ACM Multimedia 2020 Tutorial
Effective and Efficient: Toward Open-world Instance Re-identification

IS INSTANCE SEARCH A SOLVED PROBLEM?

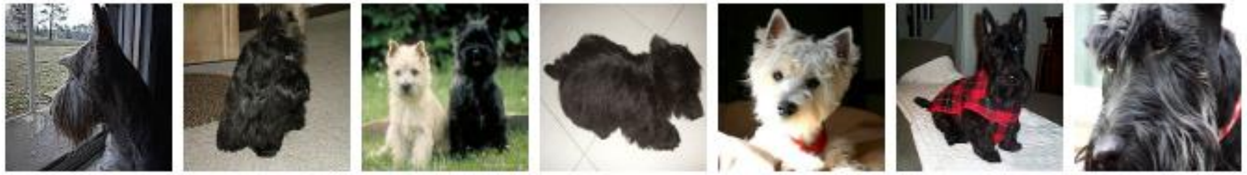
Shin’ichi Satoh

National Institute of Informatics, Japan

Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

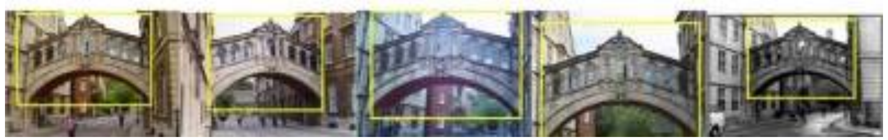
What is Instance Search?

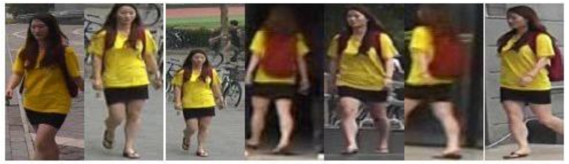
- Instance Search is different from image similarity search



NOT Instance Search

- Instance Search is to search for instances of specific objects, such as…
- Particular objects, individuals,





- Manufactured products whose appearance is indistinguishable,

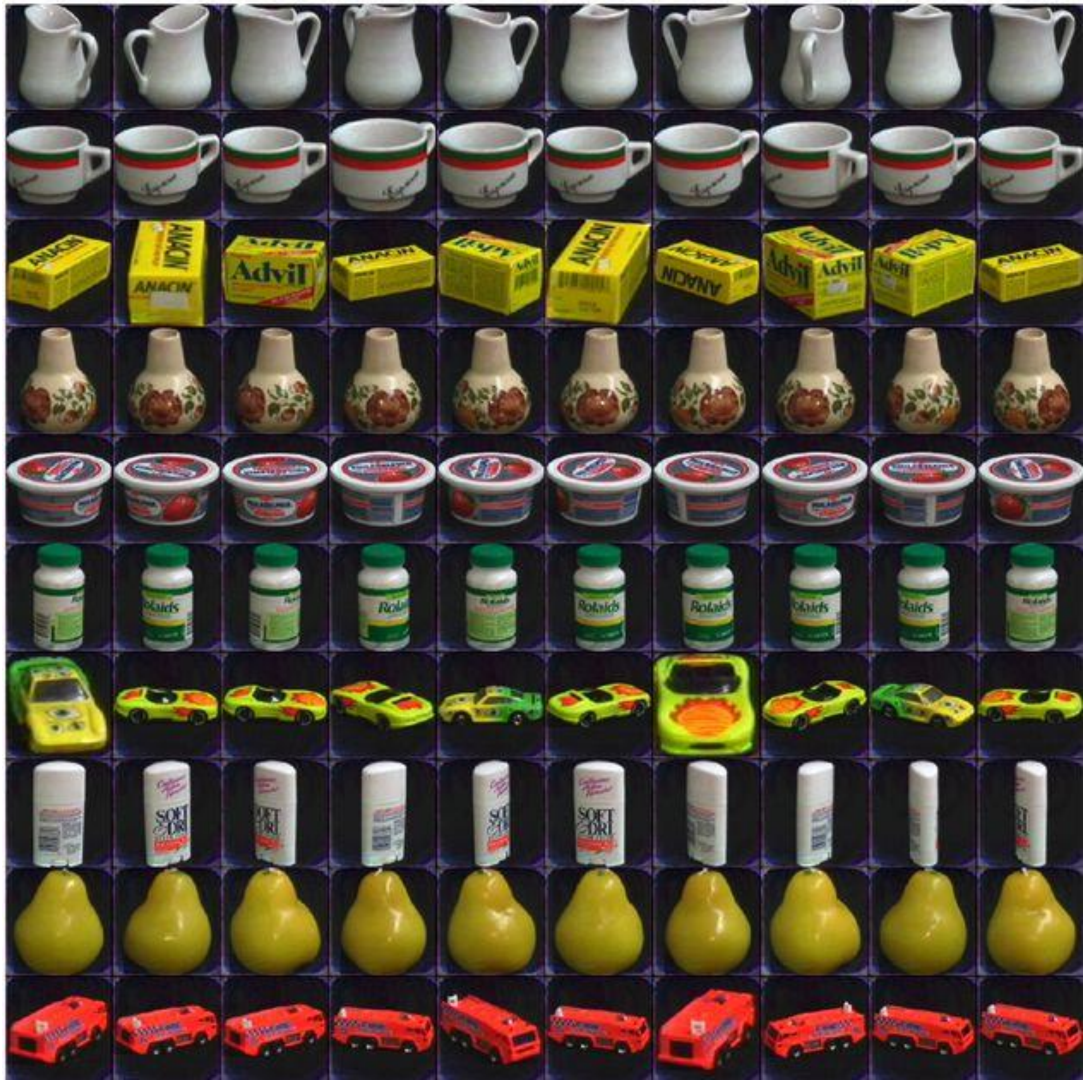
- Logos, etc.

Properties of Instance Search

- Ground truth is “well-defined”: if images contain the same object to the query, the images are relevant
- In contrast, image similarity is complicated to define
- Well studied from the early stage of image semantic analysis
- Many reasonable techniques work as expected: nice research topics
- Request from society: primary target of users of image retrieval, e.g., search for images of this person, search for images of this product, search for images of this landmark, etc.
- Good cue for multimedia analytics: market research of specific products by companies
- Surveillance, safety, security

COIL-100: Early Dataset for Object Recognition

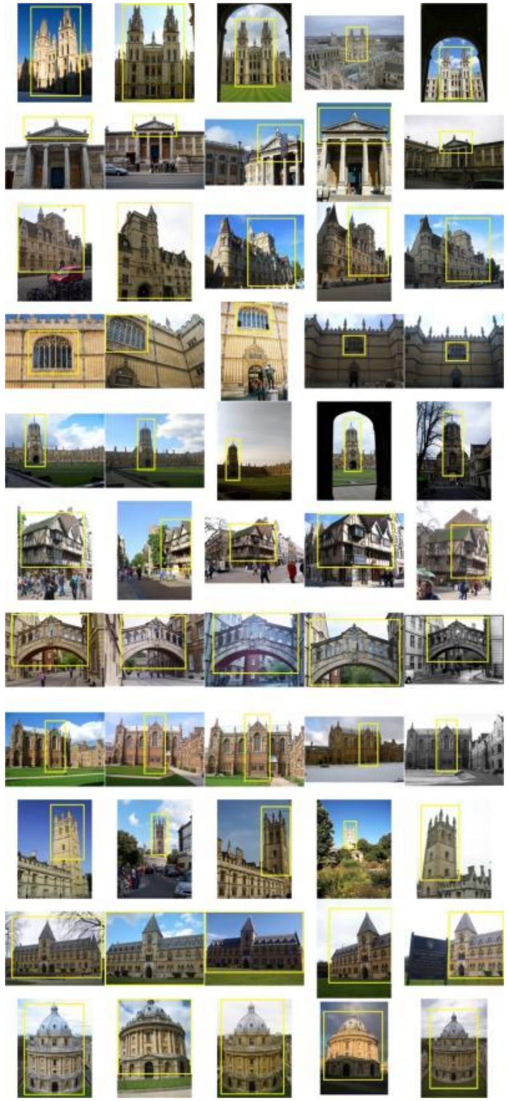
- Columbia University Image Library (COIL-100) appeared in 1990s for object recognition
- The goal is to identify objects, despite the view angle differences
- No background (simple)
- Before Caltech, PASCAL VOC, ImageNet
- “Easier” than these datasets
- One reason: very small intra-class variation



"Columbia Object Image Library (COIL-100),"
S. A. Nene, S. K. Nayar and H. Murase,
Technical Report CUCS-006-96, February 1996.

Landmark Retrieval: Oxford and Paris

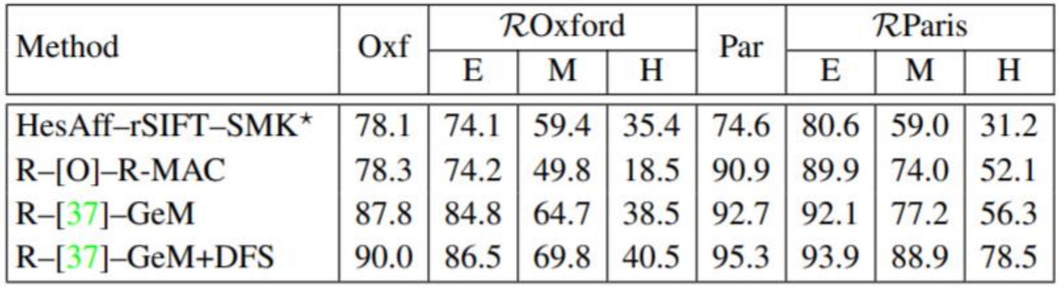
- Landmark retrieval is very well studied
- The definition of the ground truth is very clear: whether the target landmark appears in images or not (c.f., whether images are similar or not)
- 5k Oxford landmarks and 6k Paris landmarks
- Oxford (CVPR07): 3k references, Paris (CVPR08): 1.5k references
- Many important image retrieval techniques have been studied based on these datasets
- Query expansion, database-side augmentation, query-side augmentation, geometric verification, diffusion, etc…
- Recently very high performance reported (over 90% map): already solved?



Oxford: Object retrieval with large vocabularies and fast spatial matching, CVPR07
Paris: Lost in quantization: Improving particular object retrieval in large scale image databases, CVPR08

Revisiting Oxford and Paris (CVPR18)

- Cleansed annotations for Oxford and Paris with new labels: Easy, Hard, Unclear, Negative
- New, larger, cleaner distractors
- There still remain many challenging situations: "*image retrieval appears far from being solved*"



Method	Oxf	$\mathcal{R}$Oxford			Par	$\mathcal{R}$Paris		
		E	M	H		E	M	H
HesAff–rSIFT–SMK*	78.1	74.1	59.4	35.4	74.6	80.6	59.0	31.2
R–[O]–R-MAC	78.3	74.2	49.8	18.5	90.9	89.9	74.0	52.1
R–[37]–GeM	87.8	84.8	64.7	38.5	92.7	92.1	77.2	56.3
R–[37]–GeM+DFS	90.0	86.5	69.8	40.5	95.3	93.9	88.9	78.5

Table 3. Performance (mAP) on Oxford (Oxf) and Paris (Par) with the original annotation, and $\mathcal{R}$Oxford and $\mathcal{R}$Paris with the newly proposed annotation with three different protocol setups: Easy (E), Medium (M), Hard (H).

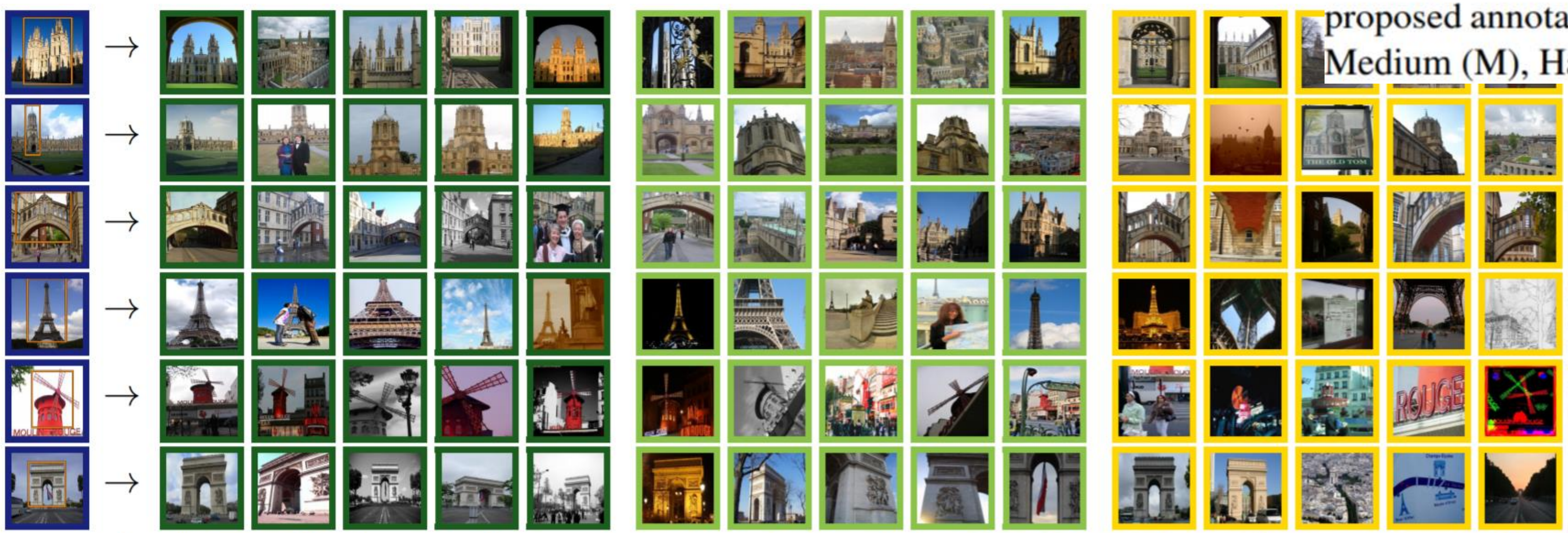
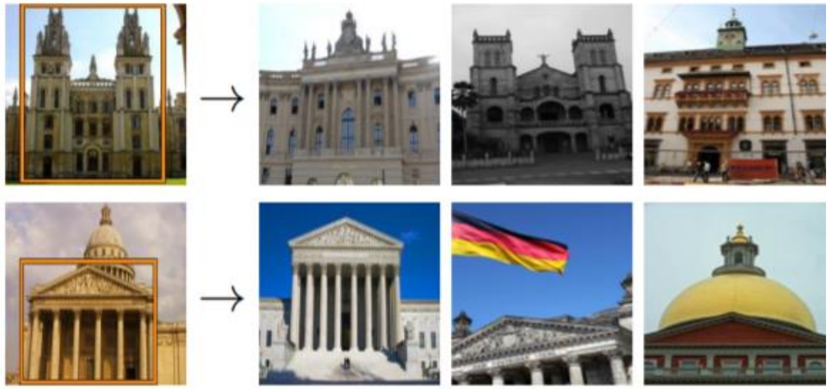


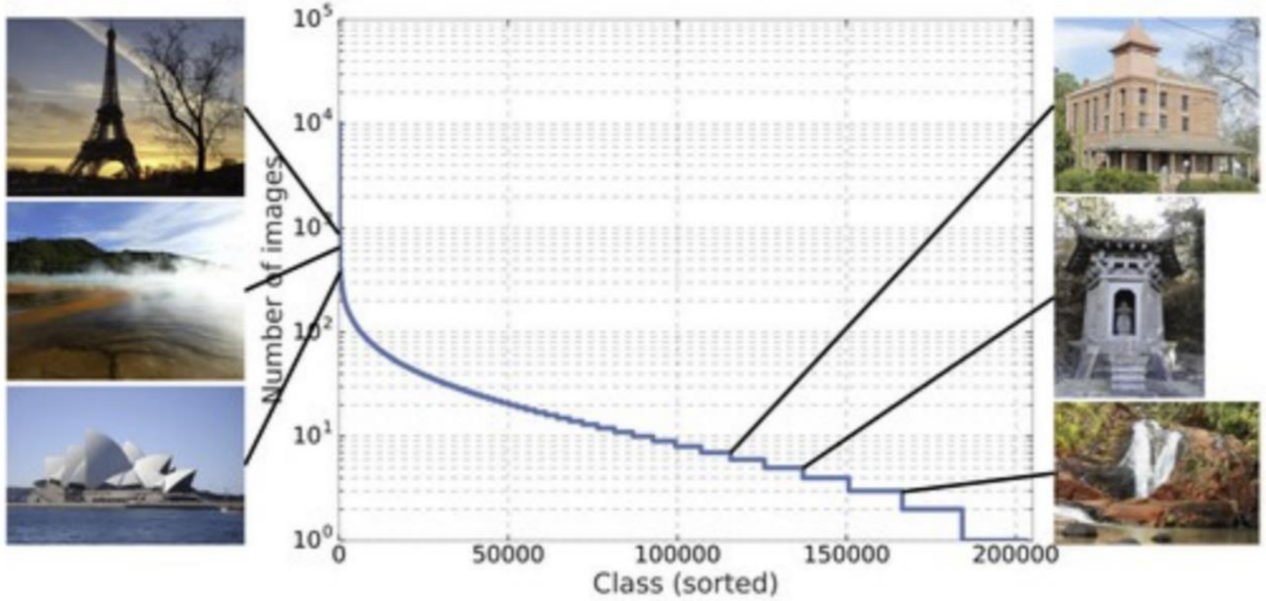
Figure 3. Sample **query (blue)** images and images that are respectively marked as **easy (dark green)**, **hard (light green)**, and **unclear (yellow)**. Best viewed in color.



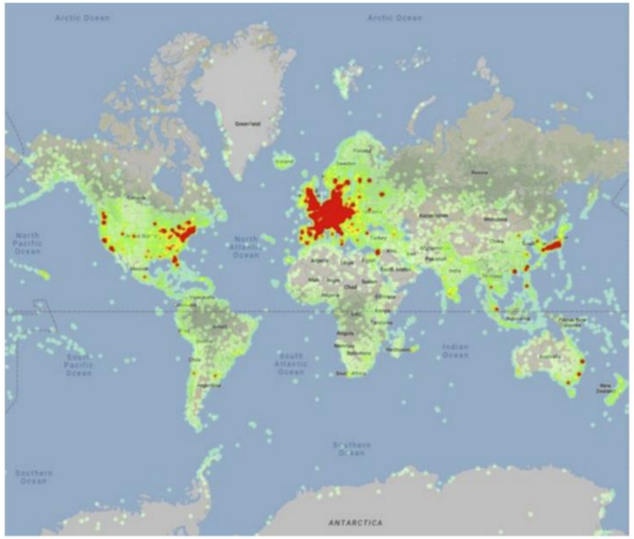
Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking, CVPR18

Google Landmark V2 (CVPR20)

- Large-scale, long-tailed landmark retrieval dataset
- “ImageNet” of landmark retrieval
- 5M images, 200k instance labels



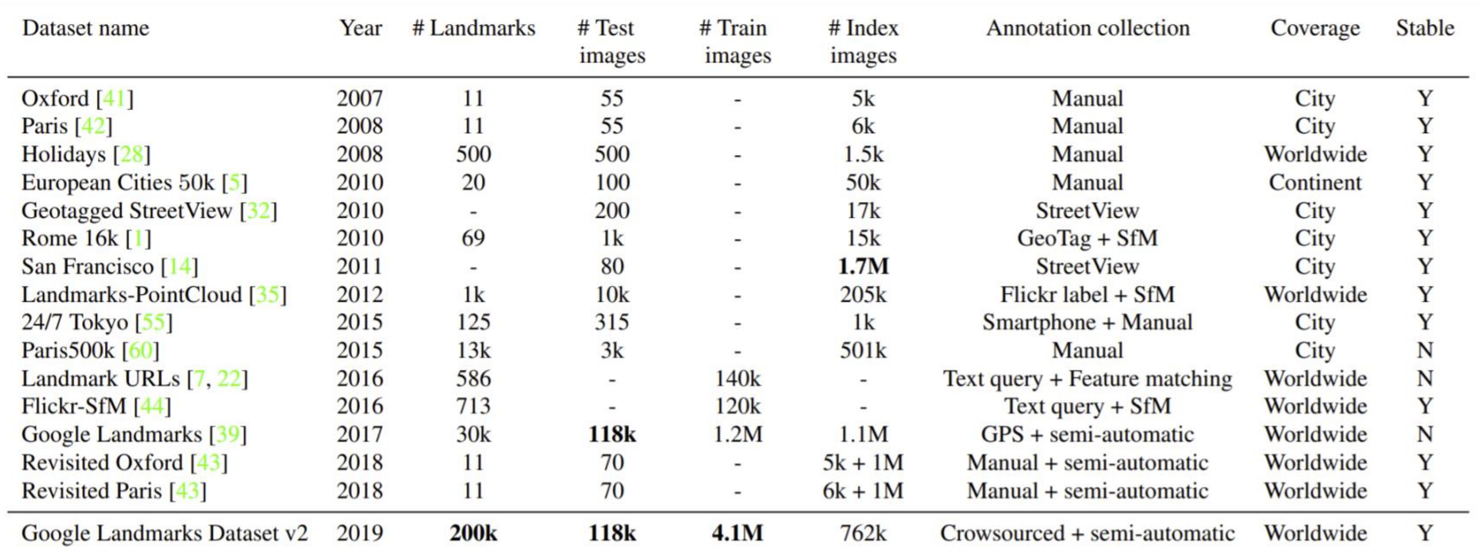
the class distribution is very long-tailed, the dataset contains a large number of lesser-known local landmarks



Heatmap of the places in the Google Landmarks Dataset v2

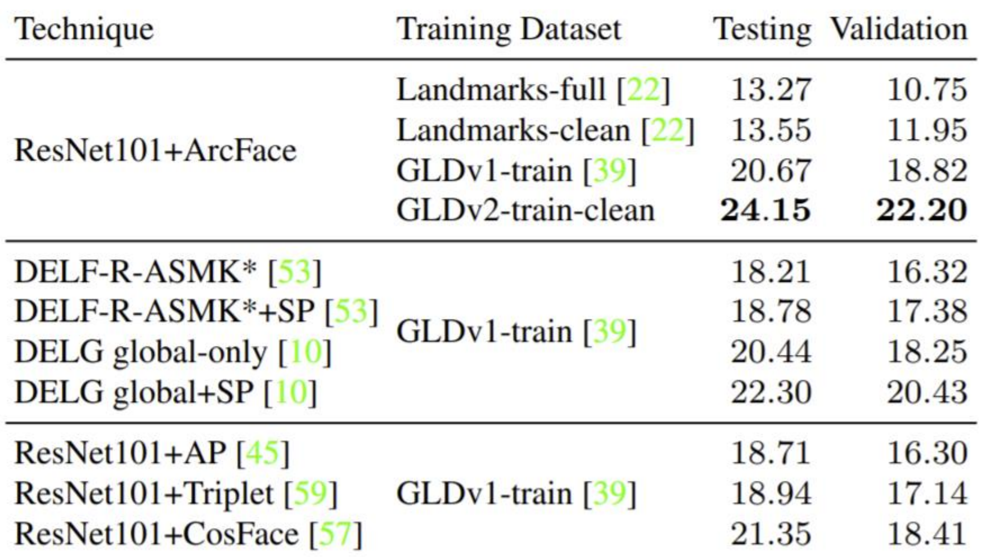
Google Landmarks Dataset v2 A Large-Scale Benchmark for Instance-Level Recognition and Retrieval, CVPR20

Google Landmark V2 (CVPR20)



Dataset name	Year	# Landmarks	# Test images	# Train images	# Index images	Annotation collection	Coverage	Stable
Oxford [41]	2007	11	55	-	5k	Manual	City	Y
Paris [42]	2008	11	55	-	6k	Manual	City	Y
Holidays [28]	2008	500	500	-	1.5k	Manual	Worldwide	Y
European Cities 50k [5]	2010	20	100	-	50k	Manual	Continent	Y
Geotagged StreetView [32]	2010	-	200	-	17k	StreetView	City	Y
Rome 16k [1]	2010	69	1k	-	15k	GeoTag + SfM	City	Y
San Francisco [14]	2011	-	80	-	**1.7M**	StreetView	City	Y
Landmarks-PointCloud [35]	2012	1k	10k	-	205k	Flickr label + SfM	Worldwide	Y
24/7 Tokyo [55]	2015	125	315	-	1k	Smartphone + Manual	City	Y
Paris500k [60]	2015	13k	3k	-	501k	Manual	City	N
Landmark URLs [7, 22]	2016	586	-	140k	-	Text query + Feature matching	Worldwide	N
Flickr-SfM [44]	2016	713	-	120k	-	Text query + SfM	Worldwide	Y
Google Landmarks [39]	2017	30k	**118k**	1.2M	1.1M	GPS + semi-automatic	Worldwide	N
Revisited Oxford [43]	2018	11	70	-	5k + 1M	Manual + semi-automatic	Worldwide	Y
Revisited Paris [43]	2018	11	70	-	6k + 1M	Manual + semi-automatic	Worldwide	Y
Google Landmarks Dataset v2	2019	**200k**	**118k**	**4.1M**	762k	Crowsourced + semi-automatic	Worldwide	Y

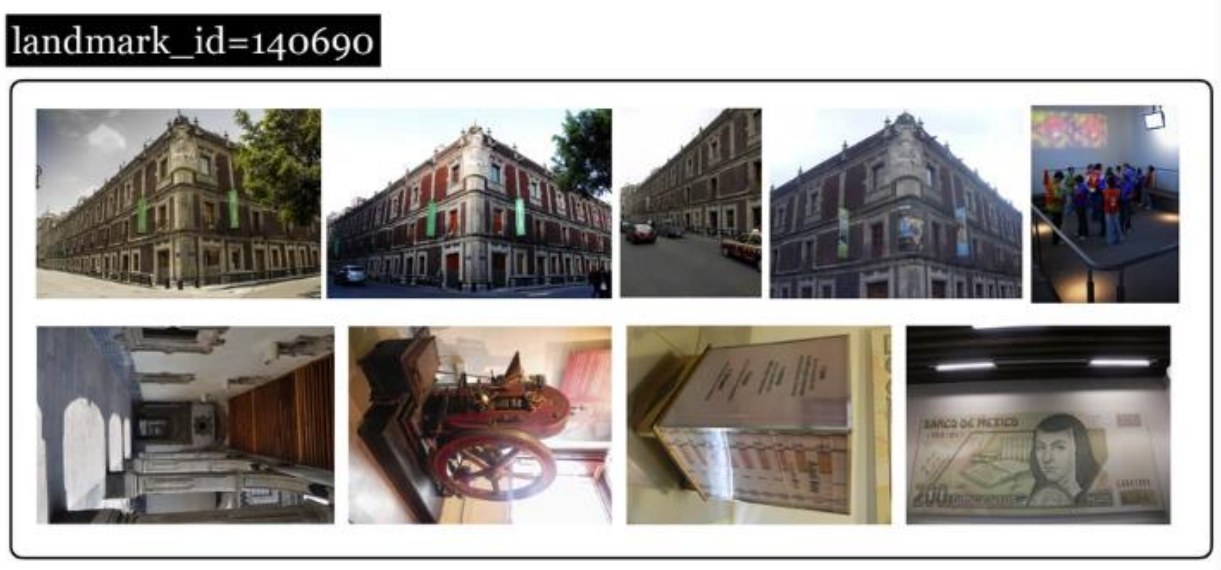
Comparison against existing landmark datasets

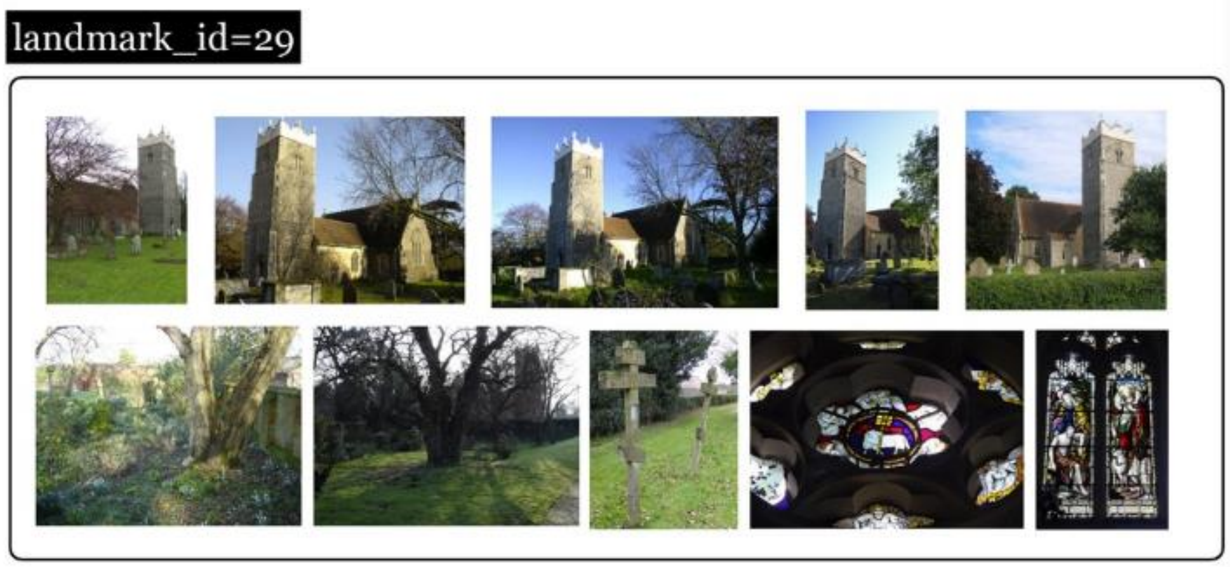


Technique	Training Dataset	Testing	Validation
ResNet101+ArcFace	Landmarks-full [22]	13.27	10.75
	Landmarks-clean [22]	13.55	11.95
	GLDv1-train [39]	20.67	18.82
	GLDv2-train-clean	**24.15**	**22.20**
DELF-R-ASMK* [53]	GLDv1-train [39]	18.21	16.32
DELF-R-ASMK*+SP [53]		18.78	17.38
DELG global-only [10]		20.44	18.25
DELG global+SP [10]		22.30	20.43
ResNet101+AP [45]	GLDv1-train [39]	18.71	16.30
ResNet101+Triplet [59]		18.94	17.14
ResNet101+CosFace [57]		21.35	18.41

Baseline results (% mAP@100) for the Google Landmarks Dataset v2 (GLDv2) retrieval task.

Google Landmark V2: Noise problem

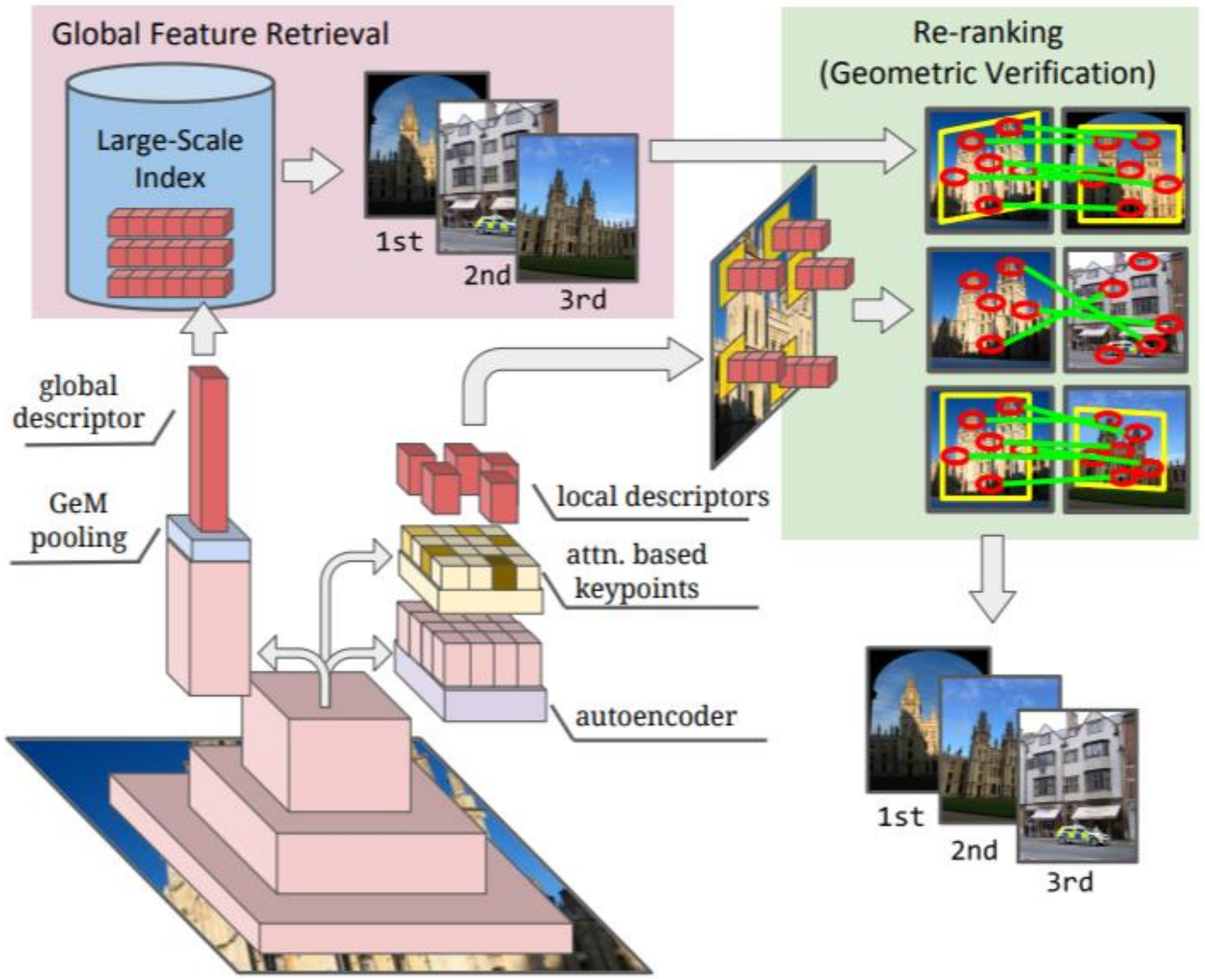




[Large-scale Landmark Retrieval/Recognition under a Noisy and Diverse Dataset](https://arxiv.org/abs/1906.04087)
https://arxiv.org/abs/1906.04087

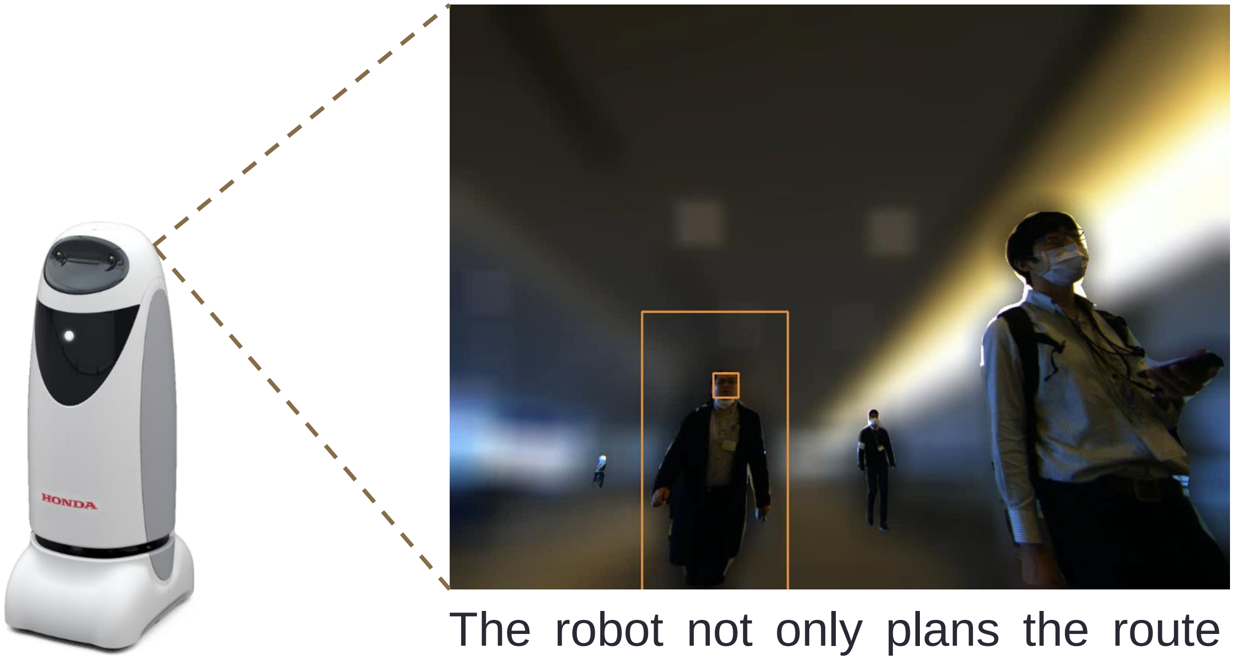
Approaches

- Deep local feature-based methods are very high-performant in recent landmark retrieval attempts
- MAC, CroW, R-MAC, DELF, DIR…
- “Old” encoding techniques for local features such as VLAD, SMK, ASMK (originally proposed for SIFT) are still effective
- Spatial verification, query expansion, diffusion, are also effective
- SIFT has been replaced by deep, but overall pipeline is kept: new breakthrough?



DELF (Large-Scale Image Retrieval with Attentive Deep Local Features, ICCV17)

New Challenge in Person Re-ID: Guiding Robot w/ Follower



The robot not only plans the route and pilots the way (using the front view camera), but also always **focuses the follower (using the back view camera)**, so that the follower could catch up the robot and keep in service.

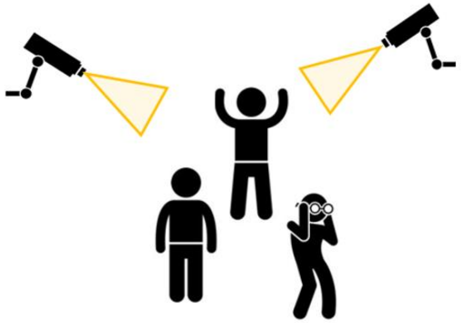
Face Recognition

⇩

Person Re-identification

ACMMM20 Demo
Progressive Domain Adaptation for
Robot Vision Person Re-identification

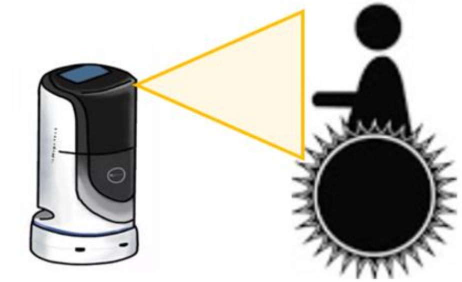
Security Vision *vs.* Robot Vision

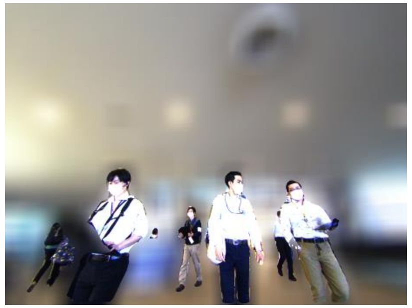


Camera is fixed. Background of camera view maintains. A stable illumination condition.
The view of camera is from top to bottom.

Persons have relatively stable appearance.

Sufficient annotations



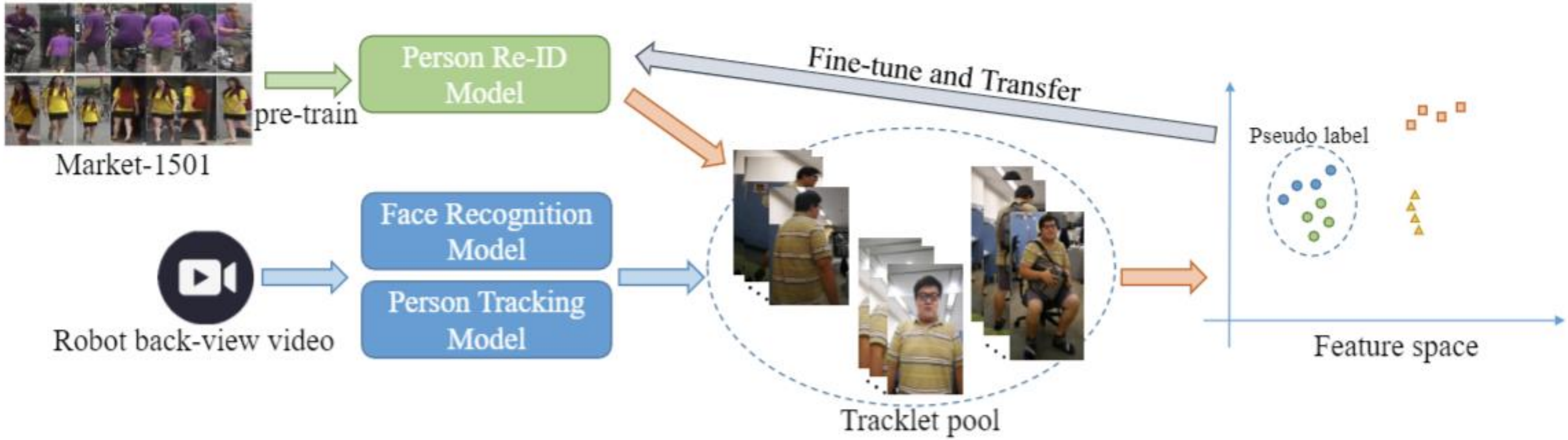


The camera view changes. Person images encounter shaking and changing illuminations.
The view of camera is from bottom to top.

Persons have more variations, including clothes change, frequently occlusion, half of the body.

No annotation

Framework

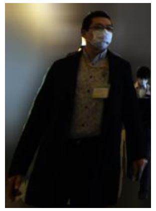


- **Step 1:** Re-ID models are pre-trained on a camera network dataset (Market-1501).
- **Step 2:** The robot back-view video is input into the existing face recognition and person tracking models, then a batch of tracklets will be generated.
- **Step 3:** In the feature space, the similarities of images from different tracklets are calculated, and thus pseudo labels are generated.
- **Step 4:** The selected tracklets (some batches of images) and their pseudo labels are used to fine-tune the Re-ID model.

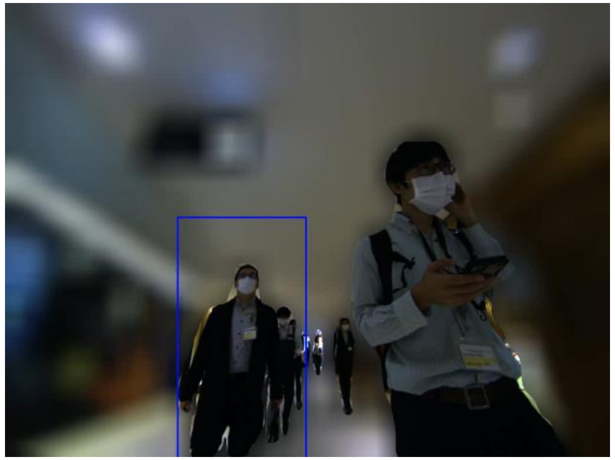
Example Videos

Face Recognition

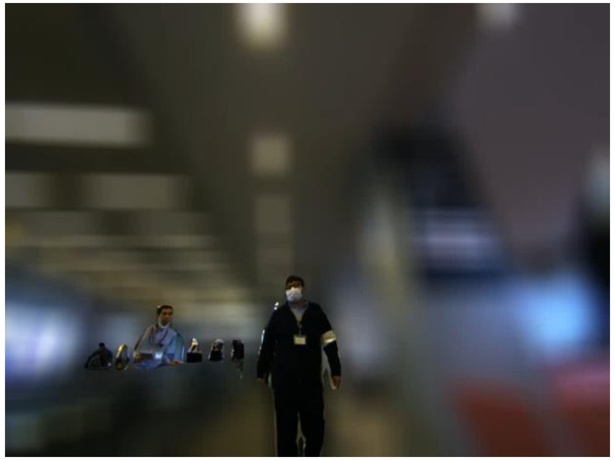
Person Re-identification

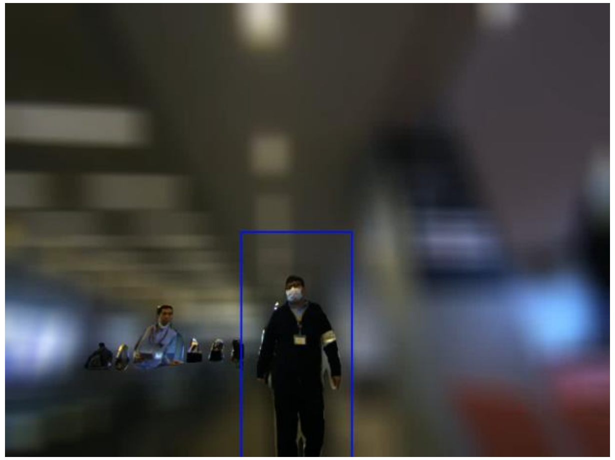


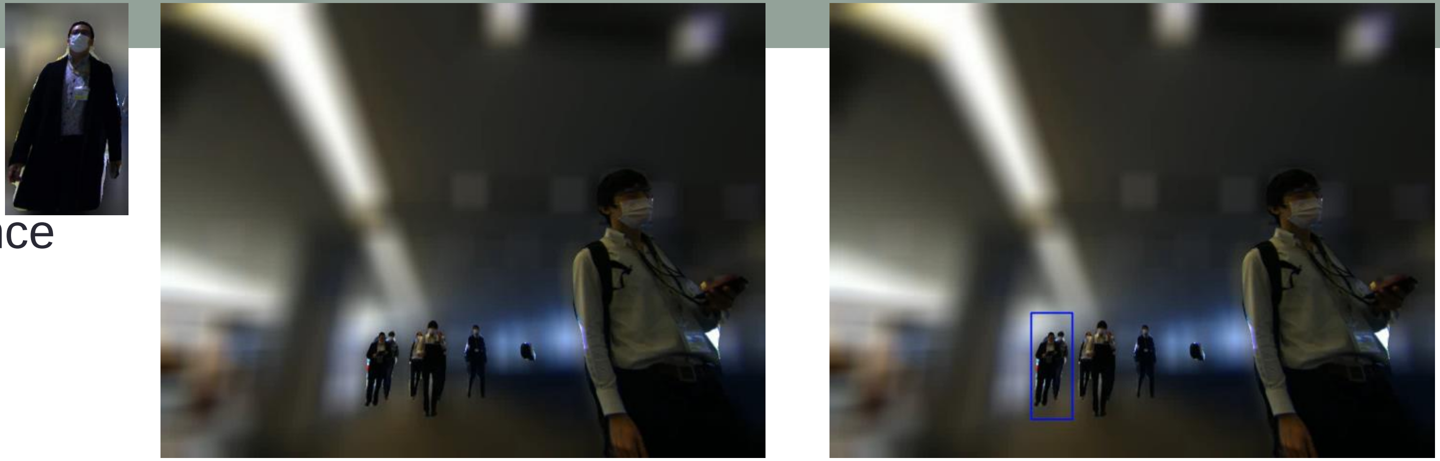
W/ Mask



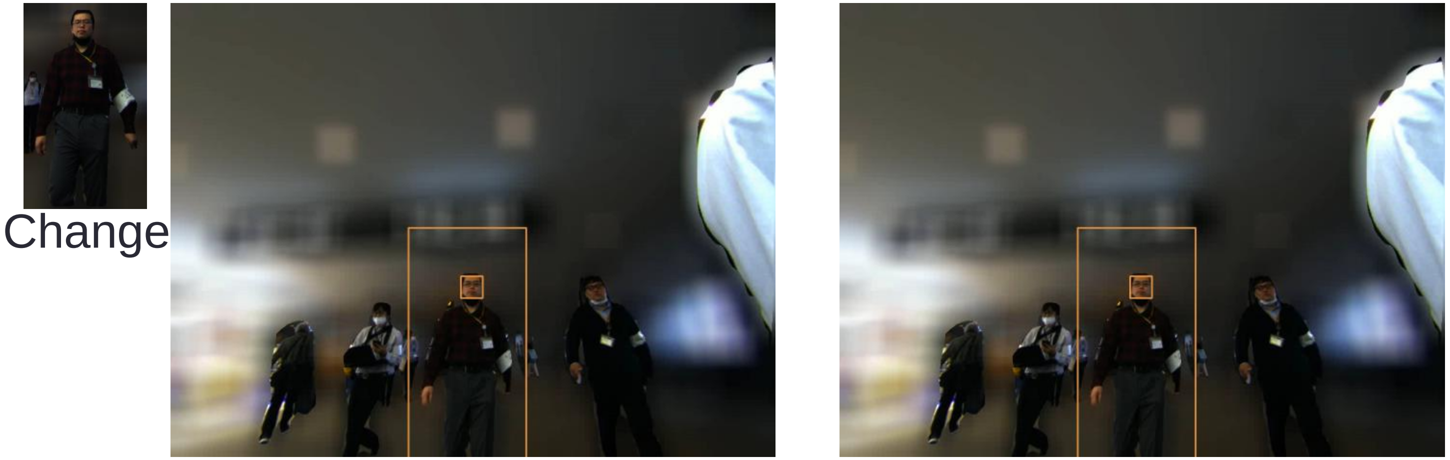
Different Poses



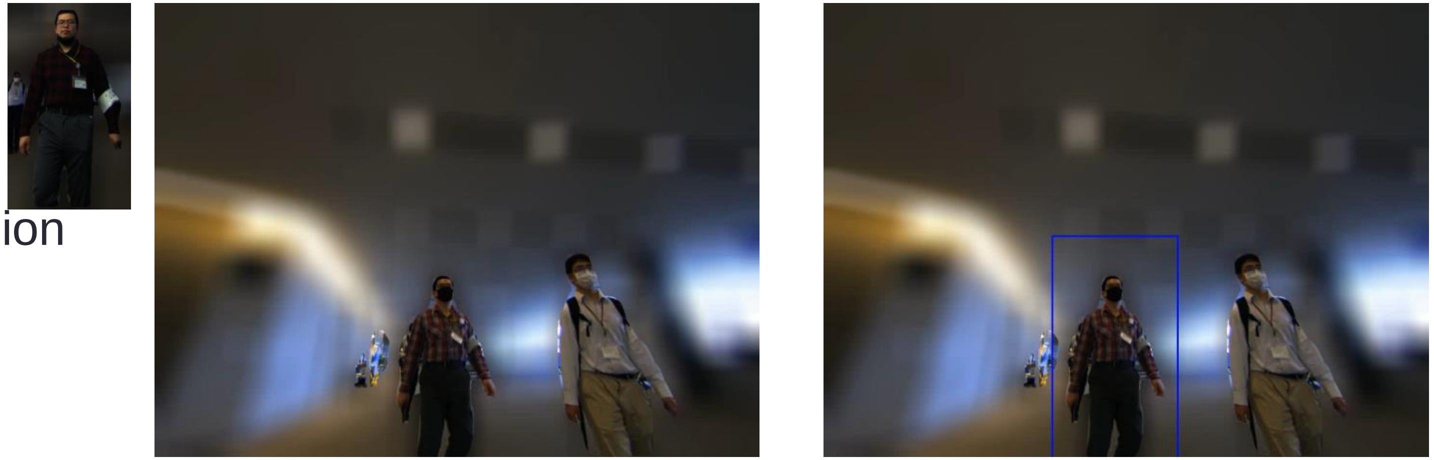




Far Distance



Illumination Change



Occlusion

Conclusion

- Instance search is an old, well-defined, and well studied task
- But it’s still important and there still are many challenges remaining
- Landmark retrieval faces now new and practical issues: large-scale, long-tail, noise issues
- Approaches: deep local feature is used in place of SIFT, but overall pipeline: local feature encoding, post processing, reranking, diffusion…
- New breakthrough is awaited
- Person Re-Identification: viewpoint changes, severe illumination changes, pose variations