Deep Transfer Learning for Visual Analysis

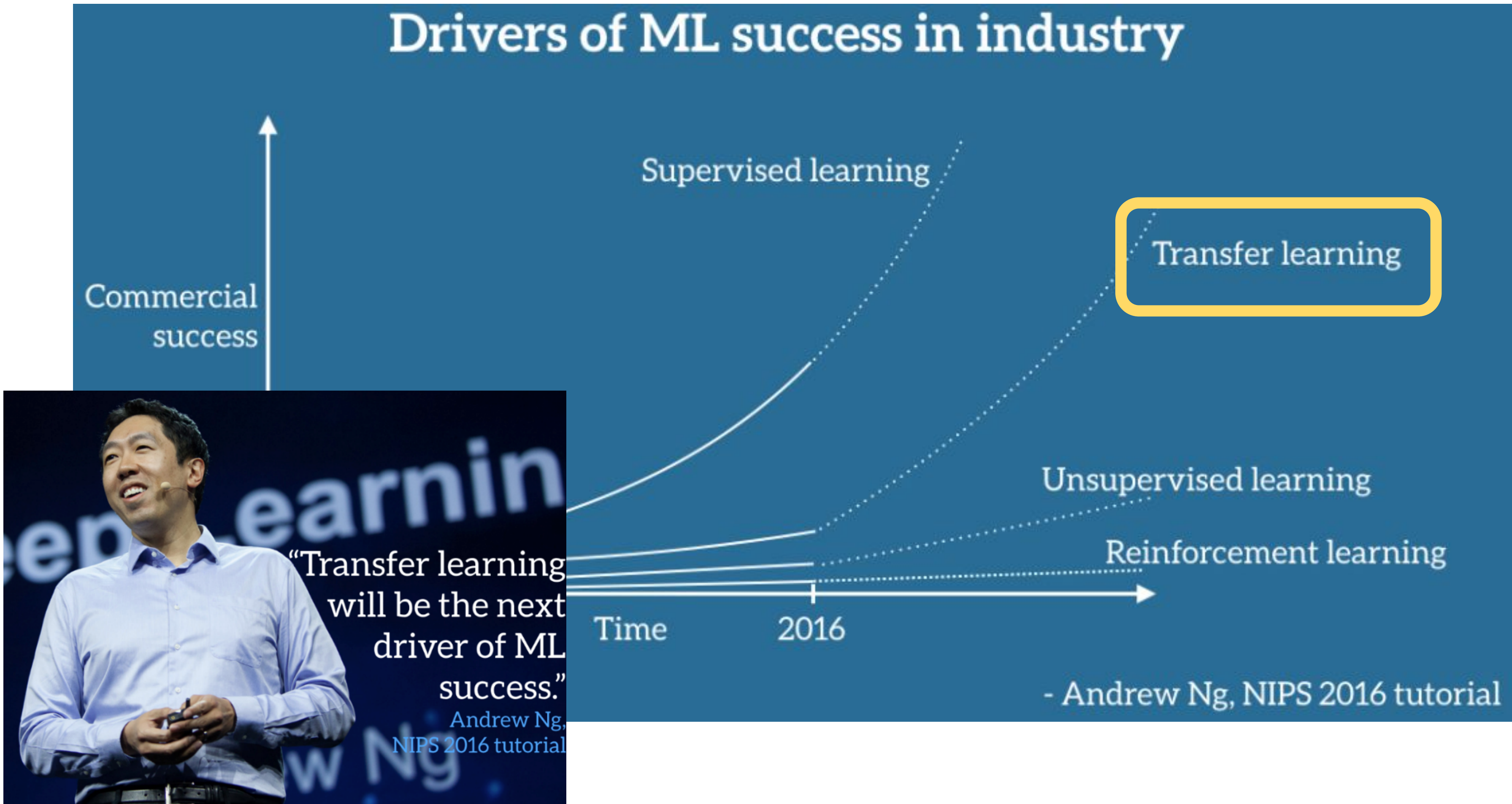
Yu-Chiang Frank Wang, Associate Professor

Dept. Electrical Engineering, National Taiwan University

Taipei, Taiwan

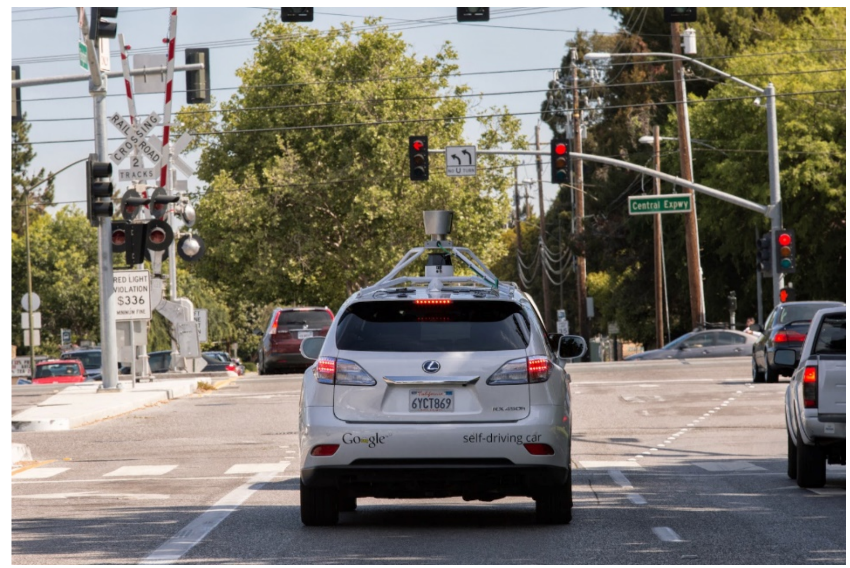
2018/5/19 2nd AII Workshop

Trends of Deep Learning



Transfer Learning: What, When, and Why? (cont'd)

- A practical example



https://techcrunch.com/2017/02/08/udacity-open-sources-its-self-driving-car-simulator-for-anyone-to-use/
https://googleblog.blogspot.tw/2014/04/the-latest-chapter-for-self-driving-car.html

Recent Research Focuses on Transfer Learning

- CVPR 2018
  Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation
- AAAI 2018
  Order-Free RNN with Visual Attention for Multi-Label Classification
- CVPR 2018
  Multi-Label Zero-Shot Learning with Structured Knowledge Graphs
- CVPRW 2018
  Unsupervised Deep Transfer Learning for Person Re-Identification

Detach & Adapt – Beyond Image Style Transfer

- Faceapp – Putting a smile on your face!
  - Deep learning for representation disentanglement
  - Interpretable deep feature representation

Input
Mr. Takeshi Kaneshiro

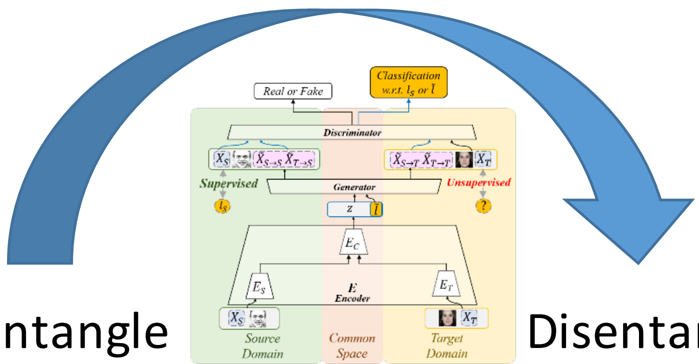
Detach & Adapt – Beyond Image Style Transfer

- Cross-domain image synthesis, manipulation & translation

With supervision

w/o supervision

Transfer



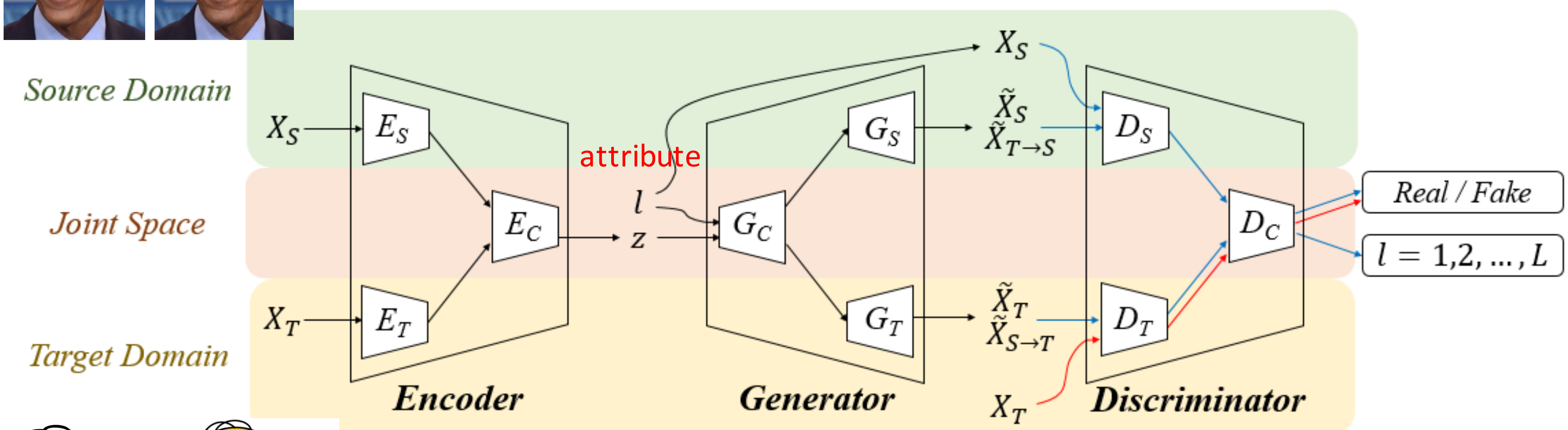
Disentangle smile from **Photo**

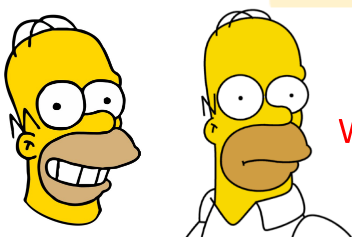
Disentangle smile from **Cartoon**

Detach & Adapt – Beyond Image Style Transfer

- Cross-domain image synthesis, manipulation & translation [CVPR'18]

With supervision

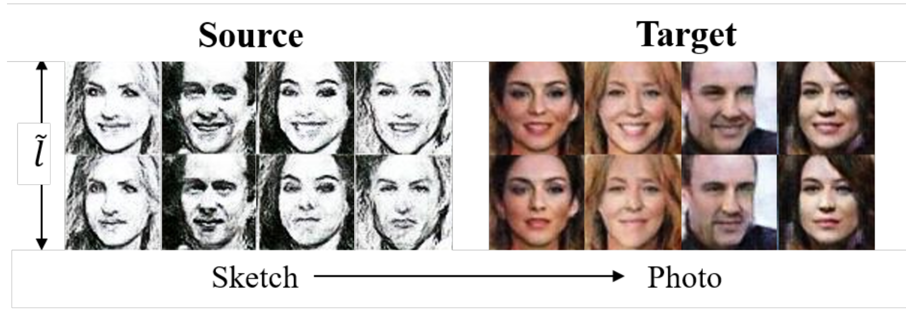




W/o supervision

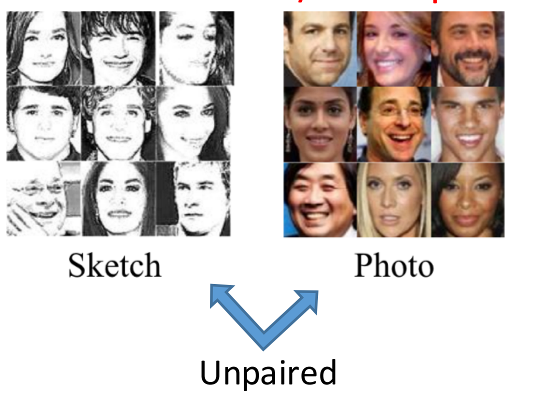
Example Results

- Face
  - Photo & Sketch



- Conditional Unsupervised Image Translation

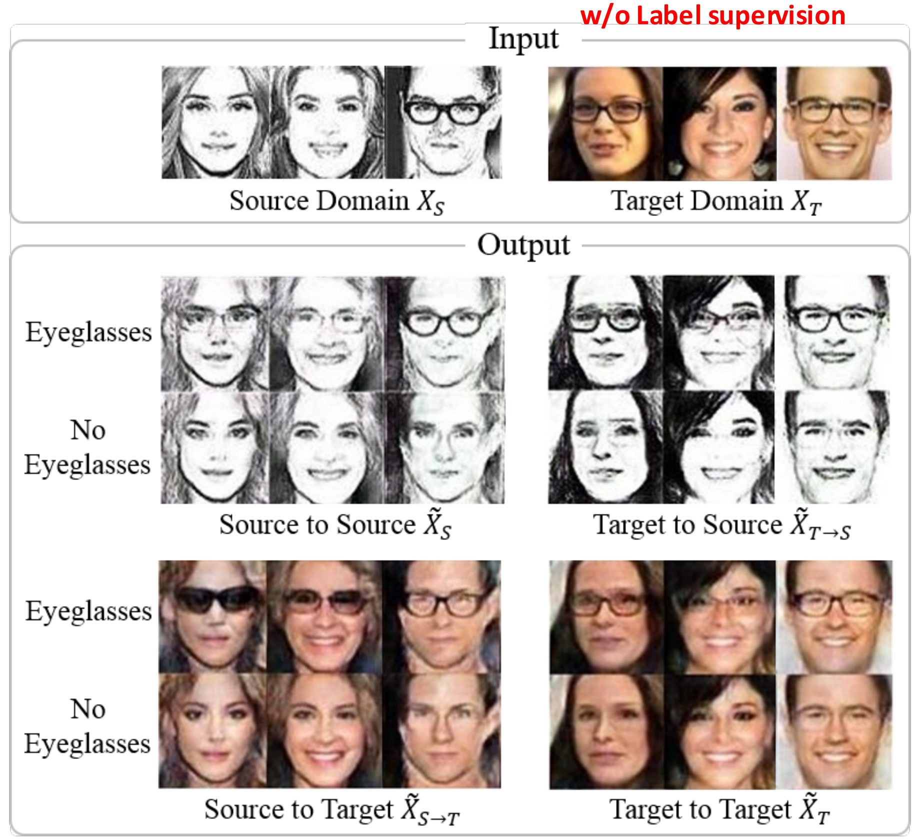
w/o Label supervision



(a) Faces.

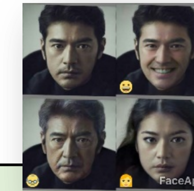
Domain	$\tilde{l}$	CoGAN	UNIT	CDRD	E-CDRD
sketch ($\mathcal{S}$)	smiling	89.50	90.10	90.19	90.01
photo ($\mathcal{T}$)	-	78.90	81.04	**87.61**	**88.28**
sketch ($\mathcal{S}$)	glasses	96.63	97.65	97.06	97.19
photo ($\mathcal{T}$)	-	81.01	79.89	**94.49**	**94.84**

w/o Label supervision



Y.-C. F. Wang et al., Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation, CVPR 2018

Comparisons



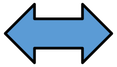
	Cross-Domain Image Translation				Representation Disentanglement	
	Unpaired Training Data	Multi-domains	Bi-direction	Joint Representation	Unsupervised	Interpretability of disentangled factor
Pix2pix	X	X	X	X	Cannot disentangle image representation	
CycleGAN	O	X	O	X		
StarGAN	O	O	O	X		
UNIT	O	X	O	O		
DTN	O	X	X	O		
infoGAN	Cannot translate images across domains				O	X
AC-GAN					X	O
CDRD (Ours)	O	O	O	O	Partially	O

Recent Research Focuses on Transfer Learning

- CVPR 2018
  Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation
- AAAI 2018
  Order-Free RNN with Visual Attention for Multi-Label Classification
- CVPR 2018
  Multi-Label Zero-Shot Learning with Structured Knowledge Graphs Appier
- CVPRW 2018
  Unsupervised Deep Transfer Learning for Person Re-Identification

Multi-Label Classification for Image Analysis

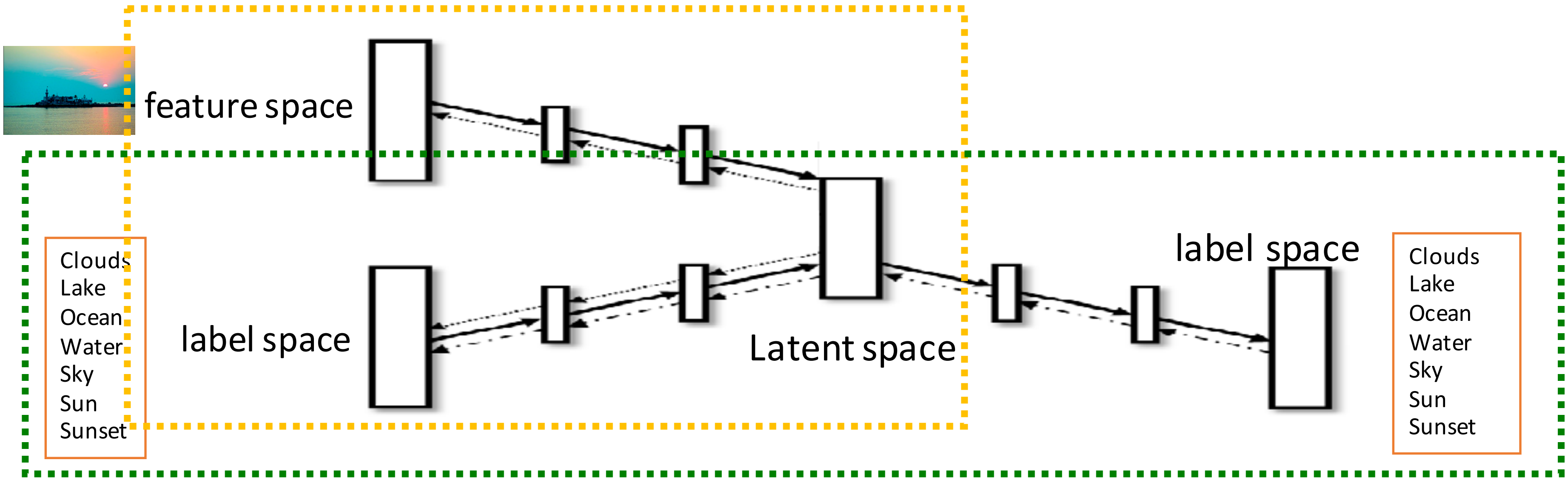
- Prediction of multiple object labels from an image
  - Learning across image and semantics domains
  - No object detectors available
  - Desirable if be able to exploit label co-occurrence info



Labels:
Person
Table
Sofa
Chair
TV
Lights
Carpet
…

DNN for Multi-Label Classification

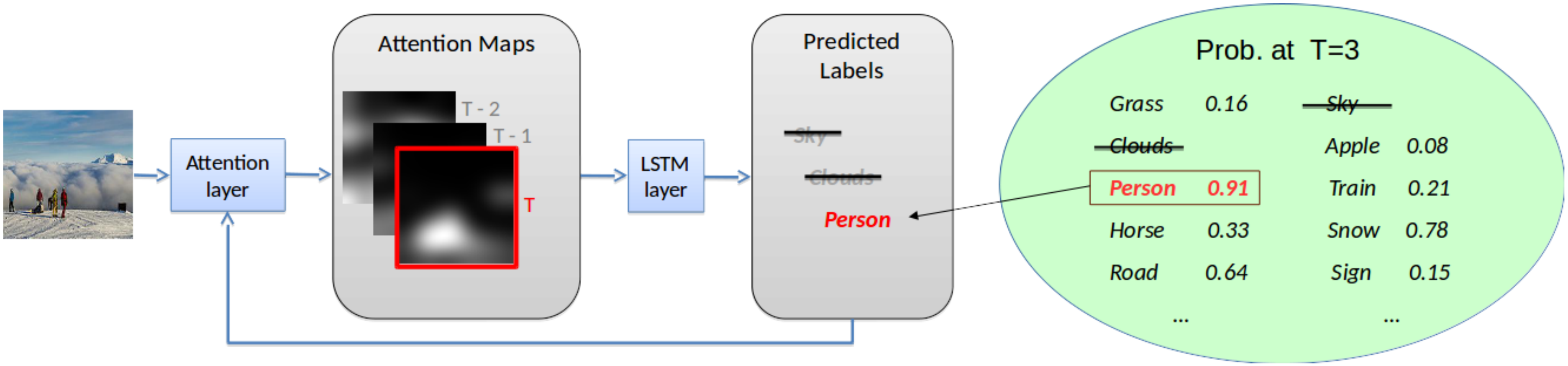
- Canonical-Correlated Autoencoder (C2AE) [Wang et al., AAAI 2017]
  - Unique integration of autoencoder & deep canonical correlation analysis (DCCA)
  - Autoencoder: label embedding + label recovery + label co-occurrence
  - DCCA: joint feature & label embedding
  - Can handle missing labels during learning

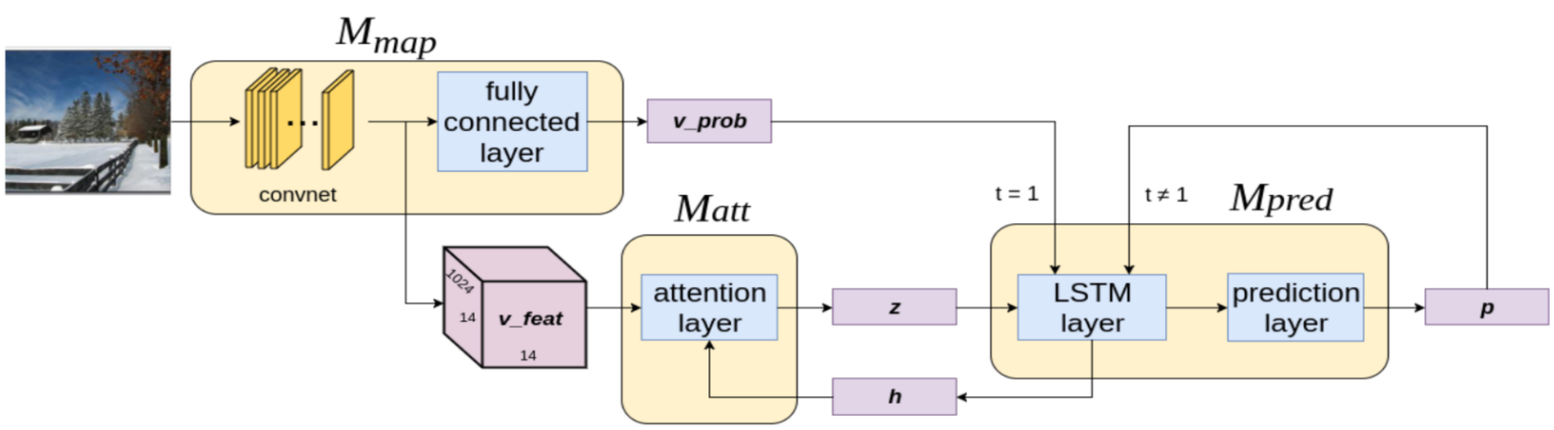


Y.-C. F. Wang et al., Learning Deep Latent Spaces for Multi-Label Classification, AAAI 2017

Order-Free RNN with Visual Attention for Multi-Label Classification [AAAI'18]

- Visual Attention for MLC [Wang et al., AAAI'18]





Y.-C. F. Wang et al., Order-Free RNN with Visual Attention for Multi-Label Classification, AAAI 2018

Order-Free RNN with Visual Attention for Multi-Label Classification

- Experiments
  - NUS-WIDE: 269,648 images with 81 labels
  - MS-COCO: 82,783 images with 80 labels
- Quantitative Evaluation

NUS-WIDE

Method	C-P	C-R	C-F1	O-P	O-R	O-F1
KNN	32.6	19.3	24.3	43.9	53.4	47.6
Softmax	31.7	31.2	31.4	47.8	59.5	53.0
WARP	31.7	35.6	33.5	48.6	60.5	53.9
CNN-RNN	40.5	30.4	34.7	49.9	61.7	55.2
Resnet-baseline	46.5	47.6	47.1	61.6	68.1	64.7
Frequency-first (w/ atten)	48.9	48.7	48.8	62.1	69.4	65.5
Rare-first (w/ atten)	53.9	51.8	52.8	55.1	65.2	59.8
Ours (w/o atten)	60.8	49.5	54.5	68.3	72.4	70.2
Ours	59.4	50.7	**54.7**	69.0	71.4	**70.2**

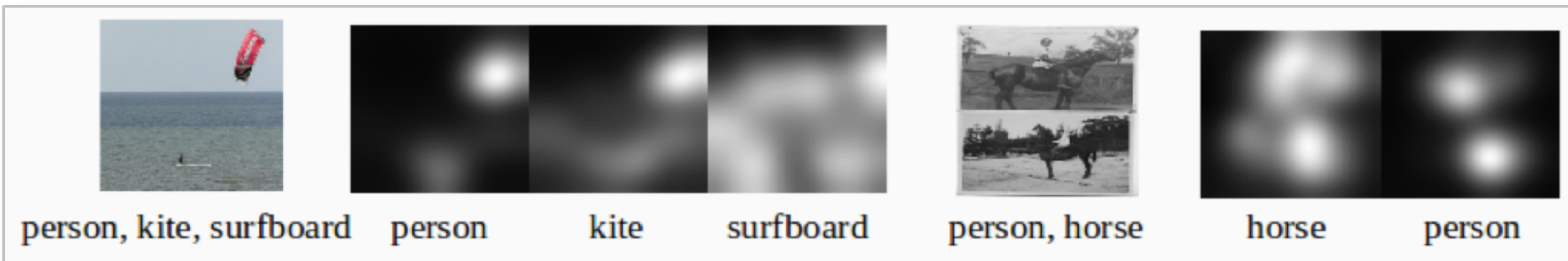
MS-COCO

Method	C-P	C-R	C-F1	O-P	O-R	O-F1
Softmax	59.0	57.0	58.0	60.2	62.1	61.1
WARP	59.3	52.5	55.7	59.8	61.4	60.7
CNN-RNN	66.0	55.6	60.4	69.2	66.4	**67.8**
Resnet-baseline	58.3	49.3	53.4	63.9	58.4	61.0
Frequency-first (w/ atten)	55.8	54.7	55.2	61.4	62.6	62.0
Rare-first (w/ atten)	59.5	56.5	58.0	57.3	56.7	57.0
Ours (w/o atten)	69.9	52.6	60.0	73.4	60.3	66.2
Ours	71.6	54.8	**62.1**	74.2	62.2	**67.7**

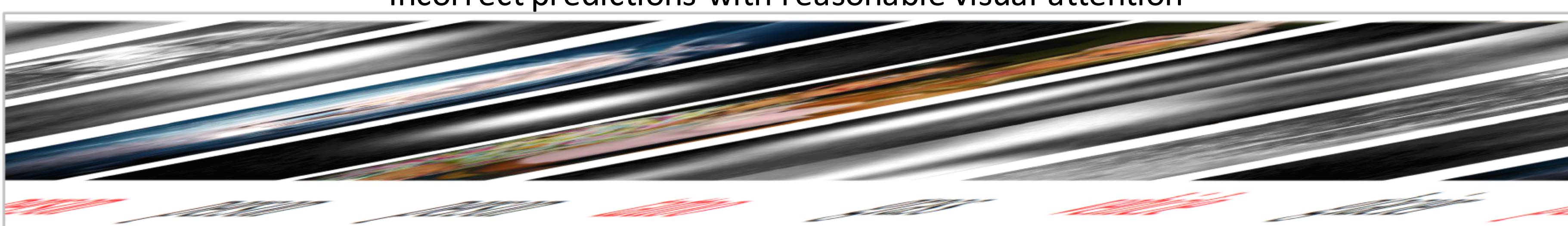
Order-Free RNN with Visual Attention for Multi-Label Classification

- Qualitative Evaluation

Example images in MS-COCO with the associated attention maps

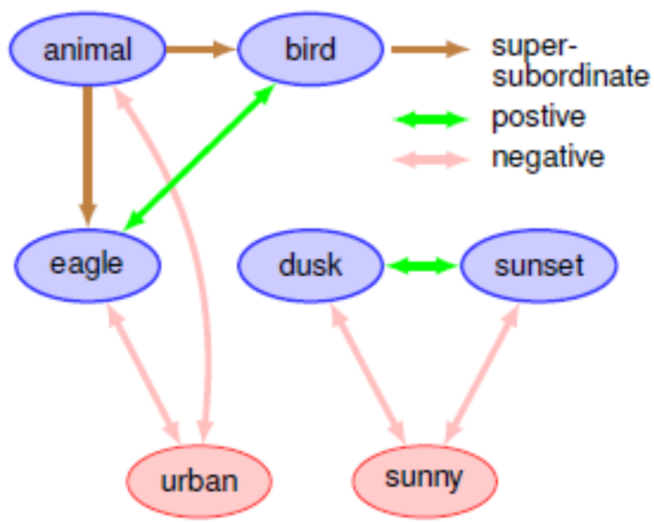


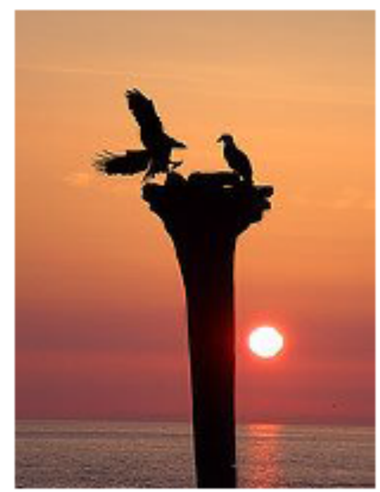
Incorrect predictions with reasonable visual attention

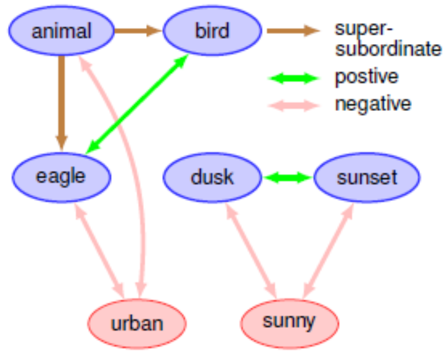


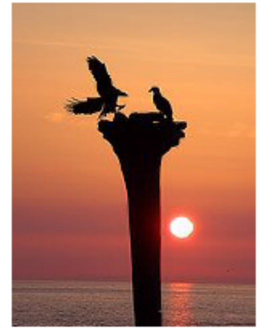
Multi-Label Zero-Shot Learning with Structured Knowledge Graphs [CVPR’18]

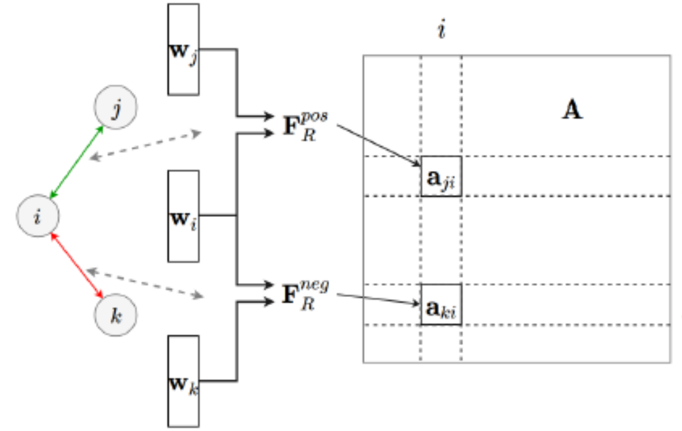
- Utilizing structured knowledge graphs for modeling label dependency



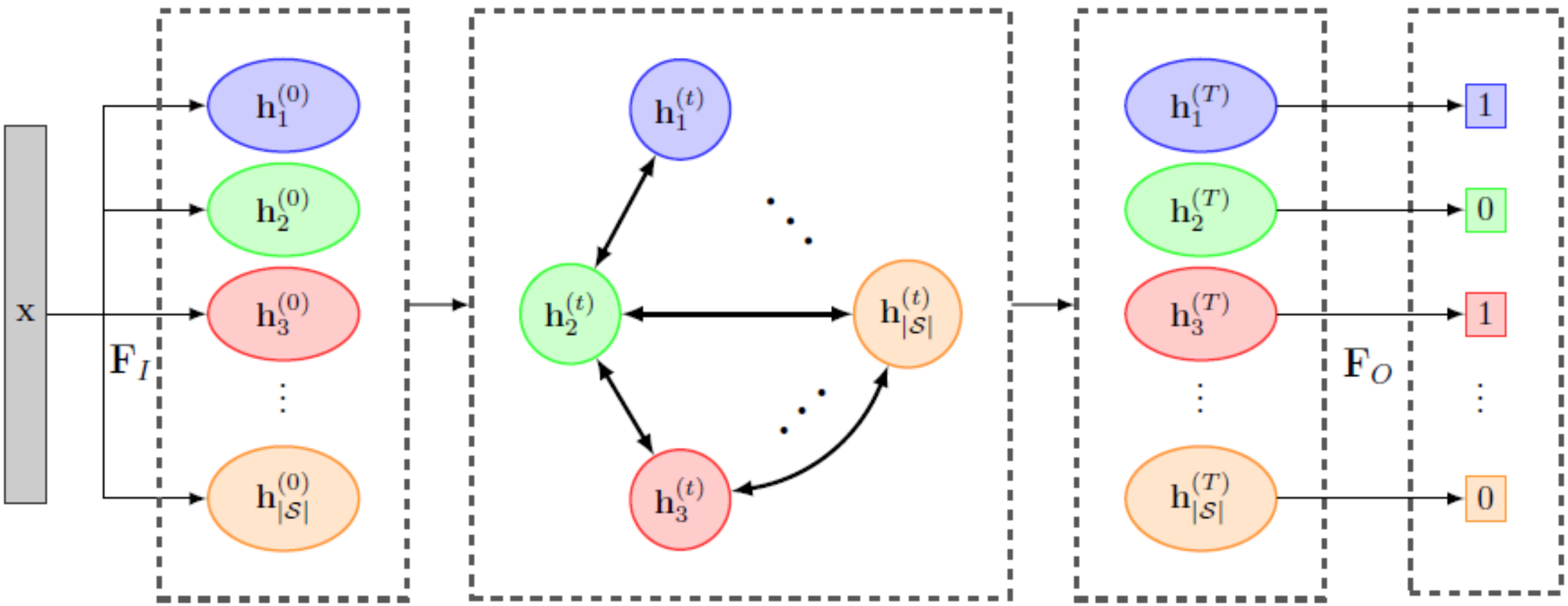


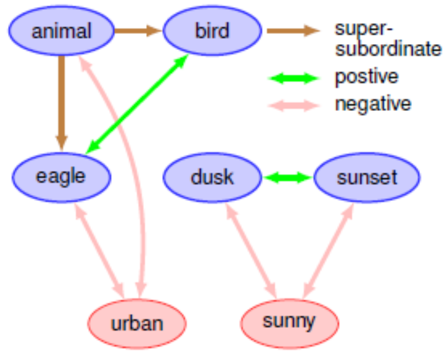


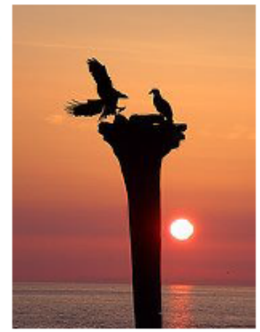


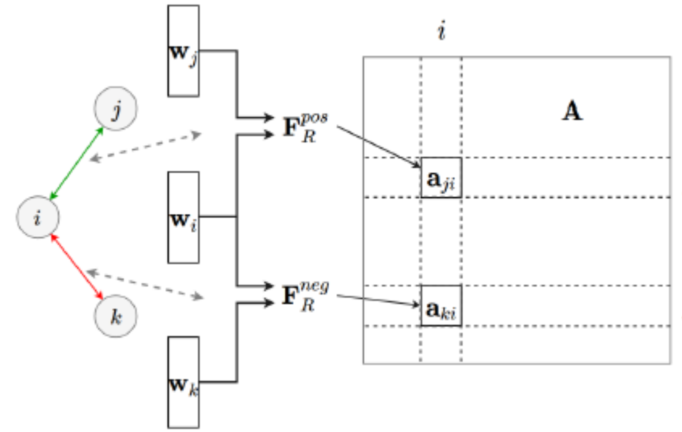


- Our Proposed Network

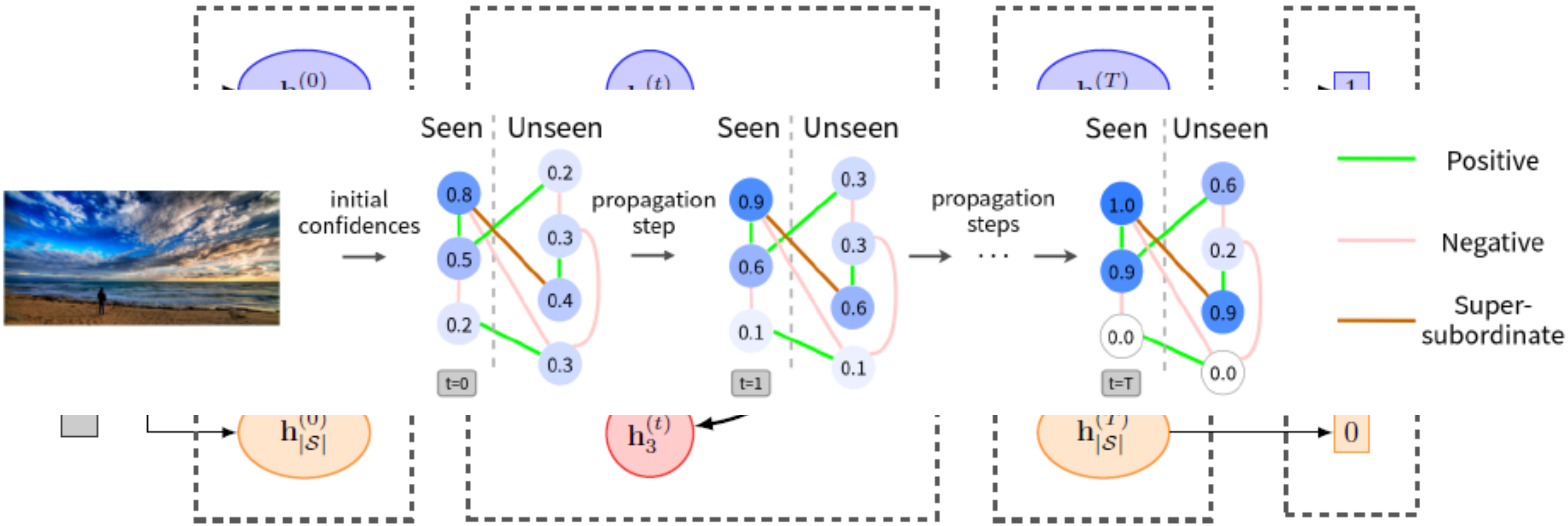








- Our Proposed Network



Order-Free RNN with Visual Attention for Multi-Label Classification

- Experiments
  - NUS-WIDE: 269,648 images with 1000 labels
  - MS-COCO: 82,783 images with 80 labels
- Quantitative Evaluation
  - ML vs. ML-ZSL vs. Generalized ML-ZSL

	NUS-81			MS-COCO		
Method	P	R	F1	P	R	F1
WSABIE	30.7	52.0	38.6	59.3	61.3	60.3
WARP	31.4	53.3	39.5	60.2	62.2	61.2
Logistics	41.9	46.2	43.9	70.8	63.3	66.9
Fast0Tag	31.9	54.0	40.1	60.2	62.2	61.2
Ours	43.4	48.2	**45.7**	74.1	64.5	**69.0**

	ML-ZSL			Generalized		
Method	P	R	F1	P	R	F1
Fast0Tag ($K = 3$)	21.7	37.7	27.2	-	-	-
Fast0Tag ($K = 10$)	-	-	-	19.5	24.9	21.9
Ours w/o Prop.	31.8	25.1	28.1	24.3	23.4	23.9
Ours	29.3	31.9	**30.6**	22.8	25.9	**24.2**

Recent Research Focuses on Transfer Learning

- CVPR 2018
  Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation
- AAAI 2018
  Order-Free RNN with Visual Attention for Multi-Label Classification
- CVPR 2018
  Multi-Label Zero-Shot Learning with Structured Knowledge Graphs
- CVPRW 2018
  Unsupervised Deep Transfer Learning for Person Re-Identification

umbo
computer vision

Introduction: Person re-identification

Camera #1

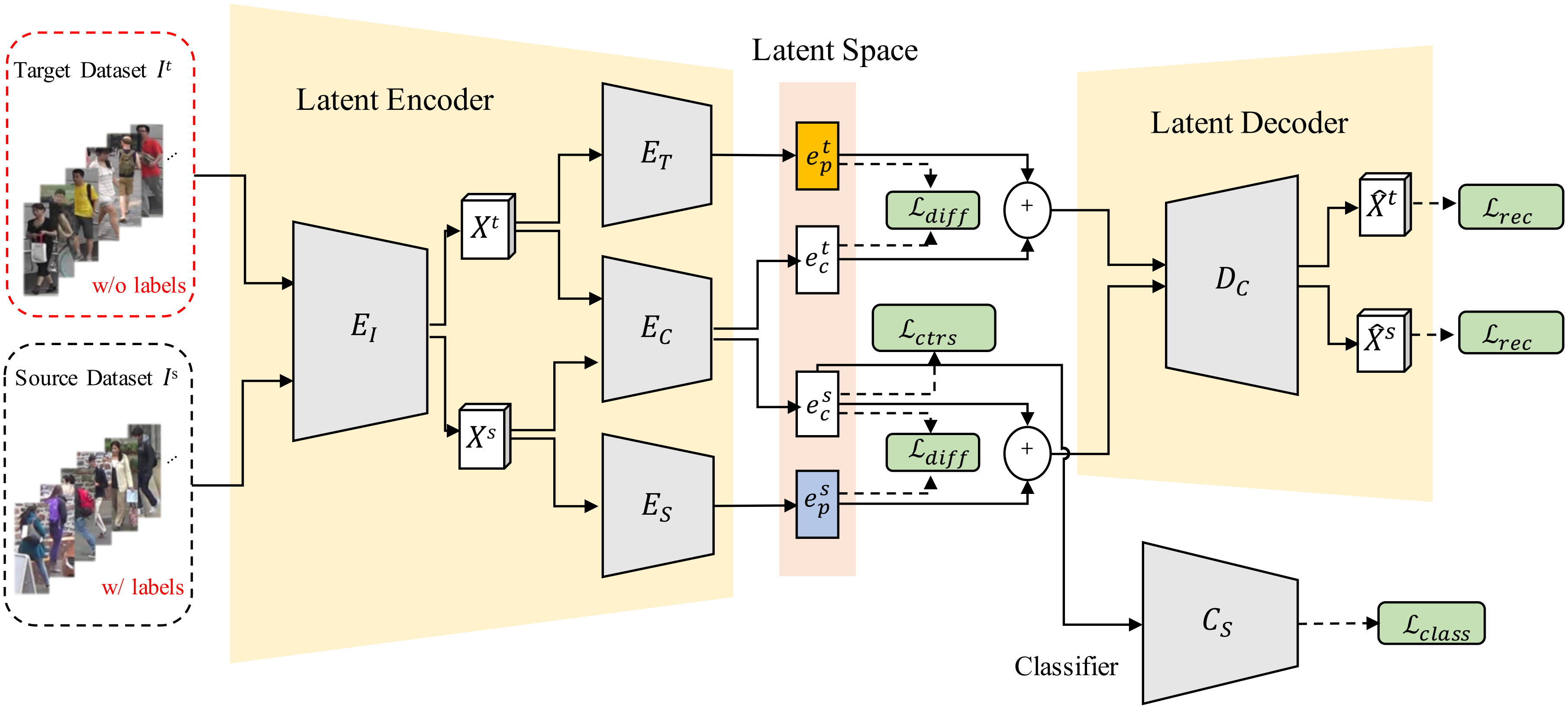
Camera #3

Camera #2

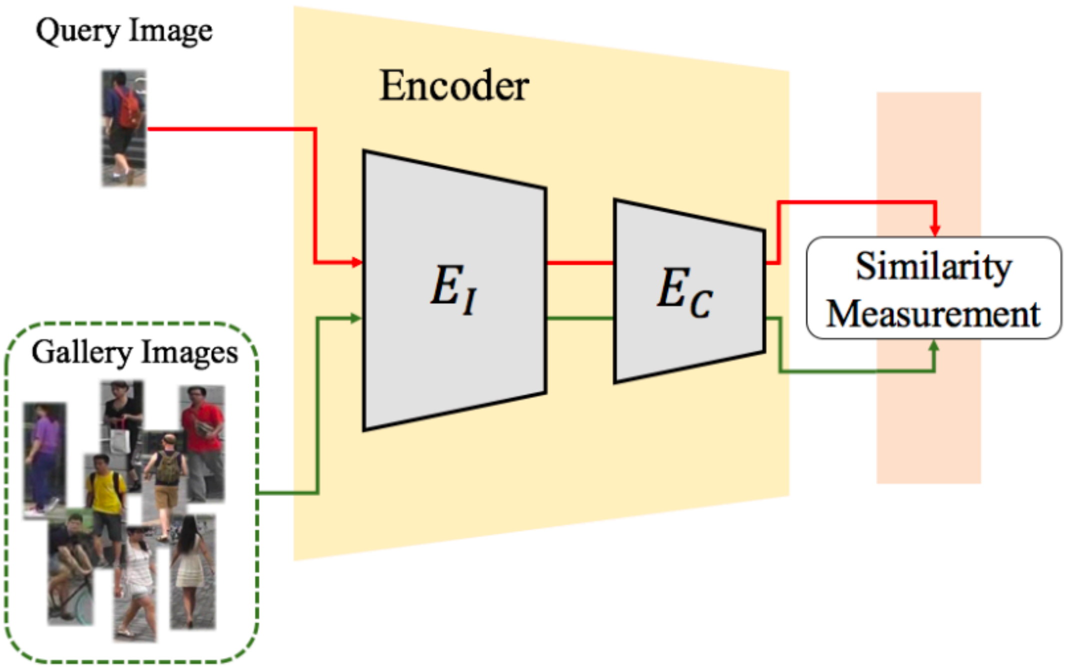
Camera #4

Person re-identification task: the system needs to **match** appearances of a **person of interest** across **non-overlapping** cameras.

Adaptation & Re-ID Network



Testing Scenario



Comparisons with Recent Re-ID Methods

Table 3: Performance comparisons on Market-1501 with supervised and unsupervised Re-ID methods.

	Method	Rank-1	Rank-5	Rank-10	mAP
Supervised	BOW [20]	44.4	-	-	20.8
	LDNS [19]	61.0	-	-	35.7
	SVDNET [15]	82.3	-	-	62.1
	TriNet [7]	84.9	-	-	69.1
	CamStyle [23]	89.5	-	-	71.6
	DuATM [14]	91.4	-	-	76.6
Unsupervised	BOW [20]	35.8	52.4	60.3	14.8
	UMDL [13]	34.5	52.6	59.6	12.4
	PUL [4]	45.5	60.7	66.7	20.5
	CAMEL [18]	54.5	-	-	26.3
	SPGAN [3]	57.7	75.8	82.4	26.7
	Ours	**70.3**	**80.4**	**86.3**	**39.4**

Table 4: Performance comparisons on DukeMTMC-reID with supervised and unsupervised Re-ID methods.

	Method	Rank-1	Rank-5	Rank-10	mAP
Supervised	BOW [20]	25.1	-	-	12.2
	LOMO [8]	30.8	-	-	17.0
	TriNet [7]	72.4	-	-	53.5
	SVDNET [15]	76.7	-	-	56.8
	CamStyle [23]	78.3	-	-	57.6
	DuATM [14]	81.8	-	-	64.6
Unsupervised	BOW [20]	17.1	28.8	34.9	8.3
	UMDL [13]	18.5	31.4	37.6	7.3
	PUL [4]	30.0	43.4	48.5	16.4
	SPGAN [3]	46.4	62.3	68.0	26.2
	Ours	**60.2**	**73.9**	**79.5**	**33.4**

Recent Research Focuses on Transfer Learning

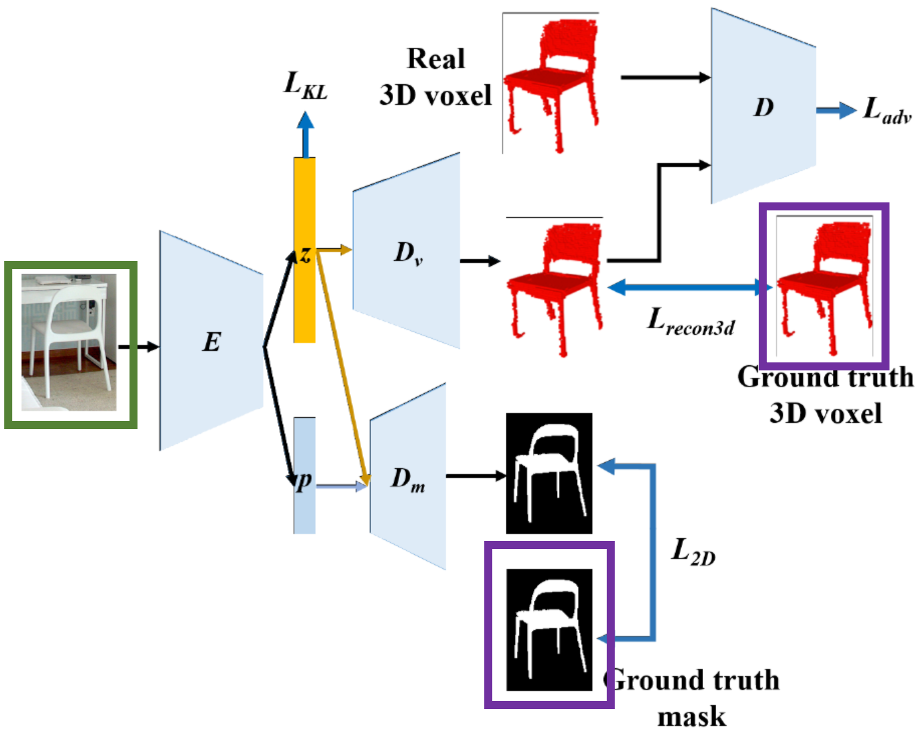
- AAAI 2018
  Order-Free RNN with Visual Attention for Multi-Label Classification
- CVPR 2018
  Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation
- CVPR 2018
  Multi-Label Zero-Shot Learning with Structured Knowledge Graphs
- CVPRW 2018
  Unsupervised Deep Transfer Learning for Person Re-Identification

Other Ongoing Research Topics

- Take a Deep Look from a Single Image
  - Single-Image 3D Object Model Prediction
  - Completing Videos from a Deep Glimpse

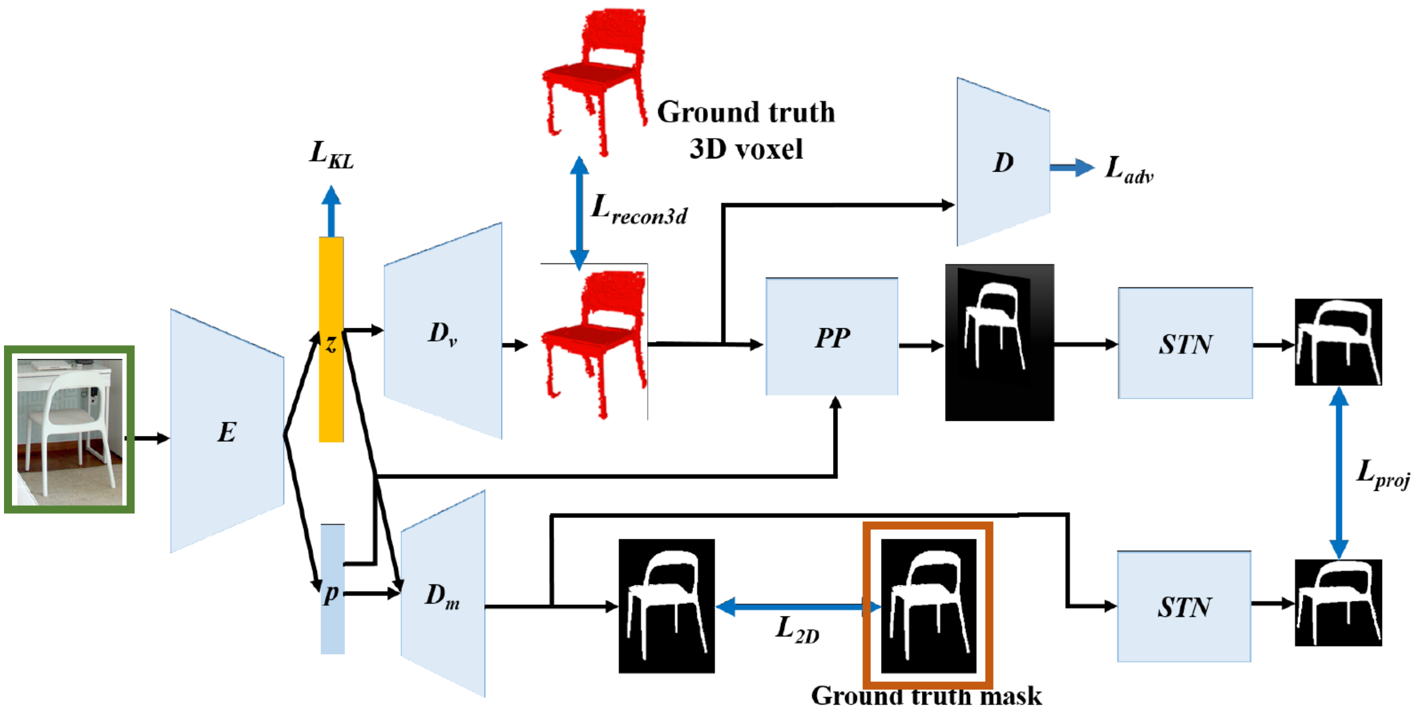
3D Shape Estimation from A Single 2D Image

- Recovering Shape from a Single Image
  - Supervised Setting
    - Input image and its ground truth 3D voxel available for training



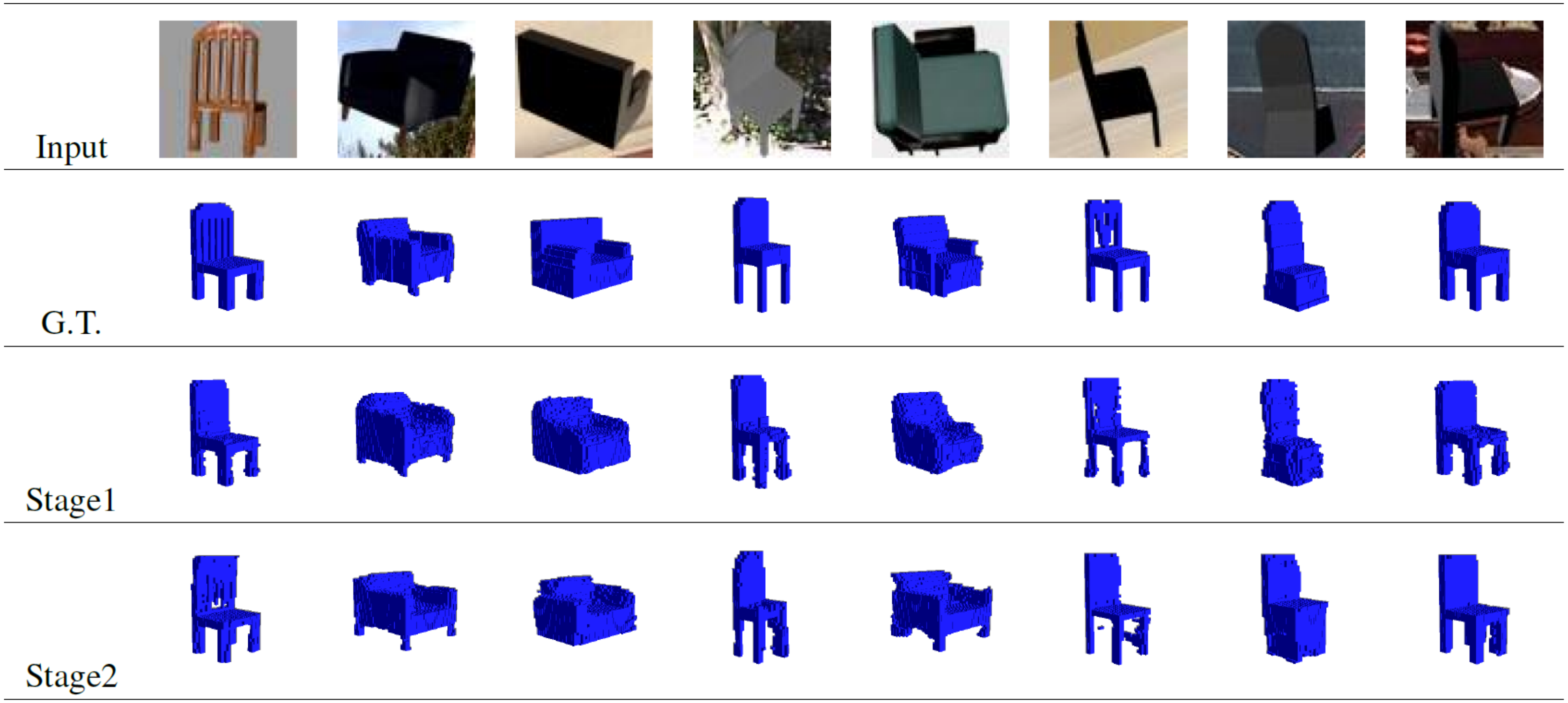
3D Shape Estimation from A Single 2D Image

- Recovering Shape from a Single Image
  - Semi-Supervised Setting
    - Input image and its ground truth 2D mask available for training



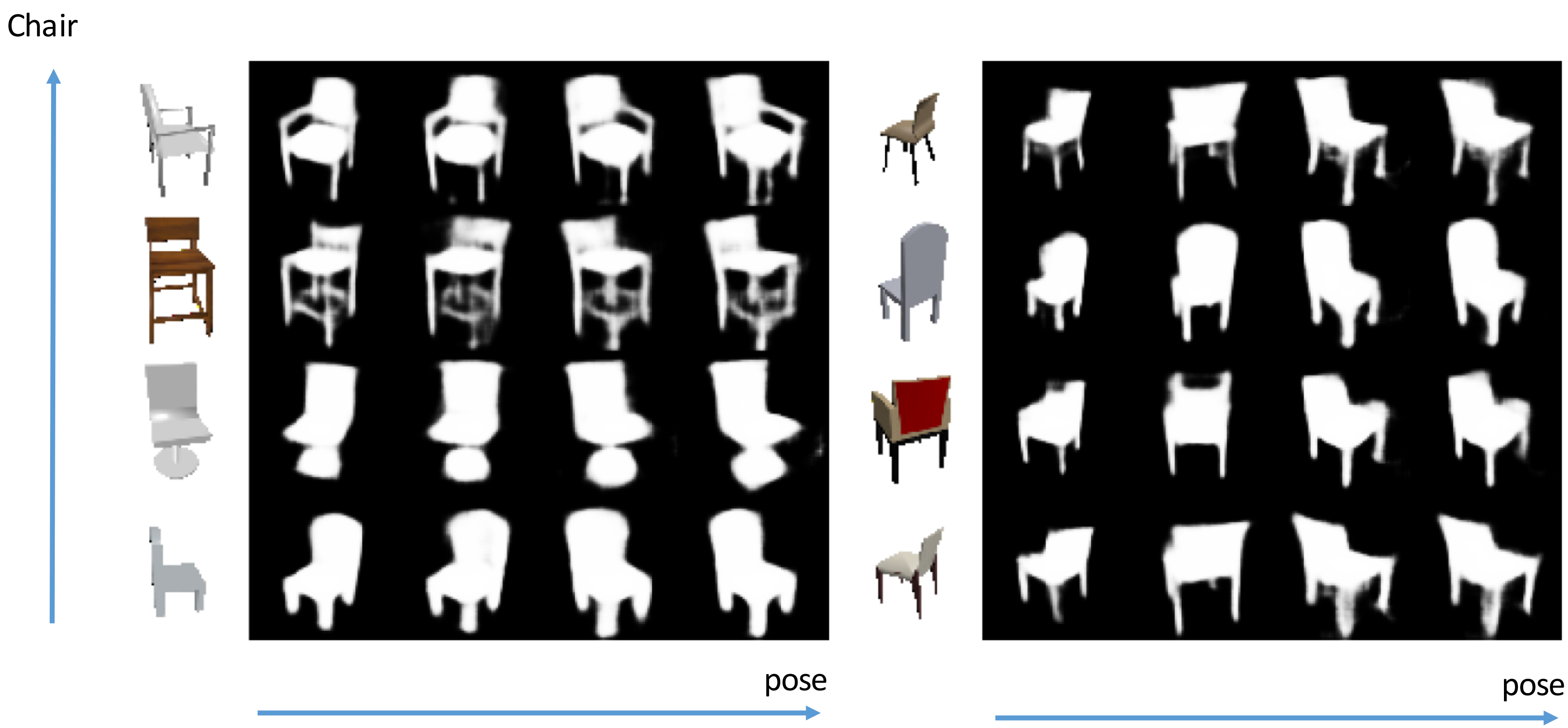
3D Shape Estimation from A Single 2D Image

- Example Results



3D Shape Estimation from A Single 2D Image

- Example Results

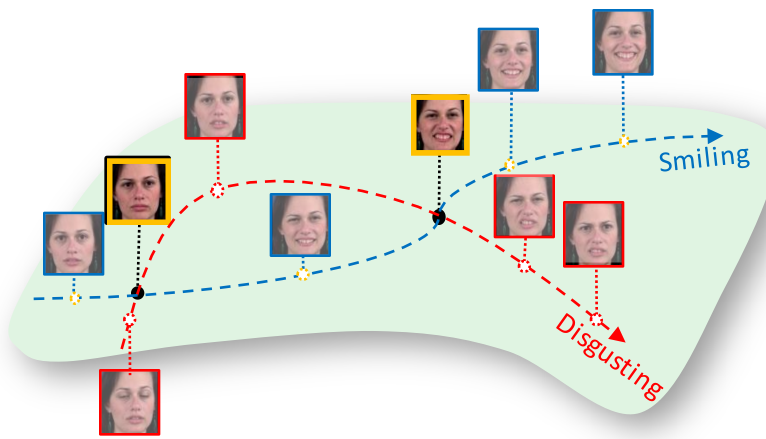


Recent Research Focuses

- Take a Deep Look from a Single Image
  - Single-Image 3D Object Model Prediction
  - Completing Videos from a Deep Glimpse

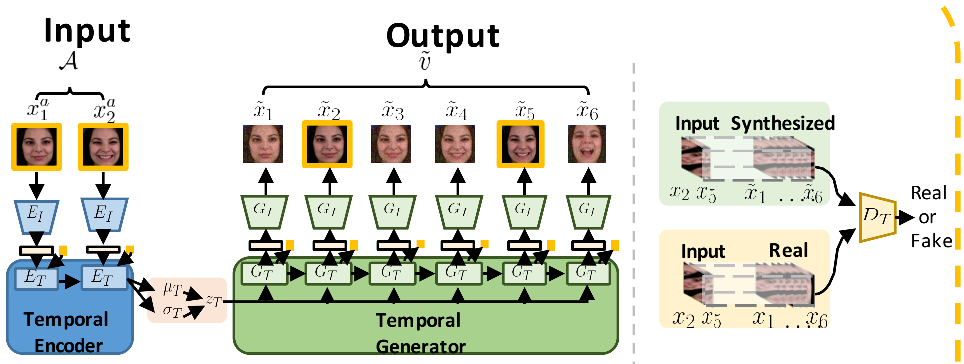
What’s Video Completion?

Input: **Anchor Images**



From Video Synthesis to Completion

- Our Proposed Network
  - Variational autoencoder, recurrent neural nets, and GAN



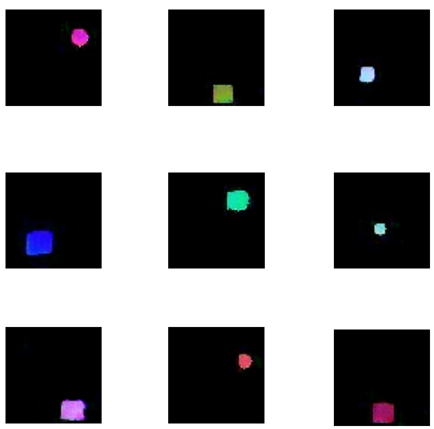
Stochastic & Recurrent Conditional-GAN (SR-cGAN)

Input: non-consecutive frames of interest
Output: video sequence
(more than one possible output)

Three Stages in Learning
1. Learning frame-based representation
2. Learning video-based representation
3. Learning video representation conditioned on input anchor frames

Video Synthesis

ACD	Shape Motion	Facial Expressions
Reference	0	0.116
VGAN [1]	5.02	0.322
TGAN [2]	2.08	0.305
MoCoGAN [3]	1.79	0.201
Ours	**1.05**	**0.137**



Shape Motion

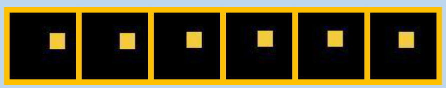
KTH

MUG

Video Completion – Example Results

Shape Motion

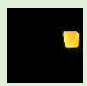
Input (Anchor Frames)



$t_i =$ 6 7 11 12 14 15

Output (Synthesized Video)

GIF



$\tilde{t}_i =$ 6 7 11 12 14 15

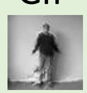
KTH

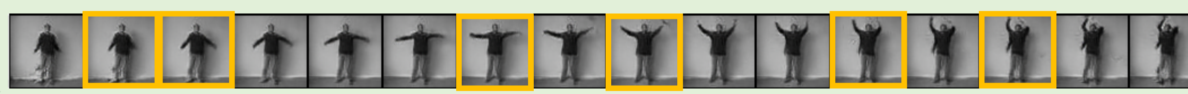
Input (Anchor Frames)

$t_i =$ 2 3 7 9 12 14

Output (Synthesized Video)

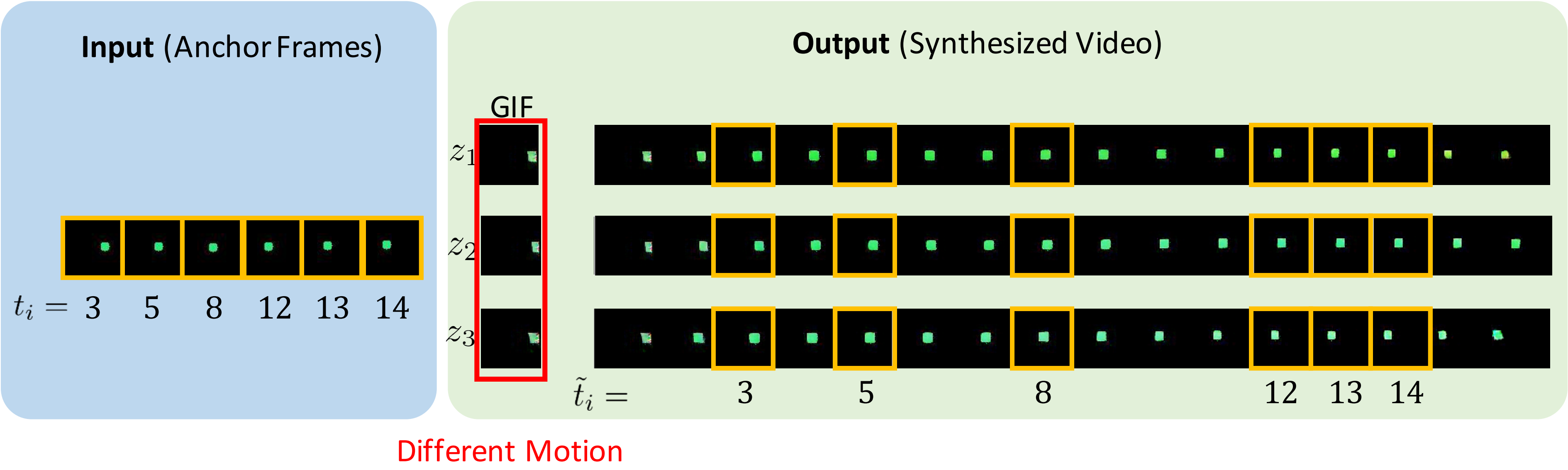
GIF





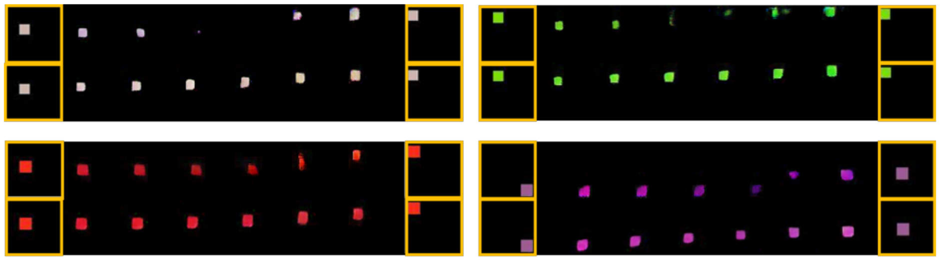
$\tilde{t}_i =$ 2 3 7 9 12 14

Video Completion - Stochasticity

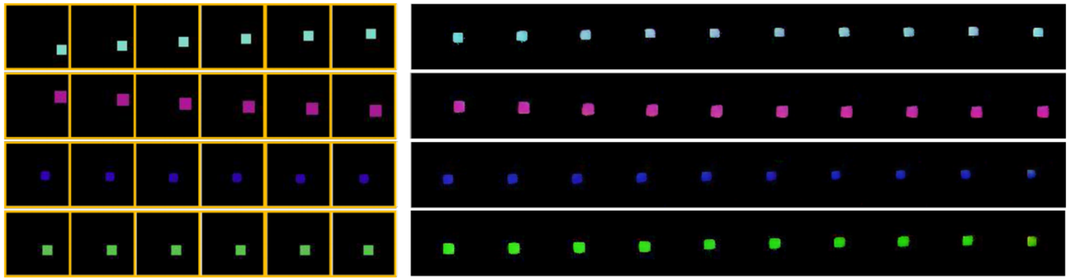


Video Interpolation & Prediction

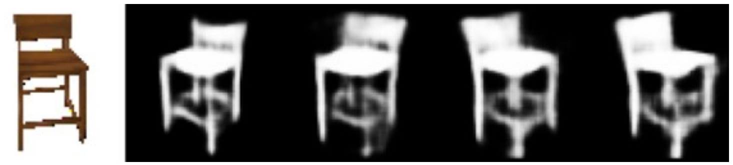
- Interpolation
  - Input:
    - 2 anchor frames
    - fixed on t=1 and 8
  - Output 8 frames



- Prediction
  - Input:
    - 6 anchor frames
    - Fixed on t=1~6
  - Output 16 frames



Summary



- Deep Transfer Learning for Visual Analysis
  - Multi-Label Classification for Image Analysis
  - Detach and Adapt – Beyond Image Style Transfer
  - Single-Image 3D Object Model Prediction
  - Completing Videos from a Deep Glimpse

Sketch Photo

Person
Table
Sofa
Chair
TV
Lights
Carpet
…

For More Information…

- Vision and Learning Lab at NTUEE (http://vllab.ee.ntu.edu.tw/)

Thank You!