Deep Learning Techniques for Digital Content Generation

Overview

- ***Deep Learning Techniques for Digital Content Generation***
- Introduction
  - Motivation
  - Methodology
- City Modeling
  - Houses
  - Road layouts
- Face Avatar
  - Animated conversations
  - Telepresence

Fig1.3. Examples of generated digital content.

Introduction: Motivation

- Creating contents manually
  - High demand
  - Tedious
  - Expensive
- Automation
  - Assist artists
  - Improve quality
  - Enable new media

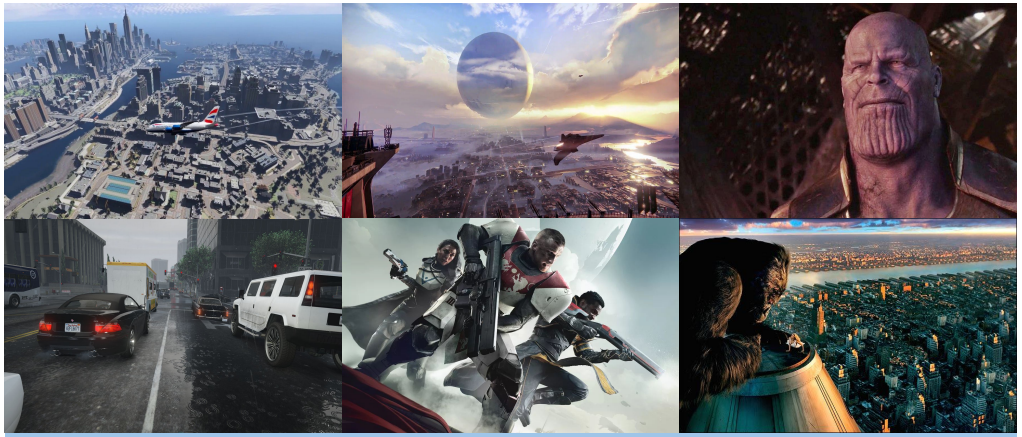


Fig1.1. GTA5 ($200mil), Destiny ($500mil), Avengers4 ($300mil), KingKong ($200mil).

Fig1.2. Existing content creation tools: SketchUp, street procedural rules, Blender rigging.

Introduction: Methodology

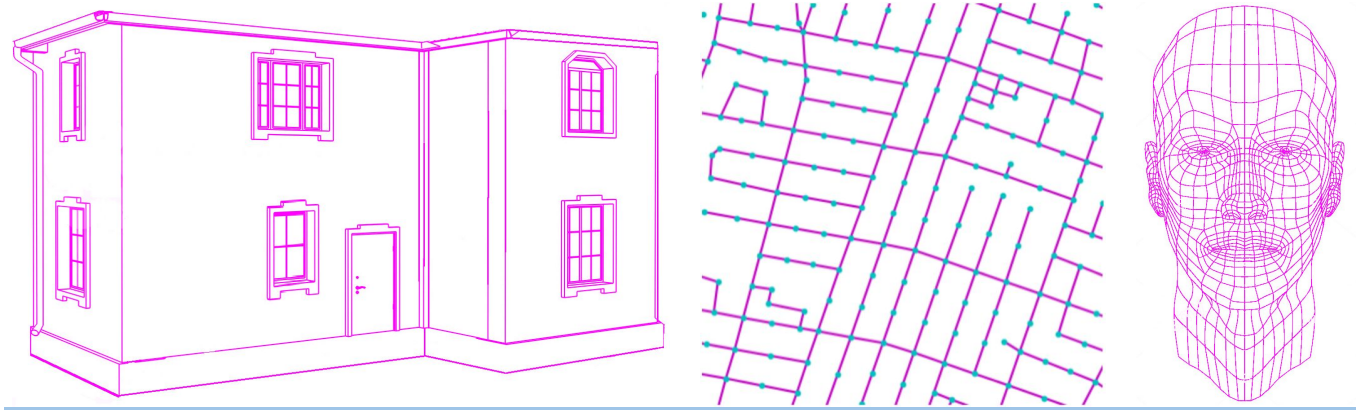


Fig1.4. The outputs of digital content generation are strongly structured, e.g. house, street, and face.

- Structured learning
  - Output is structured object $f : \mathbf{X} \rightarrow \mathbf{Y}, \mathbf{y} \in \mathbf{Y}$
  - E.g. house configuration[c2], spatial graph[c3], gesture sequence[c4], avatar parameters[c5]
- Parameterizing features
  - Measurable evidence from input $\phi(\mathbf{y}; \mathbf{x})$
  - E.g. geometric[c2,c3], object[c2], topologic[c3], context[c3,c4], perceptual[c2,c3,c5], semantic[c2,c4]

Overview

- ***Deep Learning Techniques for Digital Content Generation***
- Introduction
  - Motivation
  - Methodology
- City Modeling
  - Houses
  - Road layouts
- Face Avatar
  - Animated conversations
  - Telepresence

Fig1.3. Examples of generated digital content.

Chapter 2: HouseCraft - Introduction

- Inputs/Outputs
  - In: Floorplan+Streetview
  - Out: Position+Dimensions
- Key idea
  - Parameterize house
  - Estimate according to image evidences
  - Dataset: SydneyHouse

Fig2.1. Floorplan, streetview images, and output house horizontal position and vertical dimensions.

Fig2.2. Imprecise address position, lacking vertical dimensions.

Fig2.3. SydneyHouse annotation example.

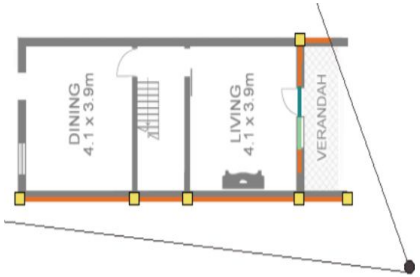
[1] HouseCraft: Building Houses from Rental Ads and Street Views H.Chu, S.Wang, R.Urtasun, S.Fidler, *ECCV* 2016

Chapter 2: HouseCraft - Method

- **Structured output**

  House configuration:

$$\mathbf{y} = \{\mathbf{y}_1, \dots \mathbf{y}_N\}, \mathbf{y}_n \in \{y_n^1, \dots y_n^{K_n}\}$$



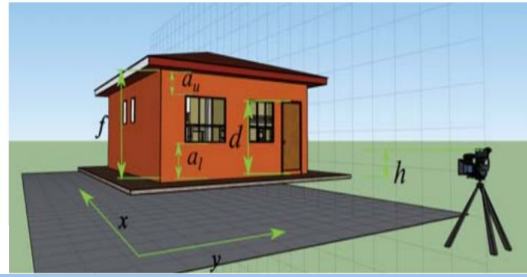


Fig2.3. Top-down view and 3D illustration of the structured outputs of the house.

- **Formulation**

$$E(\mathbf{y};\mathbf{x}) = \sum_{c=1}^{C} \mathbf{w}_c^T \phi_c(\mathbf{y};\mathbf{x})$$

  - Learning with Struct SVM
  - Brute-force inference (*O(1)* features)

- **Features**
  - Check multiple visual clues
  - Efficiency via integral geometry

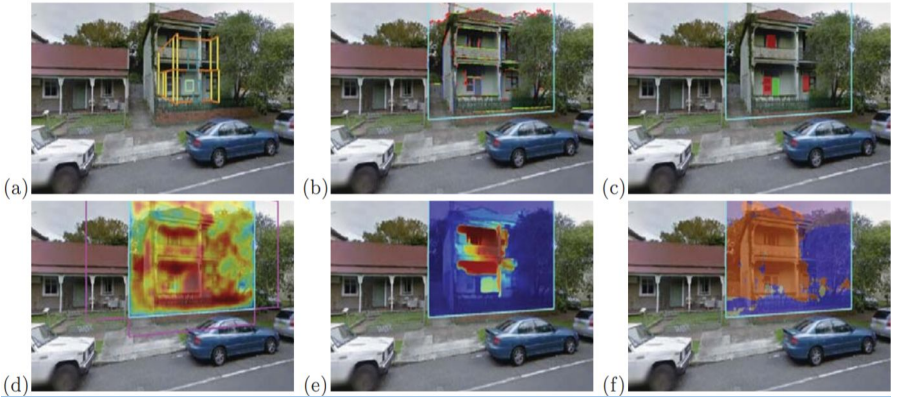


Fig2.4. House hypothesis; Edge, object, color, saliency, and semantic features.

[1] HouseCraft: Building Houses from Rental Ads and Street Views H.Chu, S.Wang, R.Urtasun, S.Fidler, *ECCV* 2016

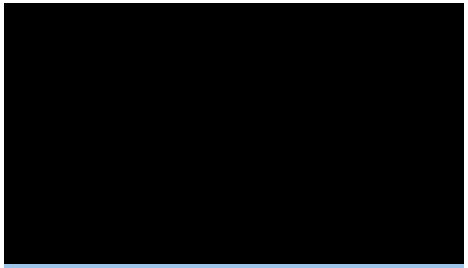
Chapter 2: HouseCraft - Results

- Results
  - Improved accuracy
  - Vertical dimensions
  - Image consistency
- Issues
  - Failure cases
  - Requires floorplan

Fig2.5 Examples of success and failure cases.

	xy/m	IOU	h/cm	f/cm	d/cm	a_l/cm	a_u/cm
random	9.07	21.04%	102.8	49.8	45.6	47.0	55.9
box-reg [40, 41]	6.68	33.31%					
google	5.01	43.46%					
ours	**2.62**	**68.29%**	**49.7**	**43.1**	**14.1**	**36.9**	**33.6**

Table2.1 Main quantitative results.



Vid2.1 Result demo.

[1] HouseCraft: Building Houses from Rental Ads and Street Views H.Chu, S.Wang, R.Urtasun, S.Fidler, *ECCV* 2016

Overview

- ***Deep Learning Techniques for Digital Content Generation***
- Introduction
  - Motivation
  - Methodology
- City Modeling
  - Houses
  - Road layouts
- Face Avatar
  - Animated conversations
  - Telepresence

Fig1.3. Examples of generated digital content.

Chapter 3: NeuralTurtleGraphics - Introduction

- Inputs/Outputs
  - In: Random seed/Aerial image
  - Out: Road layout
- Key idea
  - Road layout as spatial graph
  - Generative model for spatial graphs
  - Dataset: RoadNet+SpaceNet

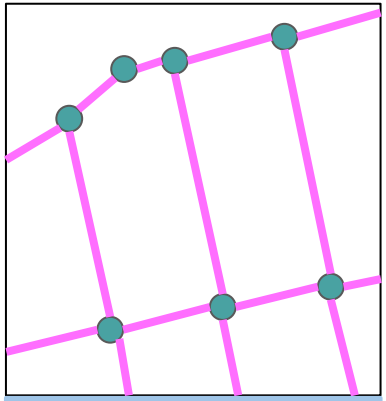
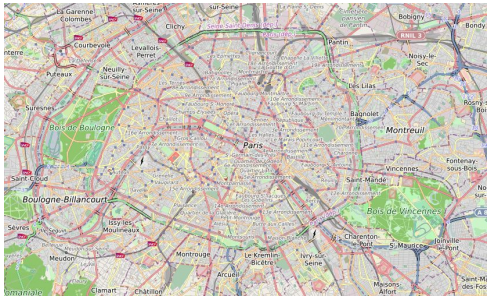


Fig3.1 An example road graph.



	Country	City	Node	Edge	Area	Length
RoadNet	13	17	233.6k	262.1k	170.0km^2	7410.7km
SpaceNet	4	4	115.8k	106.9k	122.3km^2	2058.4km

Fig3.2 Dataset statistics.

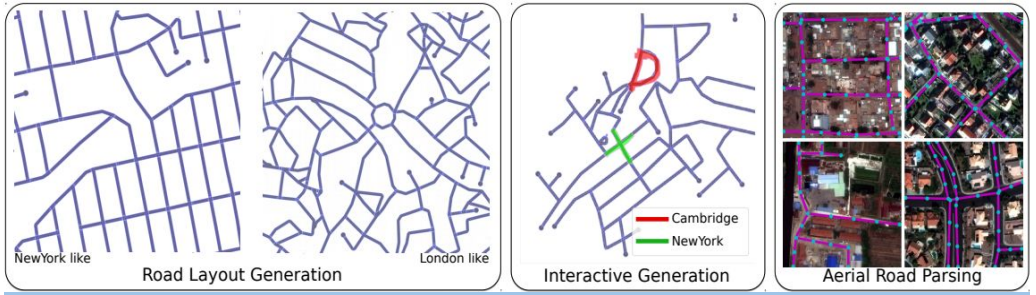


Fig3.3 Tasks and applications of generative road layout model.

[2] Neural Turtle Graphics for Modeling City Road Layouts, H. Chu, D. Li, D. Acuna, A. Kar, M. Shugrina, X. Wei, M.-Y. Liu, A. Torralba, and S. Fidler, *ICCV* 2019

Chapter 3: NeuralTurtleGraphics - Method

- **Structured output**

  Spatial graph:

  $$\mathbf{y} = \{\mathbf{V}, \mathbf{E}\}$$
  $$\mathbf{V}_i \in \{v_1, ...v_k\} \times \{v_1, ...v_k\}$$
  $$\mathbf{E}_{ij} = \mathbf{E}_{ji} \in \{0, 1\}$$

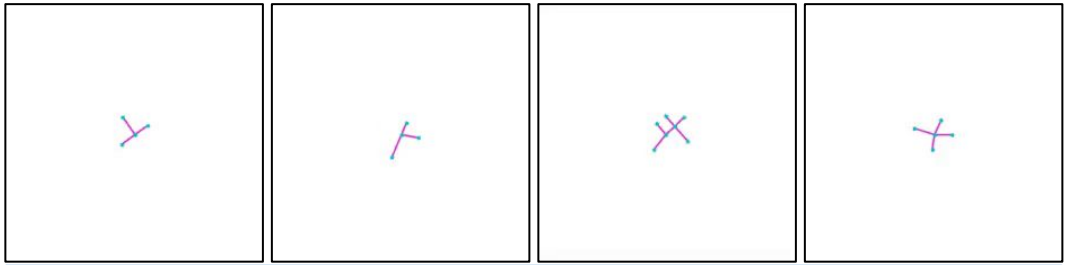


Fig3.4 Examples of iterative road layout graph generation. (Animated GIFs)

- **Formulation**

  $$P(\mathbf{y}_i | \{\mathbf{y}_{i'}, \forall i', O(i') < O(i)\})$$

  - BFS-order sequential generation
  - Parallel vertex-wise training

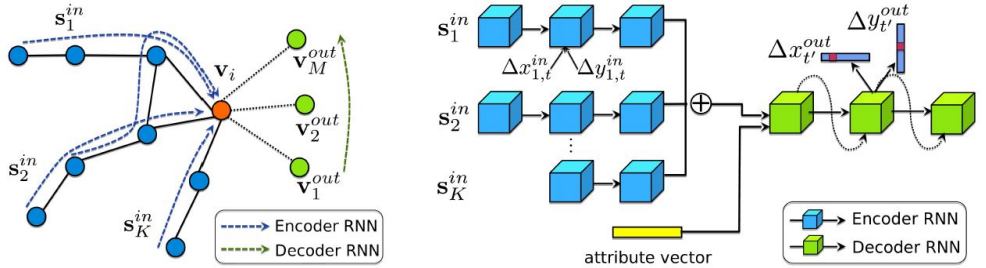


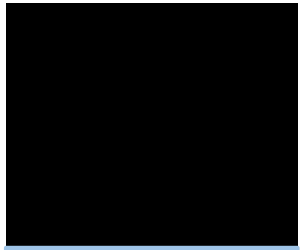
Fig3.5 Encoder-decoder neural network for a single local expansion step.

- **Features**
  - Shape and topology
  - Within local neighbourhood

[2] Neural Turtle Graphics for Modeling City Road Layouts, H. Chu, D. Li, D. Acuna, A. Kar, M. Shugrina, X. Wei, M.-Y. Liu, A. Torralba, and S. Fidler, *ICCV* 2019

Chapter 3: NeuralTurtleGraphics - Results

- Results
  - City generation
  - Interactive synthesis
  - Aerial road parsing
  - Create simulation
- Issues
  - Global layout
  - Iterative generation



Vid3.1 Result demo.

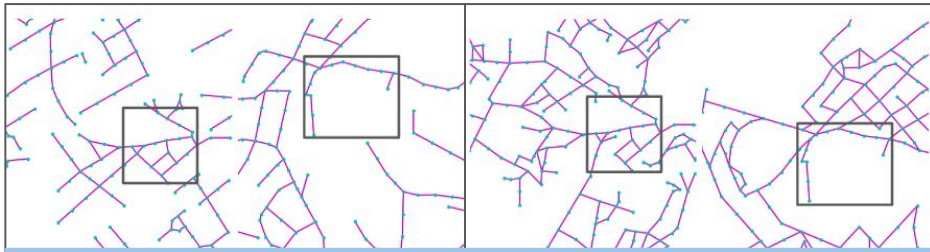


Fig3.6 Training (left) and cities generated (right) via piecewise memorization.

Method \ Metric	*Perceptual*					*Urban Planning*				
	mp1 10^{-1}	mp2 10^{0}	pa 10^{-1}	fc 10^{1}	rate	densi. 10^{1}	conne. 10^{-2}	reach 10^{5}	conve. 10^{-3}	rate
GraphRNN-2D [38, 6]	7.12	6.35	8.45	16.15	25.0	51.58	4.61	45.11	6.72	43.7
PGGAN [22]	1.98	2.15	5.34	10.51	63.2	45.77	19.48	4.33	2.94	58.9
CityEngine-5k [1]	2.74	2.71	8.34	14.78	47.1	13.59	21.66	7.61	16.66	51.7
CityEngine-10k [1]	2.55	2.56	8.23	14.17	48.9	12.43	21.79	7.05	16.82	52.1
NTG-vanilla	2.63	2.33	4.05	9.17	66.0	8.69	**1.87**	8.99	3.06	86.5
NTG-enhance	**1.52**	**1.34**	**2.83**	**6.76**	**77.3**	**3.76**	1.97	**4.13**	**1.86**	**92.4**

Table3.1 Quantitative results on perceptual and urban planning metrics.

[2] Neural Turtle Graphics for Modeling City Road Layouts, H. Chu, D. Li, D. Acuna, A. Kar, M. Shugrina, X. Wei, M.-Y. Liu, A. Torralba, and S. Fidler, *ICCV* 2019

Overview

- ***Deep Learning Techniques for Digital Content Generation***
- Introduction
  - Motivation
  - Methodology
- City Modeling
  - Houses
  - Road layouts
- Face Avatar
  - Animated conversations
  - Telepresence

Fig1.3. Examples of generated digital content.

Chapter 4: Face-to-FaceConvo - Introduction

- Inputs/Outputs
  - In: Text+Gesture+History
  - Out: Text+Gesture
- Key idea
  - Joint text-gesture model
  - Dataset: MovieChat

Fig4.1 Facial gesture conveys conversational information.

query	target	# of examples
text	text	40,200,261
text+face	text	48,475
text	text+face	48,475
text+face	text+face	24,727

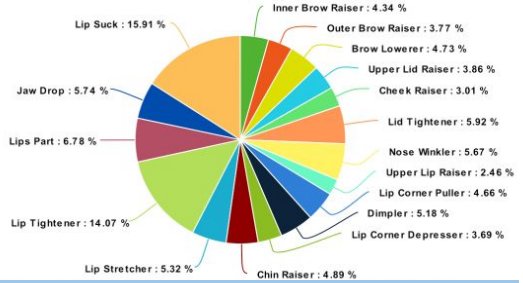


Fig4.2 Statistics of face-to-face conversations in the MovieChat dataset.

[3] A Face-to-Face Neural Conversation Model, H.Chu, D.Li, S.Fidler, *CVPR* 2018

Chapter 4: Face-to-FaceConvo - Method

- **Structured output**

  Words+Gestures:

$$\mathbf{y} = \{\mathbf{y}_1^w, \mathbf{y}_1^g, ..., \mathbf{y}_T^w, \mathbf{y}_T^g\}$$
$$\mathbf{y}_1^w \in \{1, ..., K^w\}, \mathbf{y}_1^g \in \{1, ..., K^g\}$$

- **Formulation**

$$P(\mathbf{y}_t^w, \mathbf{y}_t^g | \mathbf{y}_{<t}^w, \mathbf{y}_{<t}^g)$$

  - Sentence-level reward
  - Annealled RL training

- **Features**
  - Multi-stream
  - Hierachical sequences

Fig4.3 An example of Action Unit-based gesture representation.

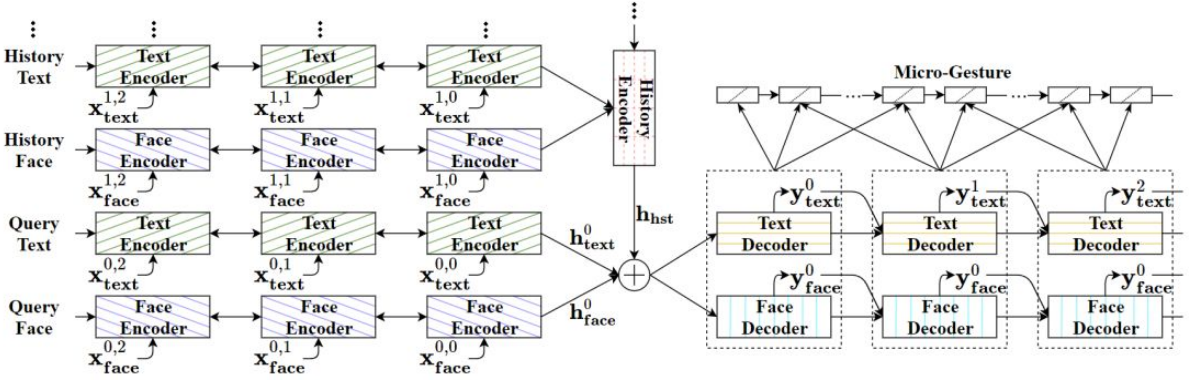


Fig4.4 The hierachical joint text-gesture model.

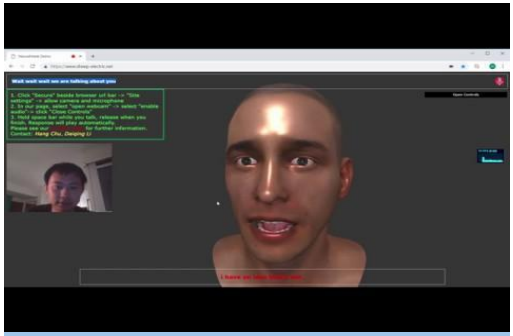
[3] A Face-to-Face Neural Conversation Model, H.Chu, D.Li, S.Fidler, *CVPR* 2018

Chapter 4: Face-to-FaceConvo - Results

- Results
  - Gesture can help text
  - Text can help gesture
  - Diverse conversation
- Issues
  - Uncanny appearance
  - Rare expressions

	text %	*face %*	*overall %*	
noMicro-noGAN	48.8	39.0	46.2	*pairwise*
Micro-noGAN	**51.2**	**61.0**	**53.8**	
noMicro-noGAN	44.8	35.3	42.4	
Ours	**55.2**	**64.7**	**57.6**	
Micro-noGAN	46.1	48.8	46.7	
Ours	**53.9**	**51.2**	**53.3**	
noMicro-noGAN	31.5	25.0	29.8	*accumu.*
Micro-noGAN	32.5	36.8	33.5	
Ours	**36.0**	**38.2**	**36.6**	

Table4.1 User study on realisticness and interestingness.



Vid4.1 Result demo.

	perp.	beam=1 *pre. %*	beam=1 *rec. %*	beam=1 *F1*	beam=3 *pre. %*	beam=3 *rec. %*	beam=3 *F1*	beam=5 *pre. %*	beam=5 *rec. %*	beam=5 *F1*
Text [12, 27]	32.53	23.18	15.58	17.12	**25.00**	17.13	18.62	**24.70**	16.91	18.34
Text+RandFace	32.65	22.92	15.99	17.27	24.74	17.32	18.57	**24.71**	17.82	18.84
Text+Face	**30.17**	24.25	17.52	18.69	**24.78**	18.60	19.40	24.34	18.74	19.37
History-RNN [23]	31.15	23.99	19.46	19.59	23.79	20.11	19.67	23.37	**20.50**	19.68
History-FC	30.39	24.49	19.61	19.88	24.38	**20.50**	**20.14**	23.70	20.45	19.91
Ours-MLE	**30.08**	25.16	19.72	20.17	24.50	**20.32**	**20.11**	23.75	**20.47**	19.89
Ours-F1	31.91	**25.16**	**20.24**	**20.42**	24.48	20.26	20.02	24.06	20.33	**19.96**
Ours-GAN	31.60	**25.23**	**20.19**	**20.44**	24.56	20.31	20.08	24.11	20.38	**19.97**

	perp.	beam=1 *pre. %*	beam=1 *rec. %*	beam=1 *F1*	beam=3 *pre. %*	beam=3 *rec. %*	beam=3 *F1*	beam=5 *pre. %*	beam=5 *rec. %*	beam=5 *F1*
Face [27]	18.98	26.48	9.83	12.96	22.41	8.18	10.82	20.74	7.55	10.02
Face+RandText	18.94	26.63	10.01	13.15	22.54	8.15	10.82	20.20	7.43	9.80
Face+Text	17.20	29.46	10.89	14.41	25.84	9.46	12.57	24.82	9.14	12.15
History-RNN [23]	20.30	20.84	7.33	9.81	20.84	7.33	9.81	20.84	7.33	9.81
History-FC	20.26	20.86	7.35	9.83	20.81	7.33	9.80	20.84	7.33	9.81
Ours-MLE	**17.18**	35.81	13.74	18.07	**30.44**	**11.43**	**15.10**	**28.25**	**10.58**	13.49
Ours-F1	17.20	**36.17**	**13.92**	**18.28**	30.42	**11.43**	**15.09**	**28.30**	**10.63**	**14.06**
Ours-GAN	**17.19**	**36.06**	**13.85**	**18.20**	**30.43**	11.38	15.05	28.12	10.52	**13.92**

Table4.2 Main quantitative results on text (left) and gesture (right).

[3] A Face-to-Face Neural Conversation Model, H.Chu, D.Li, S.Fidler, *CVPR* 2018

Overview

- ***Deep Learning Techniques for Digital Content Generation***
- Introduction
  - Motivation
  - Methodology
- City Modeling
  - Houses
  - Road layouts
- Face Avatar
  - Animated conversations
  - Telepresence

Fig1.3. Examples of generated digital content.

Chapter 5: ModularCodecAvatar - Introduction

- Inputs/Outputs
  - In: Head-mounted images
  - Out: Face mesh
- Key idea
  - Modularized face
  - Expressiveness with limited data
  - Dataset: VR-Telepresence

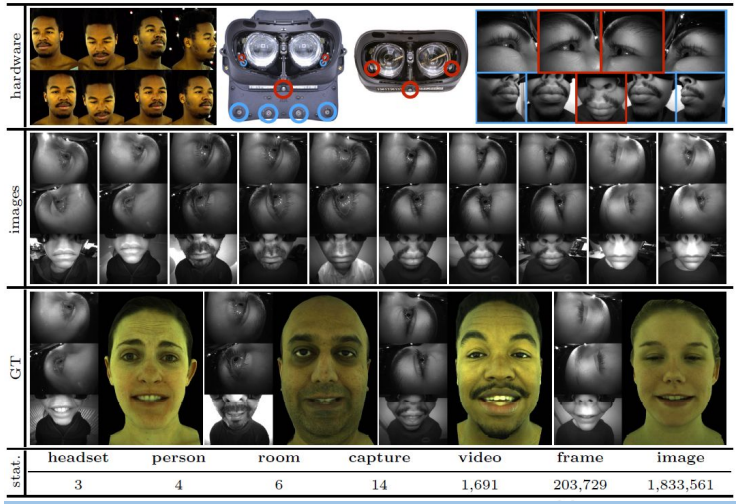


Fig5.1 VR hardware and dataset examples.

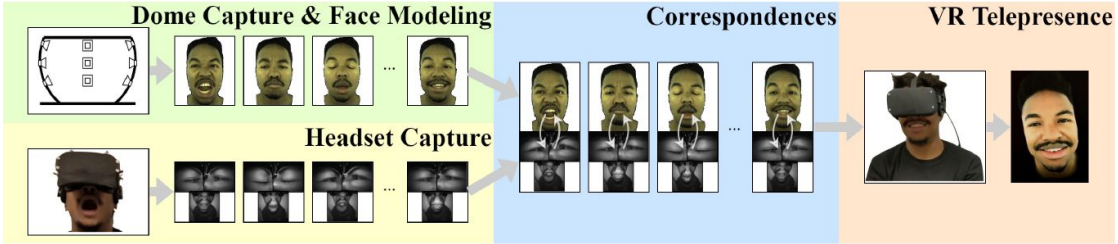


Fig5.2 Overall system pipeline.

[4] Expressive Telepresence via Modular Codec Avatar, H.Chu, S. Ma, F. De la Torre, S. Fidler, Y. Sheikh, ECCV 2020

Chapter 5: ModularCodecAvatar - Methods

- Structured output

  Decoded expressions:

  $$\mathbf{y} = \{\mathbf{y}_1, ...\mathbf{y}_N\}, \mathbf{y}_n \in f_n(\mathbf{c}_n)$$

- Formulation
  - Decomposed face
  - Adaptive blending
- Features
  - Intra-modular
  - Inter-modular

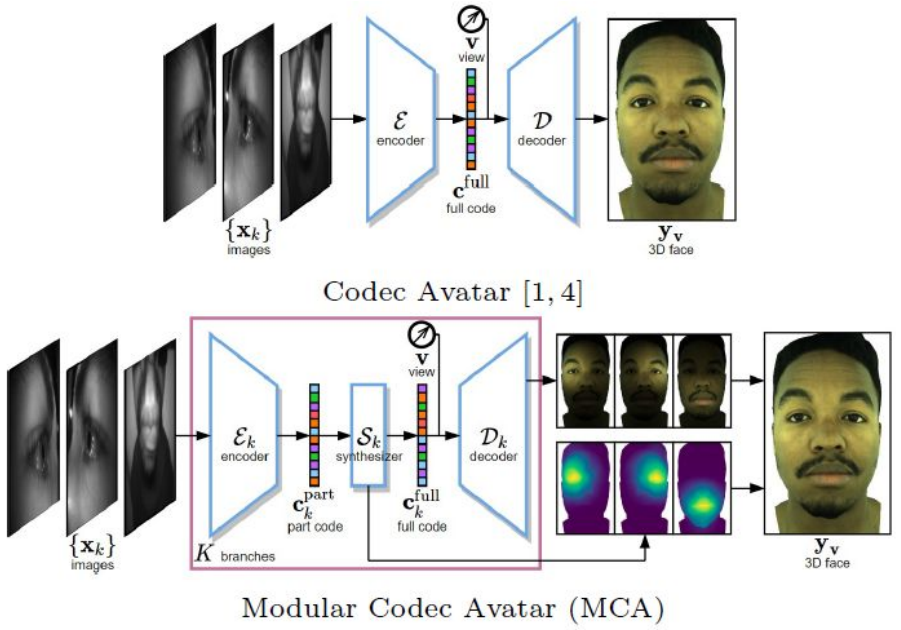


Fig5.3 The MCA model formulation diagram.

[4] Expressive Telepresence via Modular Codec Avatar, H.Chu, S. Ma, F. De la Torre, S. Fidler, Y. Sheikh, ECCV 2020

Chapter 5: ModularCodecAvatar - Results

- Results
  - Improved accuracy
  - Improved robustness
  - New applications
- Issues
  - Temporal smoothness
  - Person specific

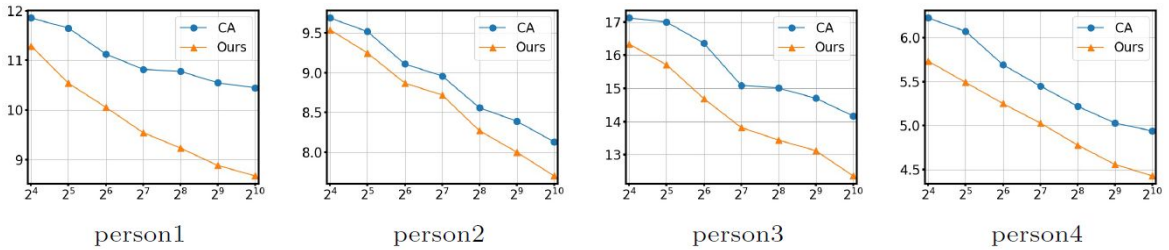
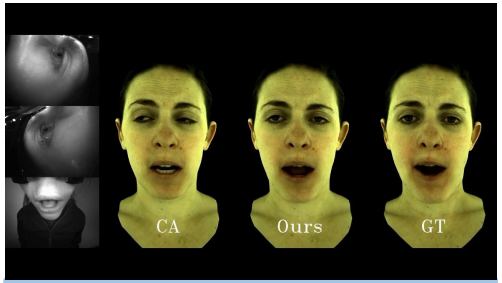


Fig5.4 Expressiveness is crutial in real-world VR telepresence.

	MAE↓		RMSE↓		Geo.↓		Tex.↓		%-better↑		SSIM↑	
method	CA	Ours	CA	Ours	CA	Ours	CA	Ours	CA	Ours	CA	Ours
person1	8.82	**8.69**	7.67	**7.47**	1.26	**1.14**	3.40	**3.02**	36.3	**63.7**	0.954	**0.957**
person2	4.44	**4.26**	4.00	**3.84**	1.82	**1.46**	2.05	**2.04**	27.3	**72.7**	0.949	**0.951**
person3	9.09	**6.97**	8.36	**6.66**	1.14	**0.84**	4.58	**3.43**	0.3	**99.7**	0.933	**0.942**
person4	3.33	**3.21**	3.08	**3.04**	**0.54**	0.64	0.86	**0.85**	41.1	**58.9**	0.984	0.984
overall	6.54	**6.17**	5.81	**5.48**	1.37	**1.17**	2.72	**2.44**	29.3	**70.7**	0.953	**0.956**

Table5.1 Main quantitative results.



Vid5.1 Result demo.

[4] Expressive Telepresence via Modular Codec Avatar, H.Chu, S. Ma, F. De la Torre, S. Fidler, Y. Sheikh, ECCV 2020