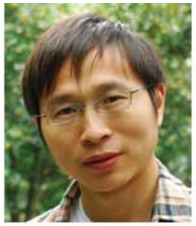
Person Re-ID: Recent Advances and Challenges

Session 3:

Benchmark and GANs in Person ReID

Shiliang Zhang
Peking University
Beijing, China



Jingdong Wang
Microsoft Research
Beijing, China

Qi Tian
University of Texas at
San Antonio, USA

Wen Gao
Peking University
Beijing, China

Longhui Wei
Peking University
Beijing, China

Presented by Longhui Wei
Peking University

Outline

- Our Benchmark Solution
- The Application of GANs in Person ReID
  - Overview of GANs
  - Our solution
- New Research Possibilities

Outline

- Our Benchmark Solution
- The Application of GANs in Person ReID
  - Overview of GANs
  - Our solution
- New Research Possibilities

1. Benchmark Solution

- **Four person re-id datasets are progressively constructed**

- Market-1501 dataset for image-based re-identification
  - 1,501 identities, 500k distractor images – 【ICCV 2015】

- MARS dataset for video-based re-identification
  - 1,261 identities, over 20k tracklets - 【ECCV 2016】

- PRW dataset for end-to-end person re-identification
  - 932 identities, # boxes depend on the detector. - 【CVPR 2017】

- MSMT17 dataset for more realistic re-identification
  - 4,101 identities, 126,441 bounding boxes. - 【CVPR 2018】

1. Benchmark Solution

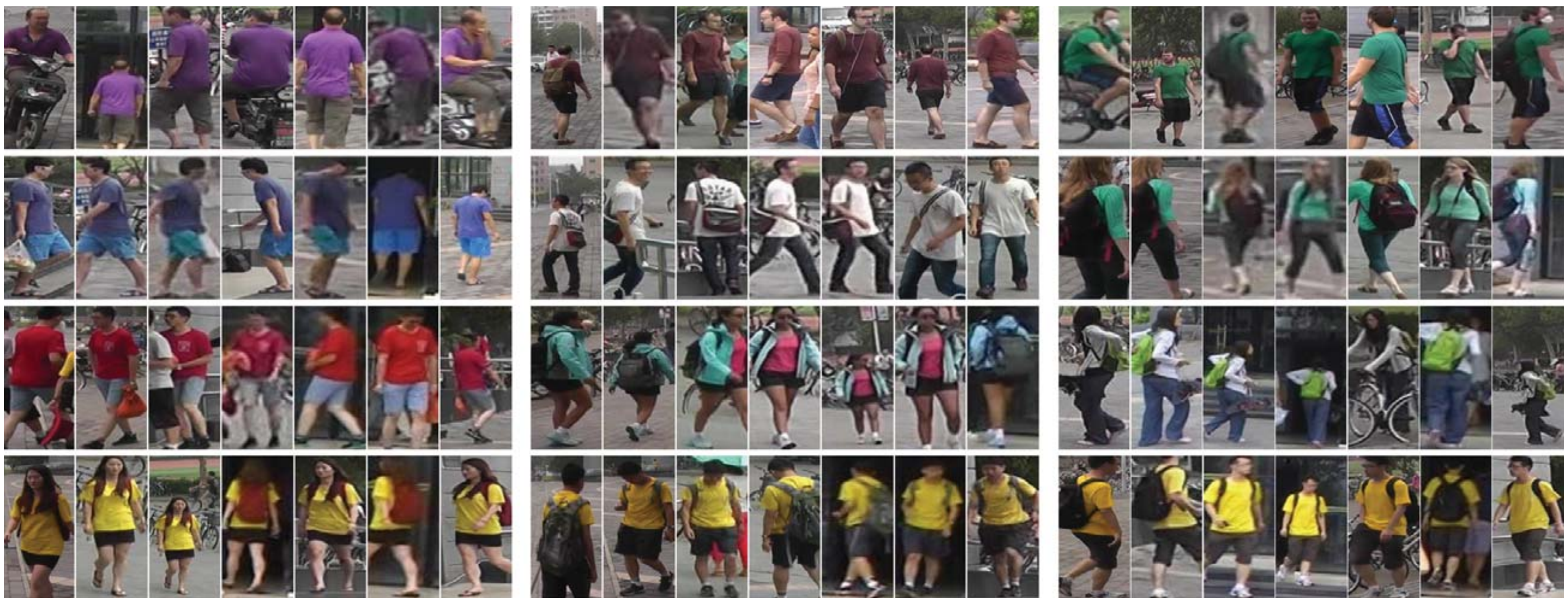
- The first three datasets are annotated from videos collected from Tsinghua University, China, in August 2014.

- We used 6 cameras
  - 5 HD (1920x1080) cameras, and 1 SD (720x576) camera
  - Moderate overlap exists among cameras

- The length of video is 10+ hours

1. Benchmark Solution – Market1501

- 1,501 identities; 32,668 bboxes by Deformable Part Model (DPM);
- 6 cameras; 3,368 queries ;



L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark", ICCV 2015.

1. Benchmark Solution – Market1501

Comparison with existing image re-id datasets

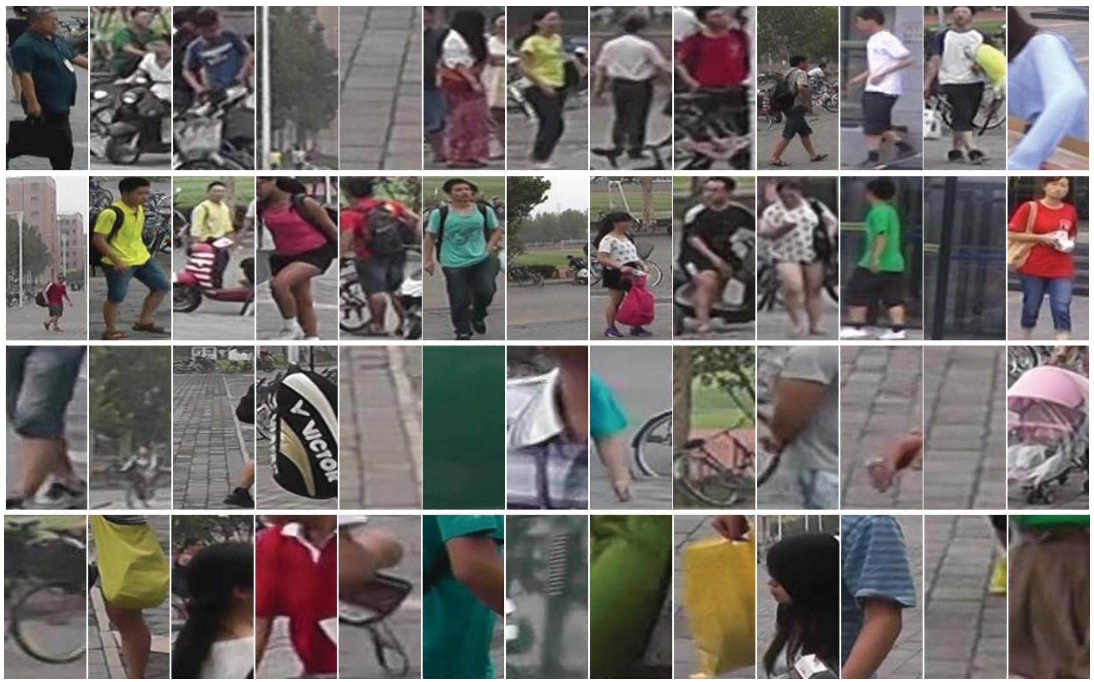
Datasets	Market-1501	RAiD	CUHK03	VIPeR	i-LIDS	OPeRID	CUHK01	CUHK02	CAVIAR
# Identities	1,501	43	1,360	632	119	200	971	1,816	72
# Bboxes	32,668	6,920	13,164	1,264	476	7,413	1,942	7,264	610
# Cam. per ID	6	4	2	2	2	5	2	2	2
DPM or Hand	DPM	Hand	DPM	Hand	Hand	Hand	Hand	Hand	Hand
Evaluation	mAP+CMC	CMC	CMC	CMC	CMC	CMC	CMC	CMC	CMC

L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark", ICCV 2015.

北京大学

1. Benchmark Solution – Market1501

- **We further add** **500k distractor images**: **Market-1501 + 500k dataset**



Distractors include:

- ✓ Incorrect detection results on the background
- ✓ Non-overlapping pedestrians with Market-1501

1. Market1501-Summary

- Summary
  - Large-scale benchmark
  - DPM detected bounding boxes
  - Multiple queries, multiple ground truths

- Limitations
  - Does not use the rich information in videos → MARS dataset [ECCV 2016]
  - No evaluation of pedestrian detectors → PRW dataset [CVPR 2017]
  - Short-time, single scene, bad detector → MSMT17 dataset [CVPR 2018]

1. Benchmark Solution – MARS

- 1,261 identities; Over 20k tracklets by DPM detector & GMMCP tracker (CVPR'15); Over 1 million frames
- 6 cameras; 2,009 queries ;

1. Benchmark Solution – MARS

- Testing Procedure
  - Given a query tracklet, we aim to search for tracklets containing the same person from other cameras.
  - mAP & CMC curve are used for evaluation

1. Benchmark Solution – MARS

Comparison with existing video re-id datasets

Datasets	MARS	**iLIDS-VID**	**PRID**	**3DPES**	**ETH**
# identities	**1,261**	300	200	200	146
# tracklets	**20,715**	600	400	1,000	146
# BBoxes	**1,067,516**	43,800	40,000	200k	8,580
# distractors	**3,248**	0	0	0	0
# cam. Per ID	**6**	2	2	8	1
Produced by	**DPM+GMMCP**	Hand	Hand	Hand	Hand
Evaluation	**mAP+CMC**	CMC	CMC	CMC	CMC

L. Zheng, Z. Bie, Y. Sun, C. Su, S. Wang, J. Wang, and Q. Tian, "MARS: A Video for Large-Scale Person Re-identification", ECCV 2016.

1. MARS - Summary

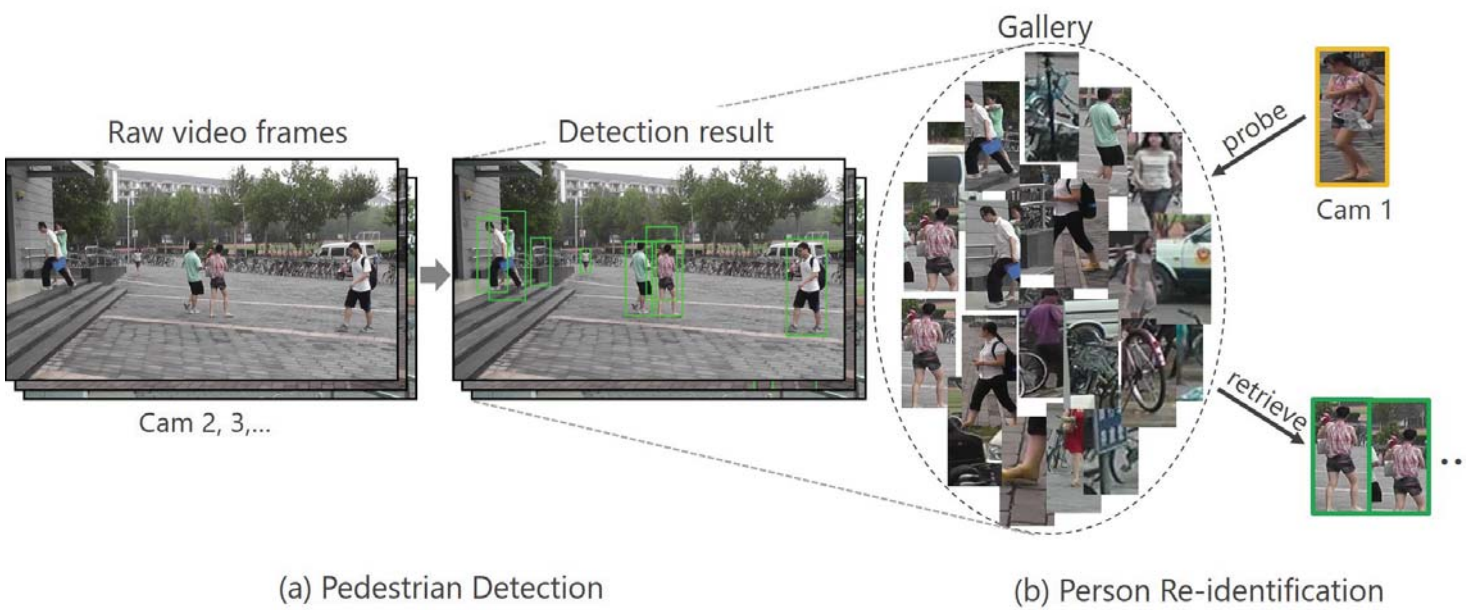
- Summary
  - Tracklets are used instead of single images
  - Tracklets contain more information
  - Large-scale

- Limitation
  - How does pedestrian detection error affect re-id accuracy?

→ PRW: [CVPR 2017]

1. Benchmark Solution – PRW

- PRW (Person Re-identification in the Wild)
  - We focus on both pedestrian detection and recognition



L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person Re-identification in the Wild", CVPR, 2017.

1. Benchmark Solution – PRW

Comparison with existing re-id datasets

Datasets	#frame	#ID	#anno. Box	#box/ID	#gallery box	#cam
PRW	11,816	932	34,304	36.8	100-500k	6
Market-1501	0	1,501	25,259	19.9	19,732	6
RAiD	0	43	6,920	160.9	6,920	4
VIPeR	0	632	1,264	2	1,264	2
i-LIDS	0	119	476	2	476	2
CUHK03	0	1,360	13,164	9.7	13,164	2

1. PRW - Summary

- Summary
  - Extensive benchmark of pedestrian detection and person re-identification
  - Tested on how detection aids re-identification

Existing Dataset vs. Real Ones

Datasets	*Duke*	*Market*	*CUHK03*	*CUHK01*	*VIPeR*	**Real World**
BBoxes	36,411	32,668	28,192	3,884	1,264	**1M +**
Identities	1,812	1,501	1,467	971	632	**10K +**
Cameras	8	6	2	10	2	**20 +**
Time Span	Short	Short	Short	Short	Short	**Long**
Scene	Outdoor	Outdoor	Indoor	Indoor	Outdoor	**Outdoor, Indoor**

- Existing public datasets differ from real data
  - Smaller scale
  - Fixed scenes
  - Shot term data, simple lighting condition

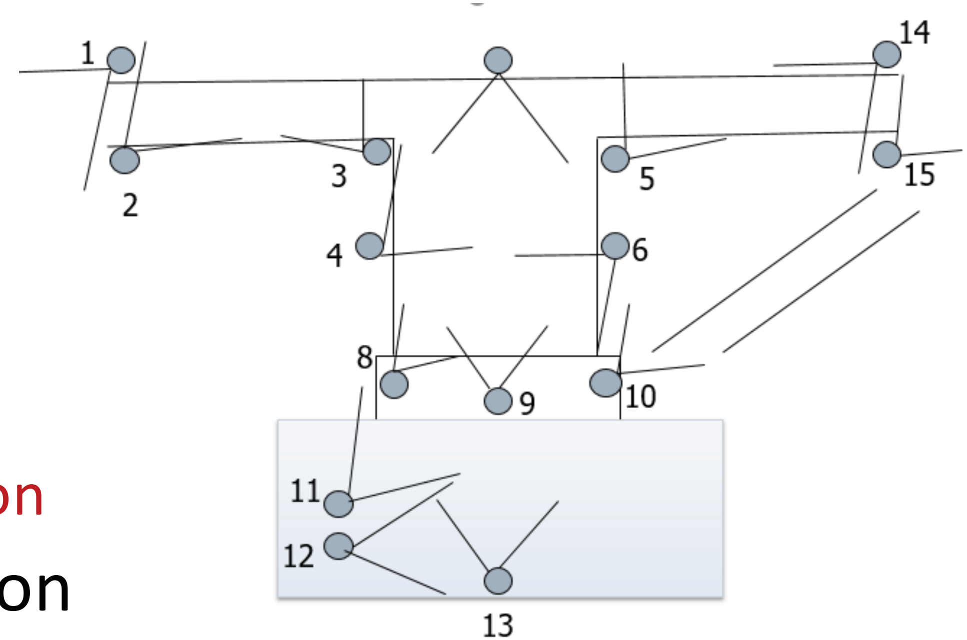
北京大学

How to push forward the research?

- We need a more realistic dataset
  - **Larger Scales**: more pedestrians, cameras, bboxes
  - **More Complex Scenes**: both indoor and outdoor
  - **Longer Time Spans**: complex lighting changes

1. Benchmark Solution – MSMT17

- 15 cameras
  - 12 outdoor, 3 indoor
- Totally 180 hours video
  - 4 days in one month
  - 3 hours each day: morning, noon, afternoon
- Faster RCNN for detection
- Annotation takes two months
  - 126,411 bounding boxes
  - 4,101 identities, 1041 for training 3060 for testing



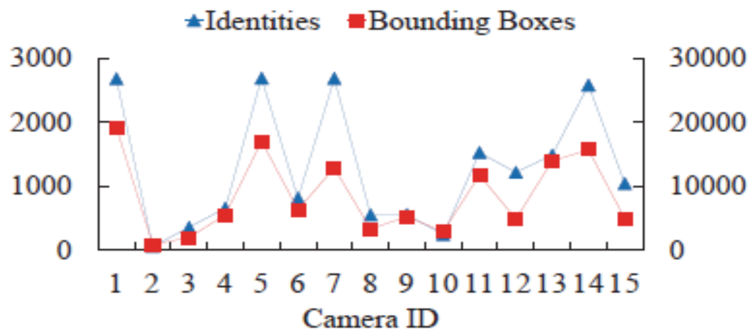
北京大学

1. MSMT17 - Comparison

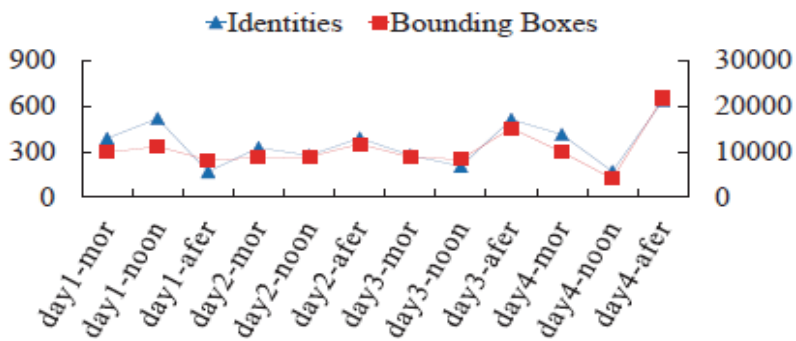
Datasets	MSMT17	Duke	Market	CUHK03	CUHK01	VIPeR	PRID
BBoxes	126,441	36,411	32,668	28,192	3,884	1,264	1,134
Identities	4,101	1,812	1,501	1,467	971	632	934
Cameras	15	8	6	2	10	2	2
Detector	Faster RCNN	Hand	DPM	DPM, Hand	Hand	Hand	Hand
Scene	Outdoor, Indoor	Outdoor	Outdoor	Indoor	Indoor	Outdoor	Outdoor
Time Span	1 month	short	short	short	short	short	short

- Largest size
- Complex scenes and backgrounds
- Multiple time slots
- State-of-the art auto detector

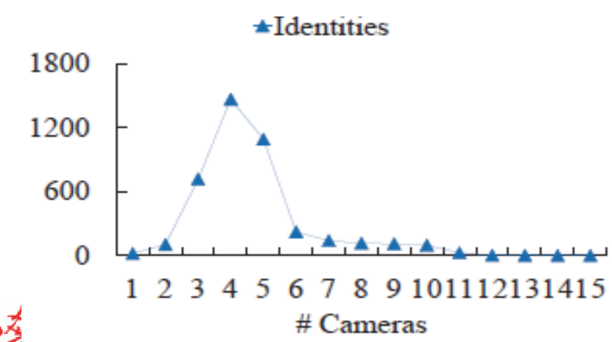
1. MSMT17 – More Statistics



Number of IDs and Bboxes on each camera



Number of IDs and Bboxes in each time slot



Number of IDs across different number of cameras

1. MSMT17 – Samples

1. MSMT17 – Performance

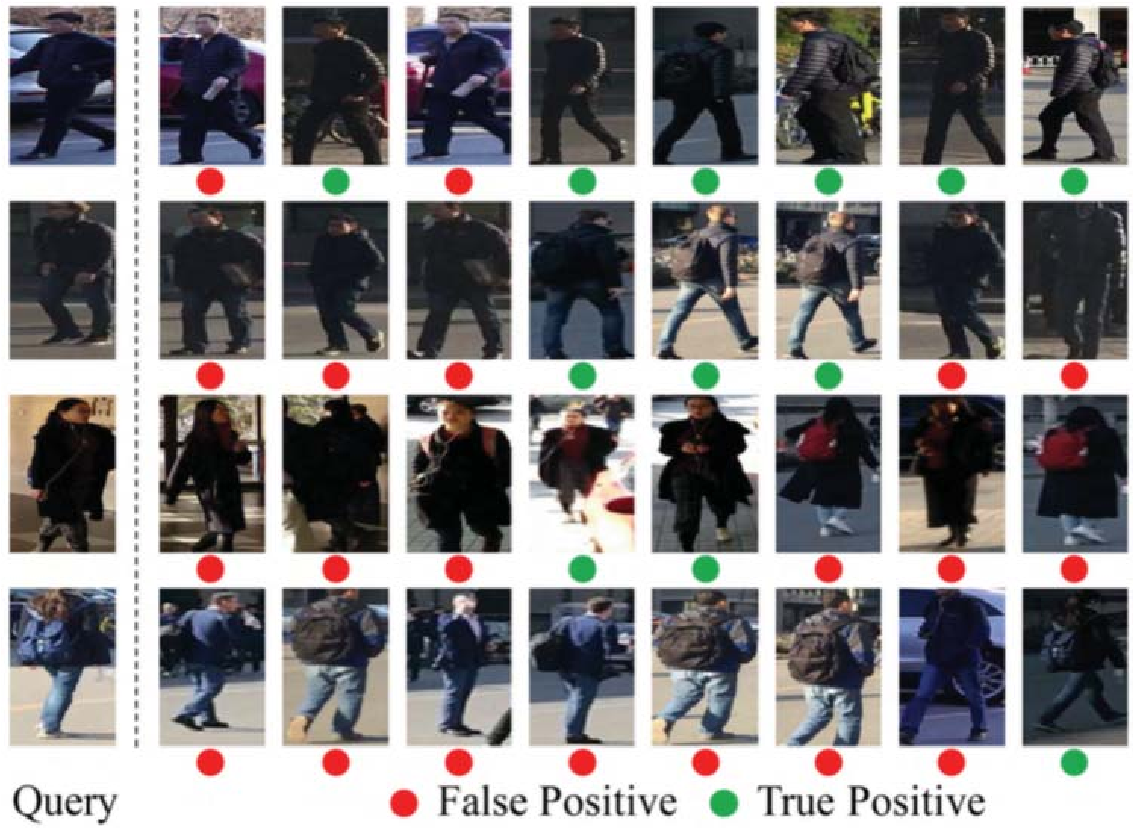
- Tested two of our recent works
  - PDC [ICCV'17], *R-1 88.7% on CUHK03*
  - GLAD [ACM MM'17], *R-1 89.9%, mAP 73.9% on Market*

Methods	R-1	R-5	R-10	R-20	mAP
GoogLeNet [1]	47.6	65.0	71.8	78.2	23.0
PDC [2]	58.0	73.6	79.4	84.5	29.7
GLAD [3]	61.4	76.8	81.6	85.9	34.0

[1] Szegedy, et al., "Going deeper with convolutions", In *CVPR*, 2015.
[2] Su, et al, "Pose- driven deep convolutional model for person re-identification", In *ICCV*, 2017.
[3] Wei, et al. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017.

北京大学

Performance on MSMT17



Sample retrieval results generated by the method of GLAD[1] on MSMT17.

[1] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval", In *ACM MM*, 2017.

1. MSMT17 - Summary

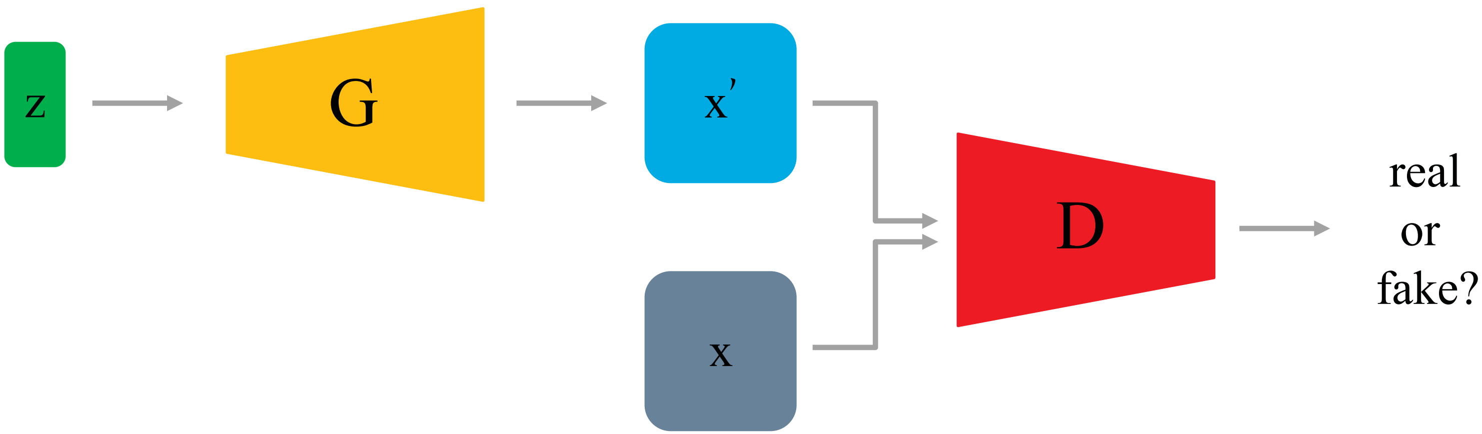
- Summary
  - Multi scenes, multi time
  - Largest, most challenging dataset
  - Faster RCNN detected boxes
  - A more realistic dataset you should try

Outline

- Our Benchmark Solution
- The Application of GANs in Person ReID
    - Overview of GANs
    - Our solution
- New Research Possibilities

2.1 Overview of GANs

- **Generative Adversarial Nets** (Goodfellow *et al.* NIPS 2014)
  - Minimax two-player game

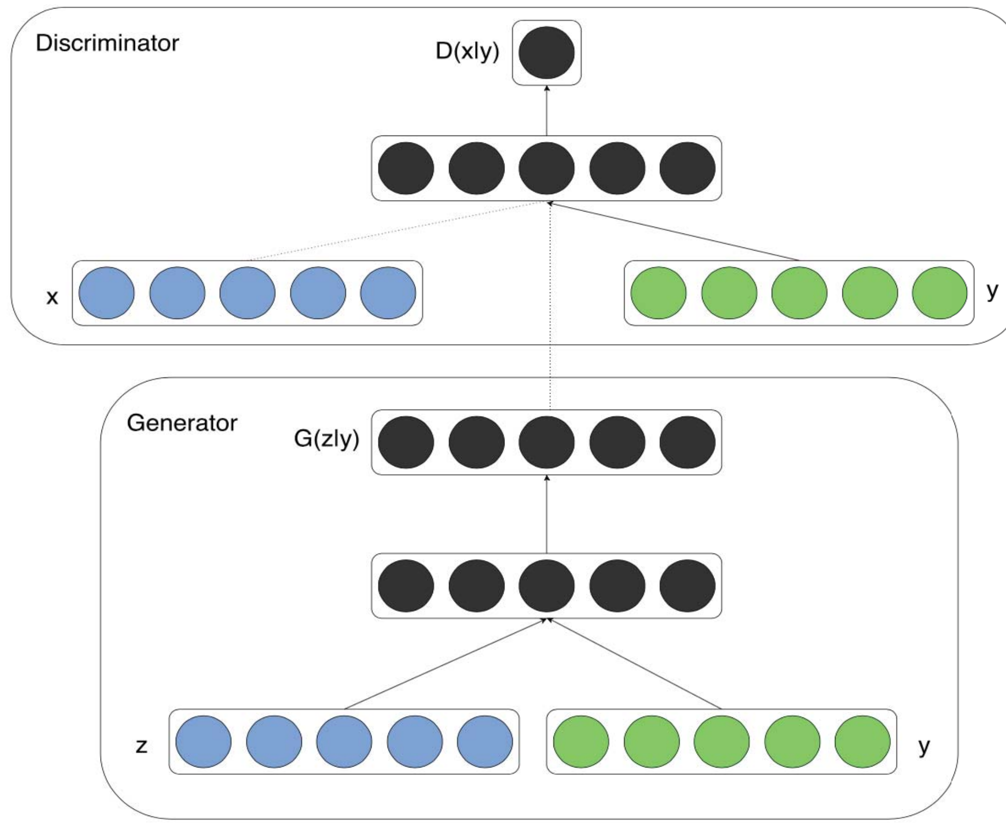


$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

2.1 Overview of GANs

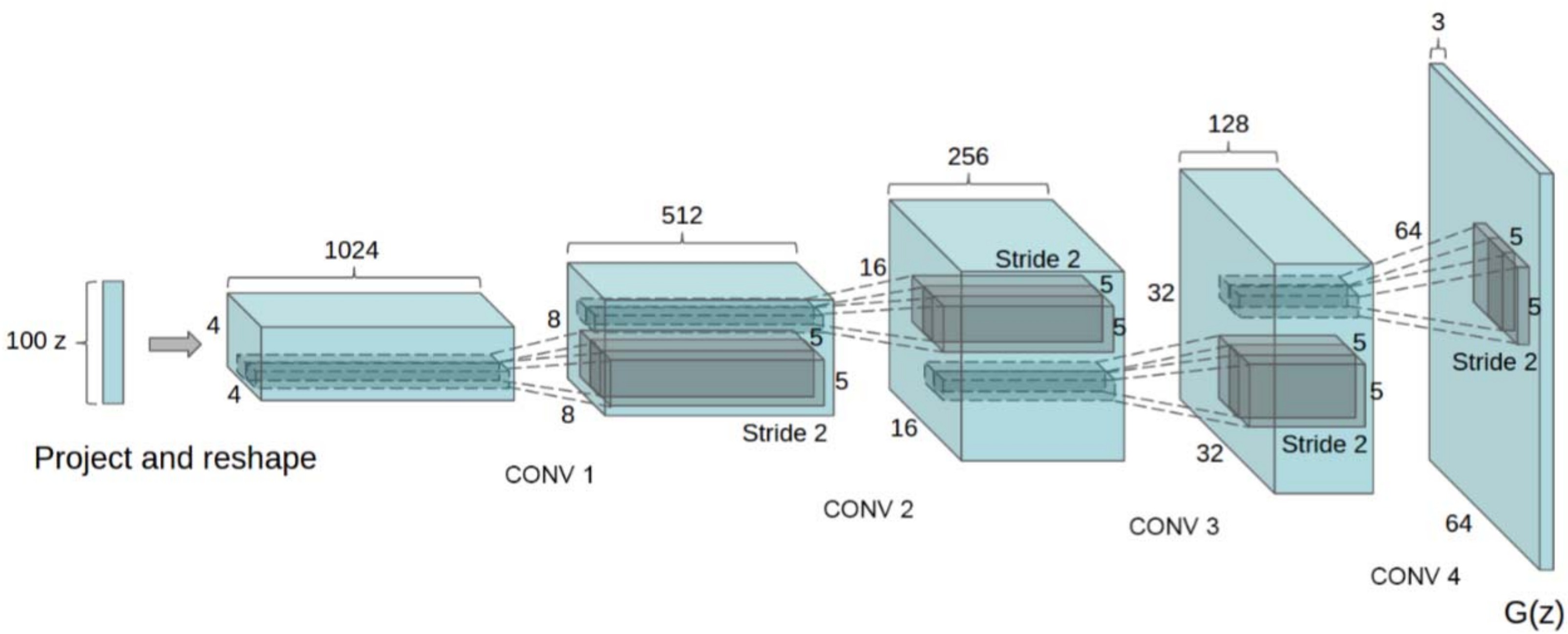
- **Conditional GANs** (Mirza *et al.* Arxiv 2014)
  - Feeding the conditional information to direct the data generation process

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))]$$



2.1 Overview of GANs

- **DCGANs** (Radford *et al.* ICLR 2016)
  - Propose certain constrains on the architecture topology of Convolutional GANs that make them stable to train
  - The first GAN model to learn to generate high resolution images in a single shot



2.1 Overview of GANs

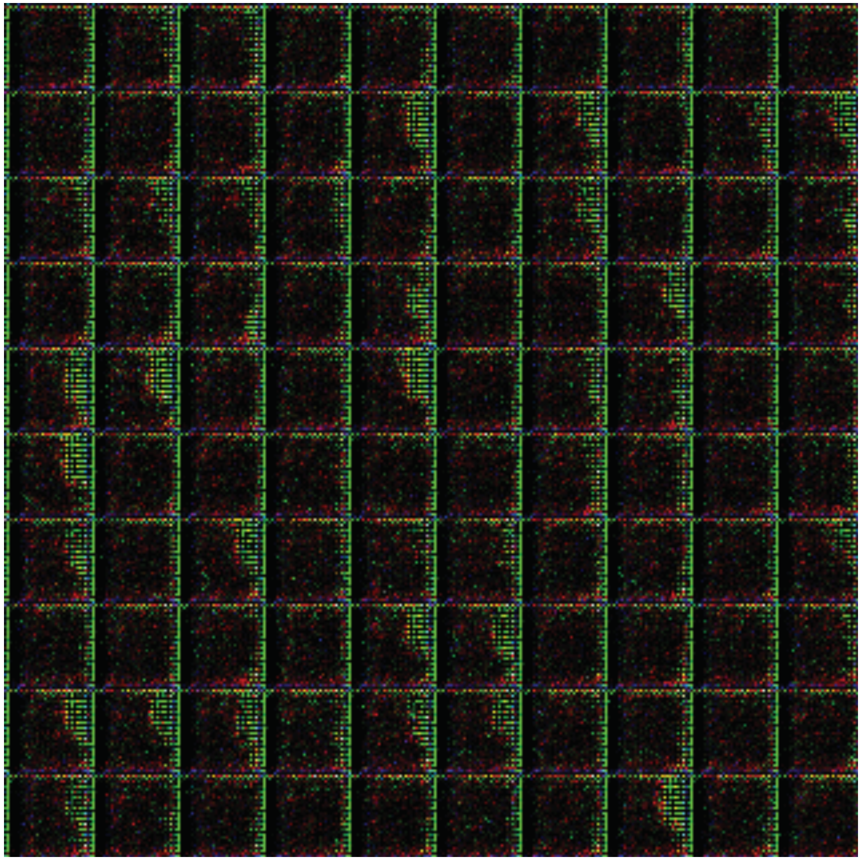
- **DCGANs** (Radford *et al.* ICLR 2016)
  - Propose certain constrains on the architecture topology of convolutional GANs that make them stable to train
  - The first GAN model to learn to generate high resolution images in a single shot

Architecture guidelines for stable Deep Convolutional GANs

- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
- Use batchnorm in both the generator and the discriminator.
- Remove fully connected hidden layers for deeper architectures.
- Use ReLU activation in generator for all layers except for the output, which uses Tanh.
- Use LeakyReLU activation in the discriminator for all layers.

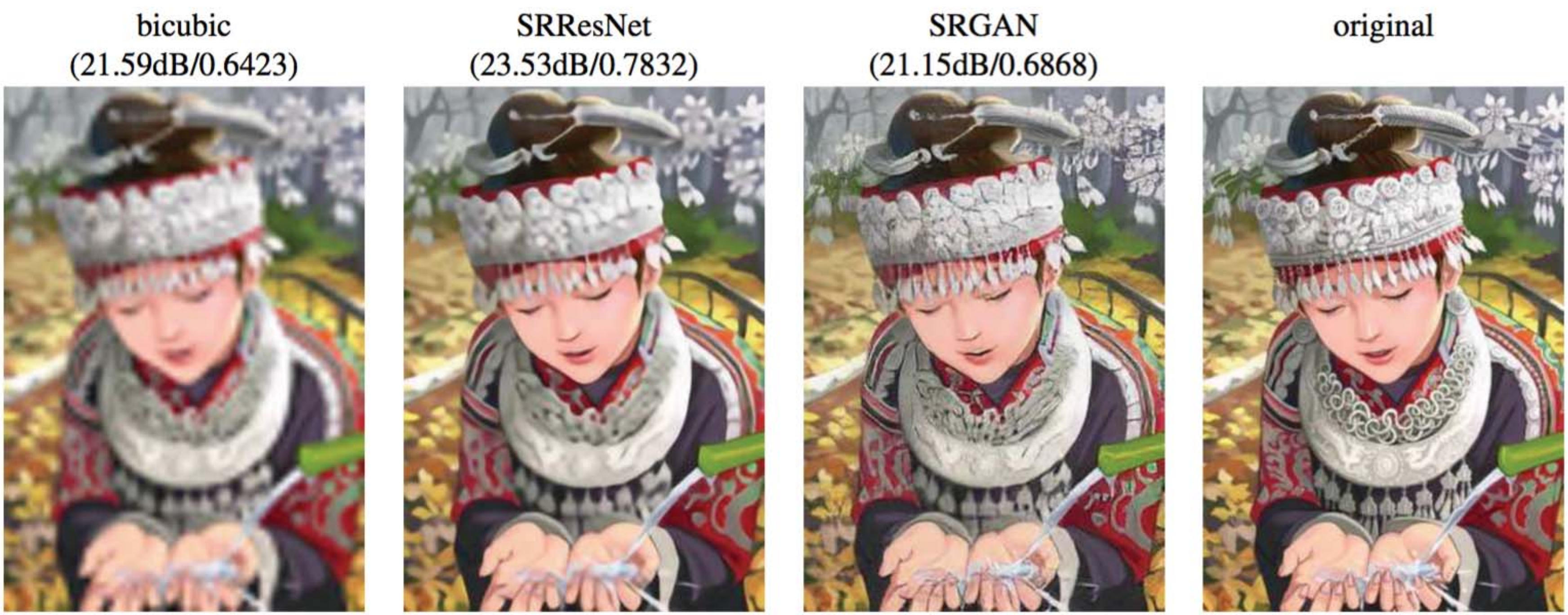
2.1 Overview of GANs

- **DCGANs** (Radford *et al.* ICLR 2016)



2.1 Overview of GANs

- **The Application of GANs in Computer Vision**
  - Image Super-Resolution[1]



[1] C. Ledig et al, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial”, in *CVPR*, 2017