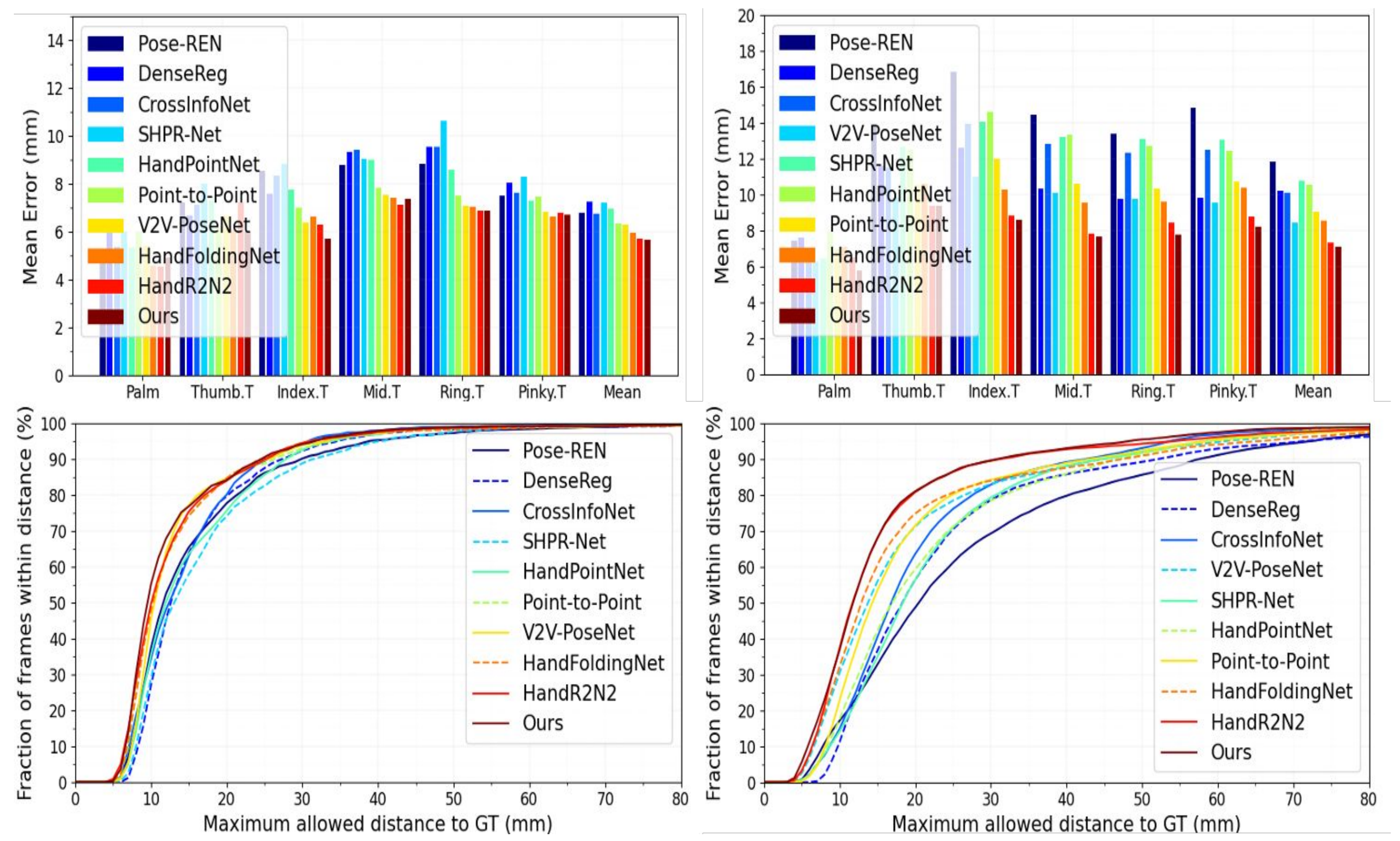
$\mathbf{D}_{T_s}$: stacked denosing transformer $\mathcal{N}$: injected noise $\mathbf{J}^*$: ground truth $\mathbf{J}_s$: s-th denoising transformer output

Experimental Results

ICVL and NYU datasets (single-hand)

Method	Mean keypoint error (mm)		Input
	ICVL	NYU	
Ren-9x6x6 [17]	7.31	12.69	D
DeepPrior++ [32]	8.1	12.24	D
Pose-Ren [4]	6.79	11.81	D
DenseReg [44]	7.3	10.2	D
CrossInfoNet [10]	6.73	10.08	D
JGR-P2O [11]	6.02	8.29	D
SSRN [37]	6.01	7.37	D
PHG [36]	5.97	7.39	D
HandPointNet [13]	6.94	10.54	P
Hand-Transformer [21]	6.47	9.80	P
Point-to-Point [16]	6.3	9.10	P
V2V [31]	6.28	8.42	V
HandFolding [8]	5.95	8.58	P
HandR2N2 [6]	5.70	7.27	P
IPNet [35]	5.76	7.17	D+P
HandDAGT (Ours)	**5.66**	**7.12**	D+P

→ HandDAGT outperforms ICVL and NYU SOTA



DexYCB dataset

→ Outperforms DexYCB SOTA

Method	Mean keypoint error (mm)					Input
	S0	S1	S2	S3	AVG	
A2J [47]	23.93	25.57	27.65	24.92	25.52	D
Spurr et al. [39]	17.34	22.26	25.49	18.44	18.44	RGB
METRO [25]	15.24	-	-	-	-	RGB
Tse et al. [42]	16.05	21.22	27.01	17.93	20.55	RGB
HandOcc [33]	14.04	-	-	-	-	RGB
IPNet [35]	8.03	9.01	8.60	7.80	8.36	D+P
HandDAGT (Ours)	**7.72**	**8.68**	**8.22**	**7.52**	**8.03**	D+P

HO3D dataset

→ Comparable performance with SOTA method

Method	Mean keypoint error (mm)	Input
Hybrik [22]	2.89	RGB
ArtiBoost [48]	2.53	RGB
HandOccNet [33]	2.49	RGB
HandVoxNet++ [30]	2.46	V
IPNet [35]	1.81	D+P
HandDAGT (Ours)	**1.81**	D+P

Qualitative results

ICVL **NYU** **DexYCB (hand - object interaction)**