MM-Hand: 3D-Aware Multi-Modal Guided Hand Generative Network for 3D Hand Pose Synthesis

Zhenyu Wu[1*], Duc Hoang[1*], Shih-Yao Lin[2], Yusheng Xie[3], Liangjian Chen[4] , Yen-Yu Lin[5], Zhangyang Wang[6], and Wei Fan[2]

[1]Texas A&M University, [2]Tencent America, [3]Amazon Web Service, [4]University of California at Irvine,

[5]National Chiao Tung University, [6]University of Texas at Austin

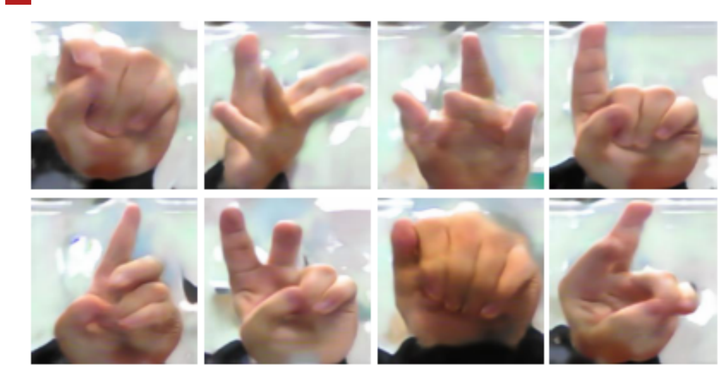
Motivation

- 3D hand pose estimation is widely used in sign language recognition, HCI, healthcare and entertainment
- Annotation of 3D hand pose on RGB hand images are difficult and prone to errors by humans
- Limitation of using synthetic hand images to train 3D hand pose estimators
  - Estimators trained on the synthetic data often fail to generalize due to the visual domain gap between non-photo-realistic synthetic data and real images
  - Building hand models with various textures or shapes requires laborious 3D geometric modeling and rendering.
- Two challenges specific to the hand domain
  - occlusion: various 3D hand movements will always make some finger parts invisible from 2D images.
  - self-similarity: the five fingers of the same hand share similar appearance and structure, making them indistinguishable.

Introduction

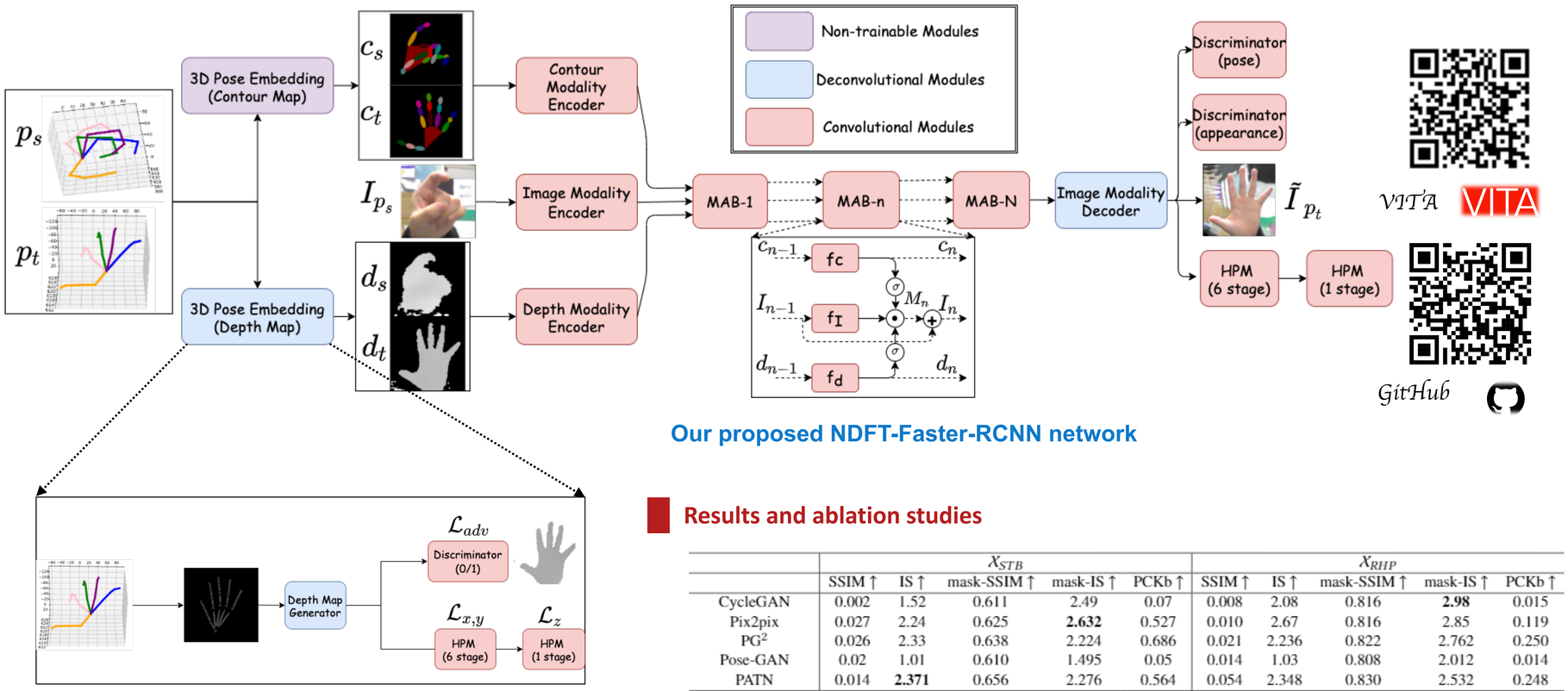
- Our proposed framework, 3D-Aware Multi-modal Guided Hand Generative Network (MM-Hand), carries out the first attempts to generate hand images under the guidance of 3D poses. The proposed MM-Hand is able to improve the realism, increase the diversity, and preserve the 3D pose of the generated images simultaneously.
- MM-Hand is trained with a novel geometry-based curriculum learning strategy. Starting with easy pose-images pairs, we gradually increase the training task difficulty.
- Extensive experiments demonstrate that our generated hand images can consistently improve 3D hand pose estimation, across two strong pose estimators and two hand pose datasets.

Visualization



We randomly pick hand images generated by MM-hand trained on STB

Methodology



Our proposed NDFT-Faster-RCNN network

A detailed look into the depth map generator

Training and inference details

- We hypothesize that the level of difficulty to generate target hand image $\widetilde{I_{p_t}}$ from source hand image I_{p_s} is positively correlated with the 3D pose distance between p_s and p_t.
- **GCT: Geometry-based Curriculum**
  - In the training stage, MM-Hand is fed by the data loader with hand pairs progressively from the easiest (smallest pose distance) pair to the hardest (largest pose distance) pair.
- **INNM: Inference with Nearest Neighbor Match**
  - In the inference stage, given a target 3D hand pose p_t, we find the best matched source hand image I_{p_s} in the training hand images whose pose p_s is closest to p_t in pose distance.

Results and ablation studies

	X_{STB}					X_{RHP}				
	SSIM ↑	IS ↑	mask-SSIM ↑	mask-IS ↑	PCKb ↑	SSIM ↑	IS ↑	mask-SSIM ↑	mask-IS ↑	PCKb ↑
CycleGAN	0.002	1.52	0.611	2.49	0.07	0.008	2.08	0.816	**2.98**	0.015
Pix2pix	0.027	2.24	0.625	**2.632**	0.527	0.010	2.67	0.816	2.85	0.119
PG^2	0.026	2.33	0.638	2.224	0.686	0.021	2.236	0.822	2.762	0.250
Pose-GAN	0.02	1.01	0.610	1.495	0.05	0.014	1.03	0.808	2.012	0.014
PATN	0.014	**2.371**	0.656	2.276	0.564	0.054	2.348	0.830	2.532	0.248
Ours (MM-Hand)	**0.115**	2.187	**0.677**	2.53	**0.688**	**0.078**	**2.376**	**0.844**	2.747	**0.619**

Quantitative comparison of the generated hand images using CycleGAN, Pix2pix, PG^2, Pose-GAN, PATN, and our proposed MM-Hand on the two benchmark datasets STB and RHP.

		X_{STB}					X_{RHP}				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
$\mathcal{M}_{Hand3D}$	None	60.28	48.84	30.23	18.74	9.81	90.12	60.27	36.58	25.29	20.52
	CycleGAN	80.27	82.57	75.39	72.56	9.81	90.29	78.29	60.29	40.29	20.52
	Pix2pix	72.57	71.27	54.13	50.25	9.81	82.39	76.48	62.16	80.29	20.52
	PG^2	74.27	68.22	58.23	52.28	9.81	85.29	72.83	64.49	40.25	20.52
	PATN	70.28	68.23	50.37	40.57	9.81	84.25	74.49	84.35	60.25	20.52
	Pose-GAN	72.58	69.27	52.85	39.56	9.81	94.59	84.38	67.83	45.59	20.52
	Ours (MM-Hand)	**52.39**	**32.37**	**27.49**	**16.48**	9.81	**80.29**	**54.38**	**28.49**	**24.38**	20.52
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
$\mathcal{M}_{3D-HPM}$	None	64.16	48.91	33.00	35.51	15.71	**52.21**	50.38	47.36	45.43	35.86
	CycleGAN	111.75	54.55	51.71	47.02	15.71	66.63	59.63	57.59	61.67	35.86
	Pix2pix	99.59	46.71	47.83	46.91	15.71	65.73	64.56	62.31	55.07	35.86
	PG^2	91.03	47.00	47.20	46.78	15.71	61.05	58.95	57.59	56.72	35.86
	PATN	99.66	46.95	47.84	40.18	15.71	56.11	50.26	50.64	51.92	35.86
	Pose-GAN	102.70	46.65	48.03	47.02	15.71	60.54	57.44	53.40	52.83	35.86
	Ours (MM-Hand)	**41.79**	**20.24**	**16.79**	**16.15**	15.71	52.47	**42.22**	**41.63**	**40.49**	35.86

The 3D hand pose estimation performance using M_{Hand3D} and M_{3D-HPM} on X_{STB} and X_{RHP}, augmented by images generated by different methods under different portion α of X as the reduced training set.

Contact: {wuzhenyu_sjtu, hoangd}@tamu.edu, {shihyaolin, davidwfan}@tencent.com, yushx@amazon.com, atlaswang@utexas.edu, liangjc2@ics.uci.edu, lin@cs.nctu.edu.tw