CollabAR: Edge-assisted Collaborative Image Recognition for Mobile Augmented Reality

Zida Liu[§], Guohao Lan[§], Jovan Stojkovic [§] [*], Yunfan Zhang[§], Carlee Joe-Wong[†], Maria Gorlatova[§]

[§] Duke University, Durham, NC, [†]Carnegie Mellon University, Moffett Field, CA
[*]University of Belgrade, Belgrade, Serbia

Outline

- **Introduction**
- CollabAR system design
- MVMDD dataset
- Evaluation

Augmented Reality (AR)

- The [virtual] content is laid out around a user in the same spatial coordinates.

- Image recognition enables seamless contextual AR experience.

Challenges

Large proportion of distorted images in real-world mobile AR scenarios.

Magic Leap One

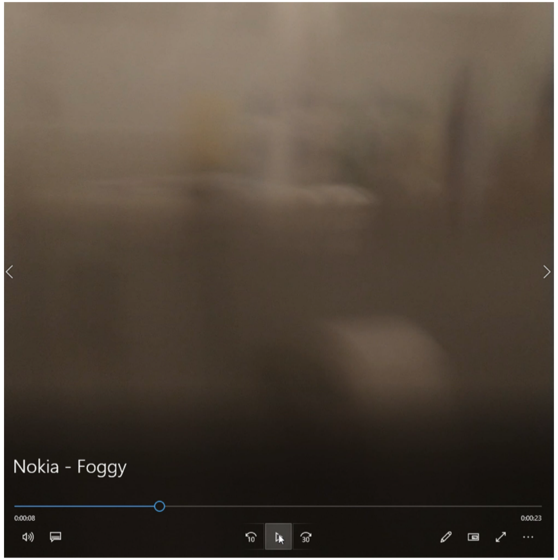
Nokia 7.1

Distortion	Setting	Hardware	
		Nokia 7.1	Magic Leap One
Gaussian noise ($\sigma_{GN}^2 \geqslant 0.003$)	Dark room (7lux)	617/1558=39.6%	infeasible
	Camera zoom-in (509lux)	1755/2185=97.2%	infeasible
Motion blur ($l \geqslant 5$)	Corridor (178lux)	2681/3452=77.6%	3760/3776=99.6%
	Sunny outdoor (9873lux)	16/2687=0.5%	957/2766=34.6%
Gaussian blur ($k \geqslant 5$)	Foggy	762/935=81.6%	infeasible
	Underwater	1250/1524=82.0%	infeasible

Measurements study in real-world mobile AR scenarios

Challenges

Large proportion of distorted images in real-world mobile AR scenarios.



Magic Leap One

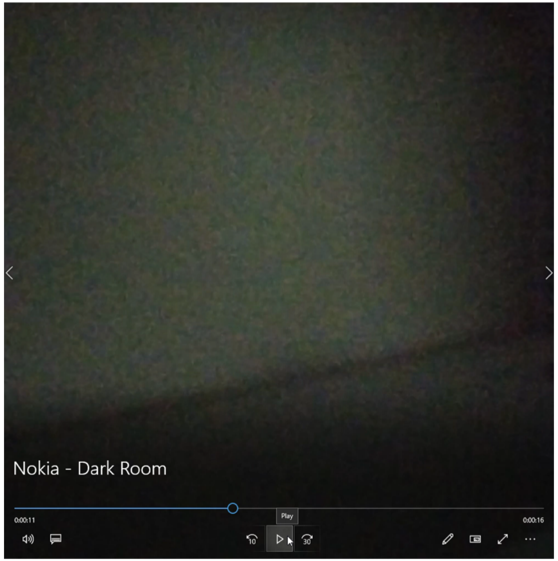
Nokia 7.1

Distortion	Setting	Hardware	
		Nokia 7.1	Magic Leap One
Gaussian noise ($\sigma_{GN}^2 \geqslant 0.003$)	Dark room (7lux)	617/1558=39.6%	infeasible
	Camera zoom-in (509lux)	1755/2185=97.2%	infeasible
Motion blur ($l \geqslant 5$)	Corridor (178lux)	2681/3452=77.6%	3760/3776=99.6%
	Sunny outdoor (9873lux)	16/2687=0.5%	957/2766=34.6%
Gaussian blur ($k \geqslant 5$)	Foggy	762/935=81.6%	infeasible
	Underwater	1250/1524=82.0%	infeasible

Measurements study in real-world mobile AR scenarios

Challenges

Large proportion of distorted images in real-world mobile AR scenarios.



Magic Leap One

Nokia 7.1

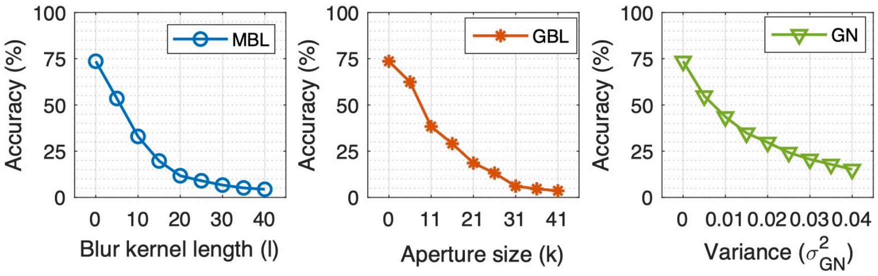
Distortion	Setting	Hardware	
		Nokia 7.1	Magic Leap One
Gaussian noise ($\sigma_{GN}^2 \geqslant 0.003$)	Dark room (7lux)	617/1558=39.6%	infeasible
	Camera zoom-in (509lux)	1755/2185=97.2%	infeasible
Motion blur ($l \geqslant 5$)	Corridor (178lux)	2681/3452=77.6%	3760/3776=99.6%
	Sunny outdoor (9873lux)	16/2687=0.5%	957/2766=34.6%
Gaussian blur ($k \geqslant 5$)	Foggy	762/935=81.6%	infeasible
	Underwater	1250/1524=82.0%	infeasible

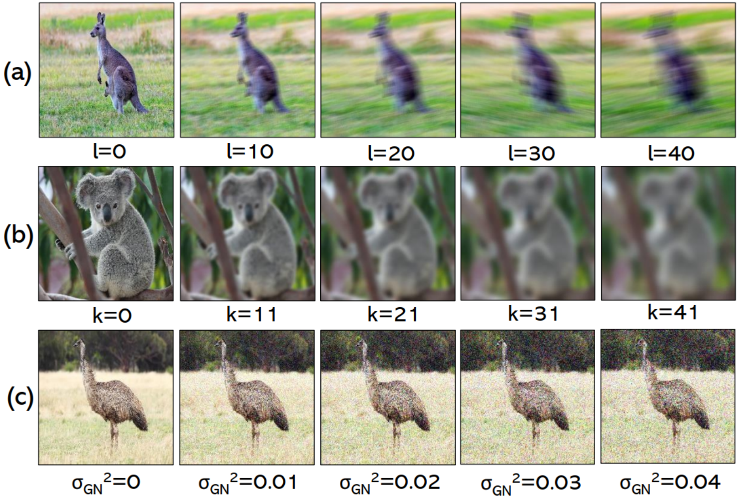
Measurements study in real-world mobile AR scenarios

Challenges

Multiple distortions cause dramatic performance degradation of well-trained DNNs for image recognition.

Impact of image distortions on MobileNetV2 trained with Caltech-256 dataset



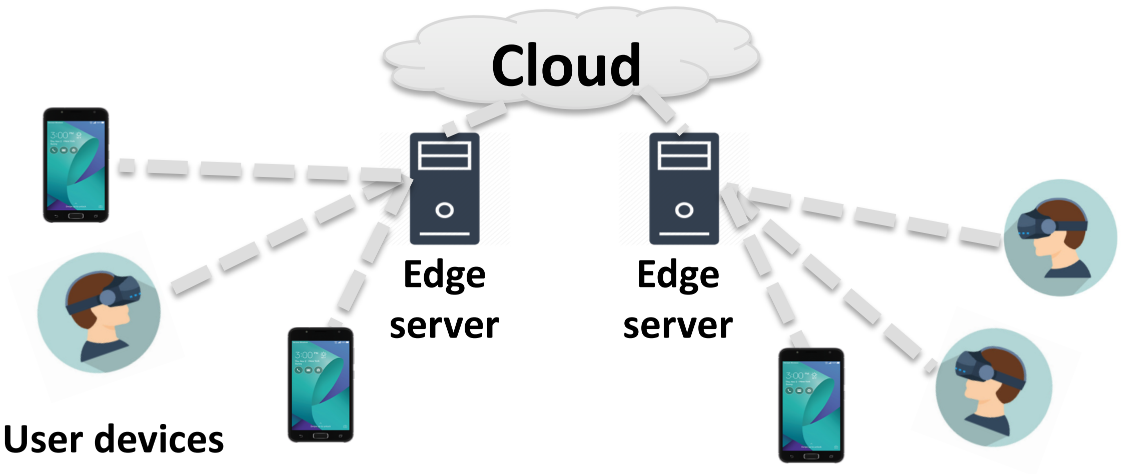


Outline

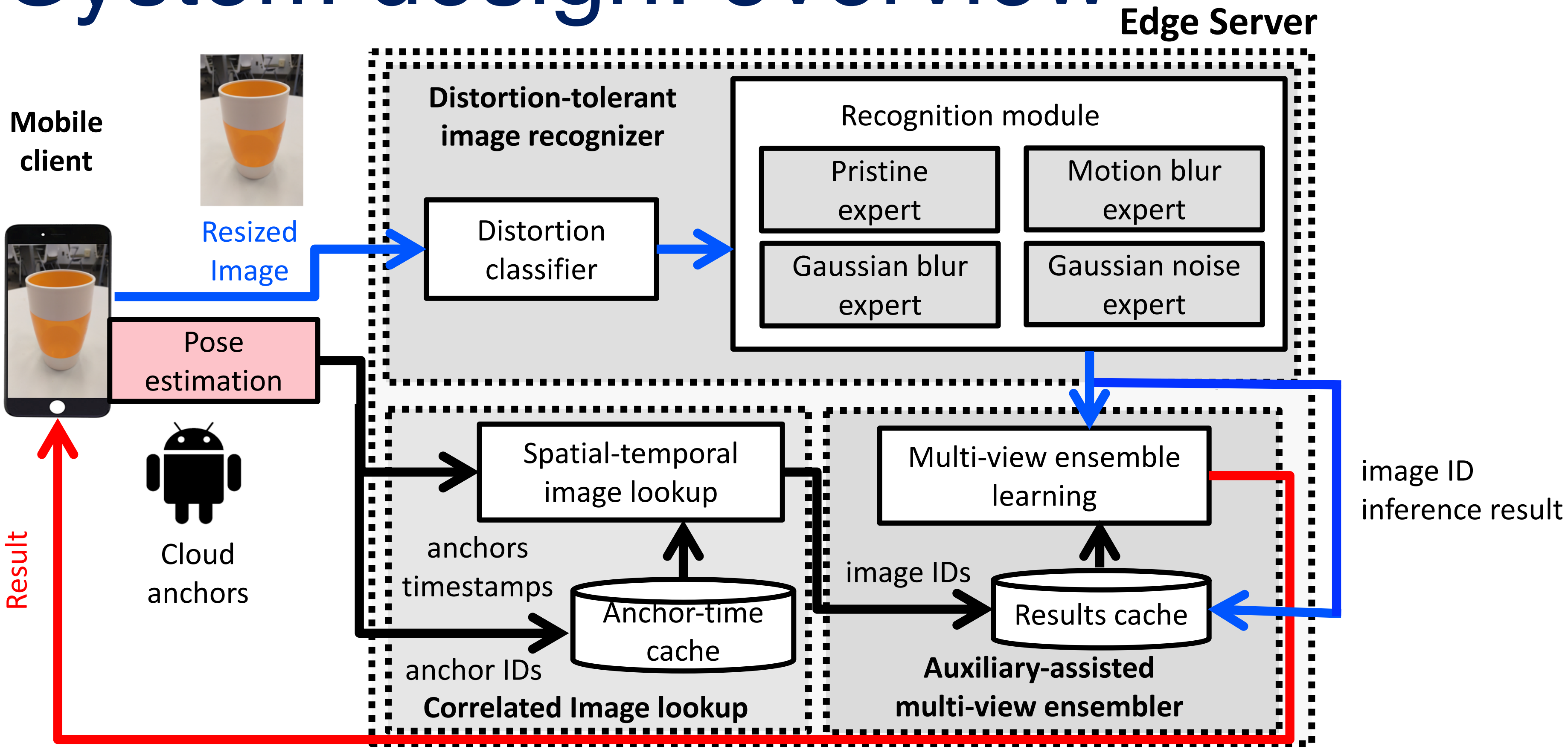
- Introduction
- **CollabAR system design**
- MVMDD dataset
- Evaluation

Our method

Collaborative image recognition: leverage the *temporally and spatially correlated images* to improve the image recognition accuracy in heterogeneous scenarios.

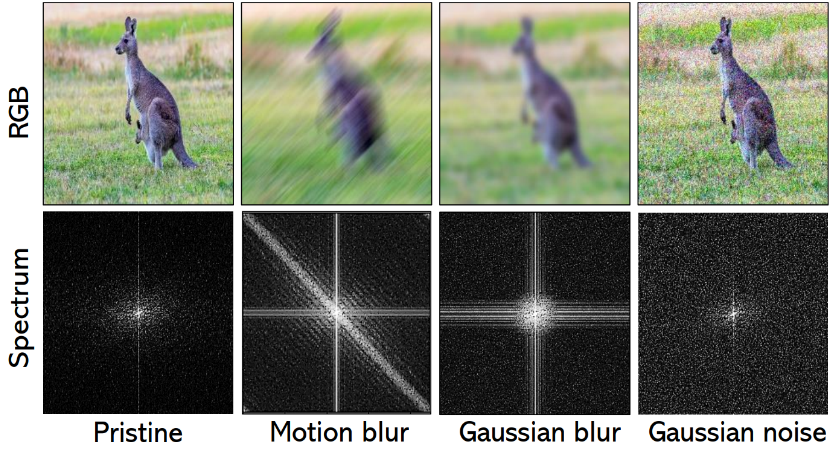


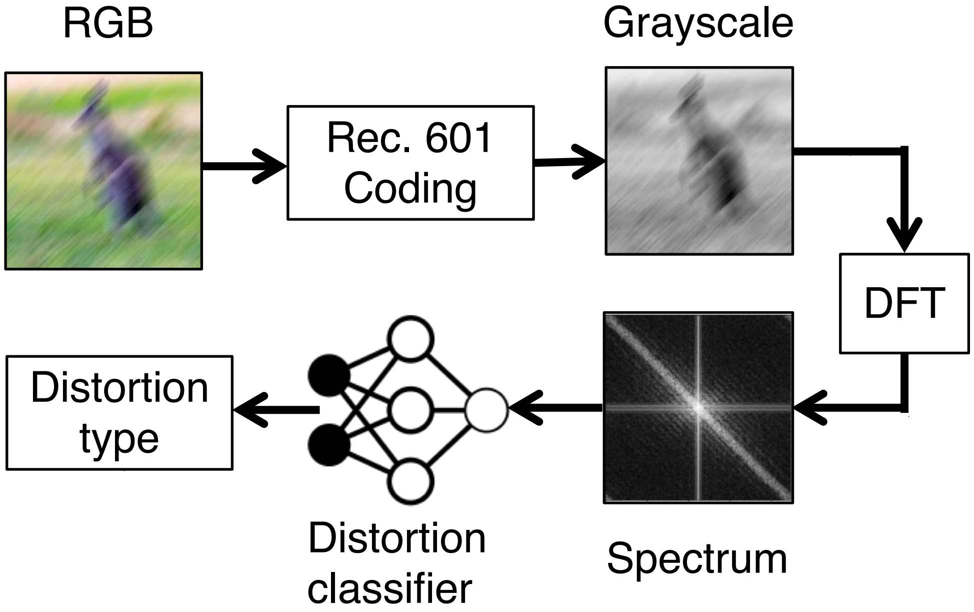
System design: overview



System design: distortion-tolerant image recognizer

1. Distortion Classifier

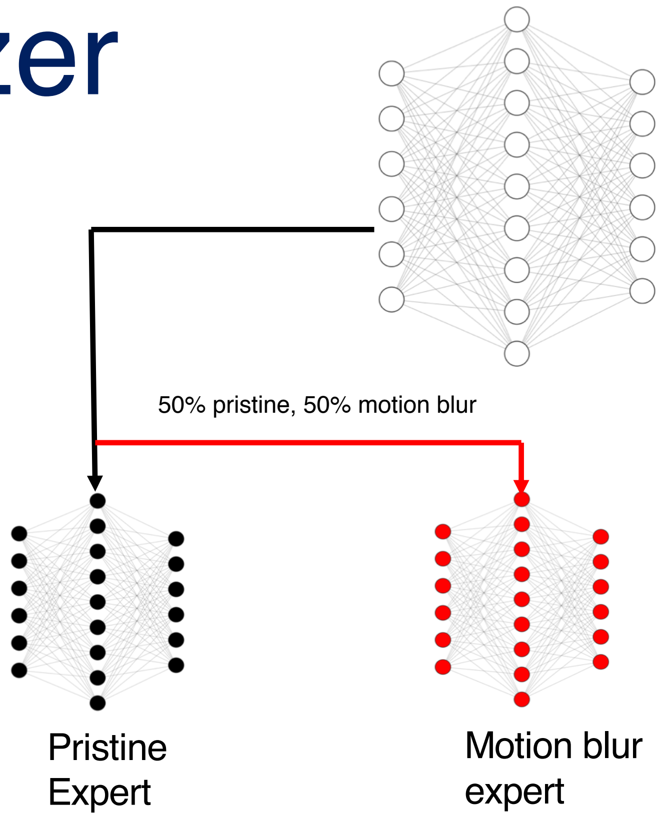




System design: distortion-tolerant image recognizer

2. Recognition Expert

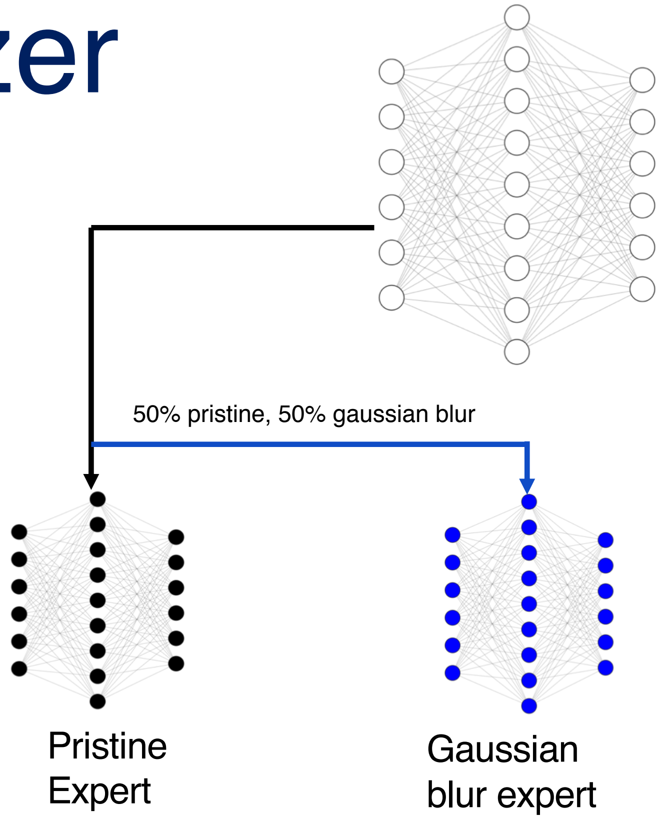
- Step 1: Initialize the CNN by training it on ImageNet dataset.
- Step 2: Fine-tuning the CNN by pristine images to get $Expert_p$.
- Step 3: Fine-tuning $Expert_p$ to get $Expert_{MB}$, $Expert_{GB}$ and $Expert_{GN}$.



System design: distortion-tolerant image recognizer

2. Recognition Expert

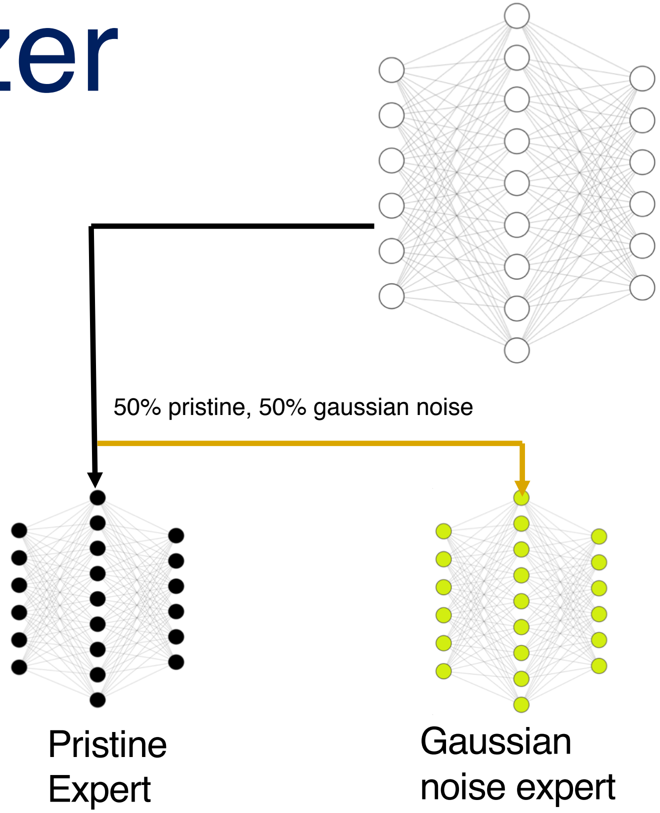
- Step 1: Initialize the CNN by training it on ImageNet dataset.
- Step 2: Fine-tuning the CNN by pristine images to get $Expert_p$.
- Step 3: Fine-tuning $Expert_p$ to get $Expert_{MB}$, $Expert_{GB}$ and $Expert_{GN}$.



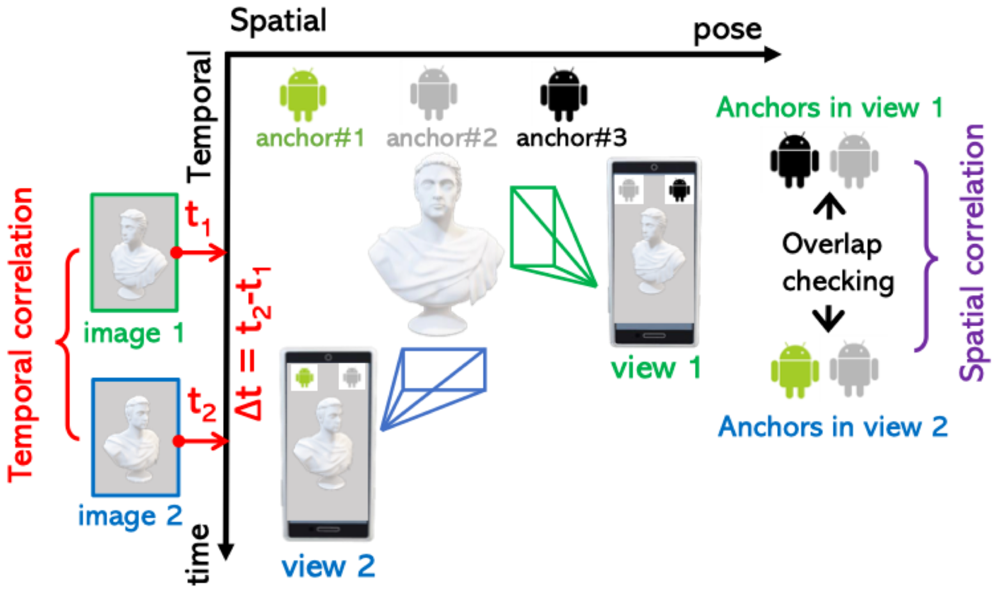
System design: distortion-tolerant image recognizer

2. Recognition Expert

- Step 1: Initialize the CNN by training it on ImageNet dataset.
- Step 2: Fine-tuning the CNN by pristine images to get $Expert_p$.
- Step 3: Fine-tuning $Expert_p$ to get $Expert_{MB}$, $Expert_{GB}$ and $Expert_{GN}$.



System design: spatial-temporal image lookup



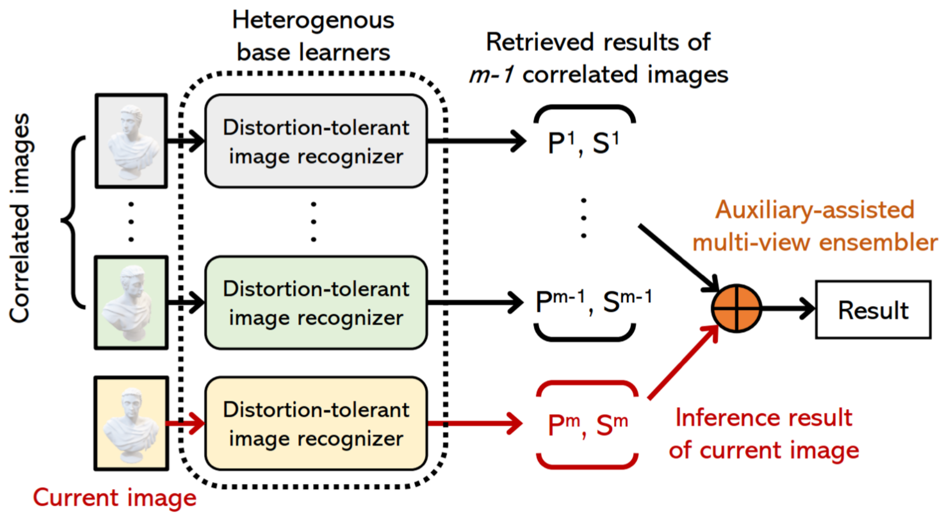
1. Spatial correlation look up: Different images contain the same anchor:

$$\{anchorIDs\}_{new} \cap \{anchorIDs\}_{Cached} \neq 0$$

2. Temporal correlation look up: Different images that are taken within a certain freshness:

$\Delta t = t_{New} - t_{Cached}$ and $\Delta t < T_{fresh}$

System design: auxiliary-assisted multi-view ensemble learning (AMEL)



1. Normalized entropy:

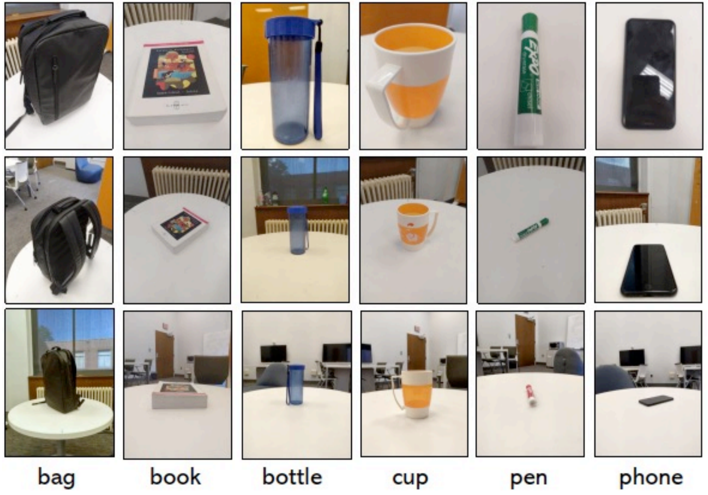
$$S^k(P^k) = \sum_{i=1}^{|C|} \frac{p^i log(p^i)}{log\,|C|}$$

- $P^k = \{p^1 \ldots p^{|C|}\}$ – probability vector of image k
- S^k – weight of image k
- C – number of classes

2. Ensemble multi-user results based on normalized entropy:

$$P = \sum_{i=1}^{m} (1 - S^i) P^i$$

Multi-view multi-distortion image dataset (MVMDD)



Examples of the pristine images that are collected in our MVMDD dataset

Pristine image set	Object categories	6
	Number of views	6
	Background complexity	2
	Size of object in image	3
	Number of instances	6
Total pristine images	6×6×2×3×6=1,296	
Augmented image set	Types of distortion	3
	Distortion levels	8
Total augmented images	1,296×3×8=31,104	
Total images	32,400	

Summary of collected MVMDD dataset

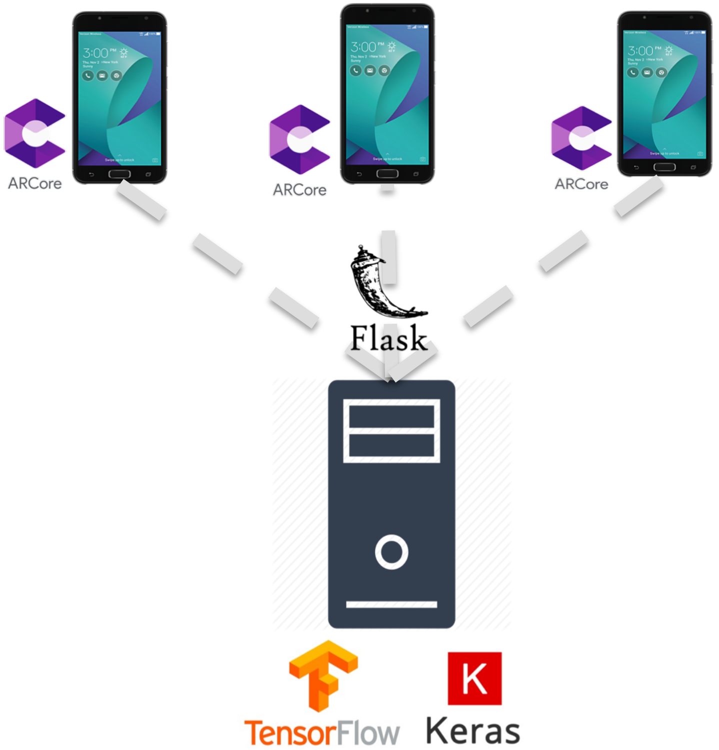
Outline

- Introduction
- CollabAR system design
- MVMDD dataset
- **Evaluation**

Evaluation: experiment setup

1. Implementation:

    - **Client:** android smartphones with Google ARCore SDK.
    - **Edge server:** a desktop with an Intel i7-8700k CPU and a Nvidia GTX 1080 GPU.
    - **Deep learning framework:** Keras 2.3 on top of the TensorFlow 2.0.
    - **Communication:** python Flask framework through HTTP protocol.



Evaluation: experiment setup

2. Benchmark datasets:

- Dataset collected by ourselves
- Existing standard dataset

- **Single-view datasets:**
  - **Caltech-256:**
    - 257 categories, 80 instances every category.
  - **MobileDistortion:**
    - 4 distortion types, 300 image instances for each distortion type.
- **Multi-view datasets:**
  - **MIRO:**
    - 12 categories, 10 objects each category, 160 views each object, black background.
  - **MVMDD:**
    - 6 categories, 6 objects each category, 6 views each object, 3 different distances, 2 background complexity levels.

Evaluation: distortion classifier performance

Real-world distortions

Accuracy: 95.50%

Output Class \ Target Class	Pristine	MBL	GBL	GN
Pristine	100.0% 100	0.0% 0	0.0% 0	0.0% 0
MBL	0.0% 0	100.0% 82	15.3% 18	0.0% 0
GBL	0.0% 0	0.0% 0	84.7% 100	0.0% 0
GN	0.0% 0	0.0% 0	0.0% 0	100.0% 100

MobileDistortion

Synthesized distortions

Accuracy: 94.21%

Output Class \ Target Class	Pristine	MBL	GBL	GN
Pristine	93.2% 723	3.5% 28	0.0% 0	5.4% 44
MBL	3.9% 30	93.1% 743	3.9% 31	0.1% 1
GBL	0.1% 1	3.4% 27	96.1% 755	0.0% 0
GN	2.8% 22	0.0% 0	0.0% 0	94.5% 775

Caltech-256

Accuracy: 92.92%

Output Class \ Target Class	Pristine	MBL	GBL	GN
Pristine	95.6% 172	2.7% 4	0.0% 0	0.0% 0
MBL	0.0% 0	97.3% 143	18.1% 39	0.0% 0
GBL	0.0% 0	0.0% 0	81.9% 176	0.0% 0
GN	4.4% 8	0.0% 0	0.0% 0	100.0% 178

MIRO

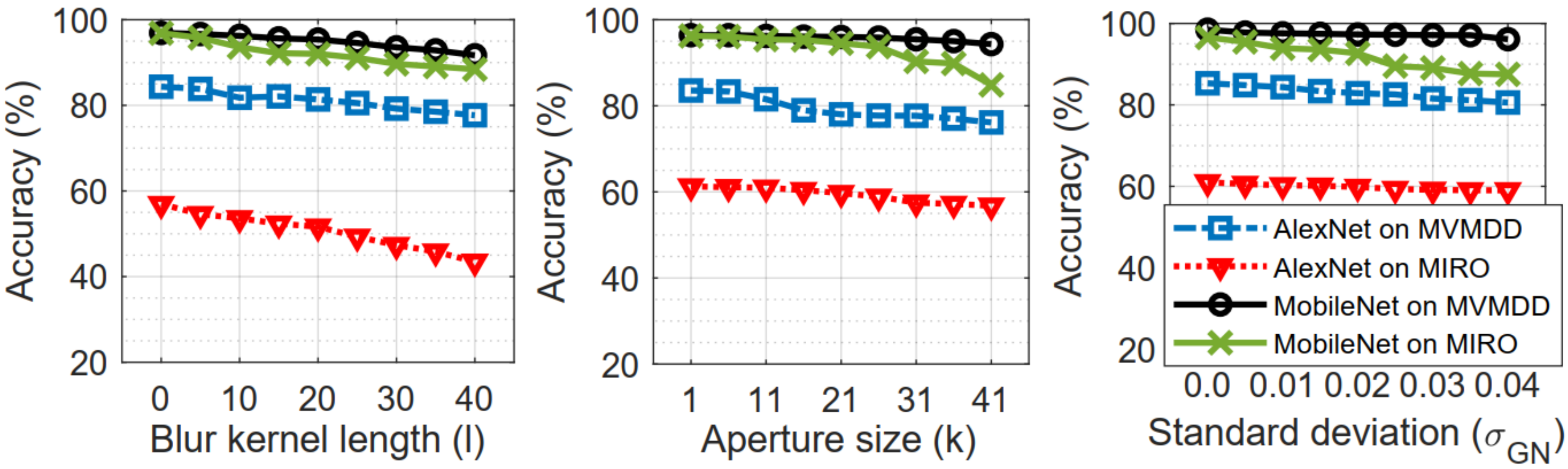
Accuracy: 93.33%

Output Class \ Target Class	Pristine	MBL	GBL	GN
Pristine	98.6% 72	0.0% 0	0.0% 0	7.0% 8
MBL	1.4% 1	100.0% 91	15.3% 17	0.0% 0
GBL	0.0% 0	0.0% 0	84.7% 94	0.0% 0
GN	0.0% 0	0.0% 0	0.0% 0	93.0% 107

MVMDD

Evaluation: image recognition accuracy

1. Performance of distortion expert.

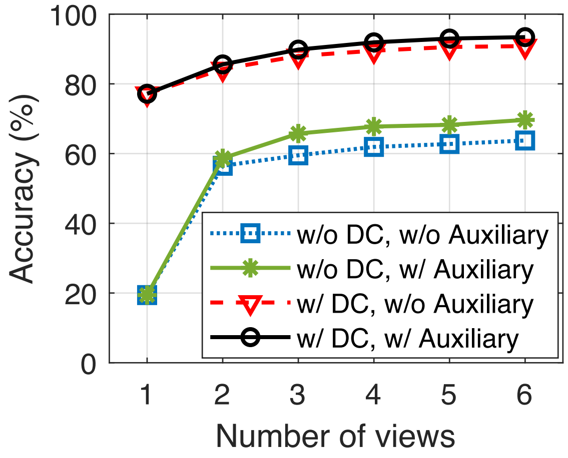


(a) Motion blur **(b) Gaussian blur** **(C) Gaussian noise**

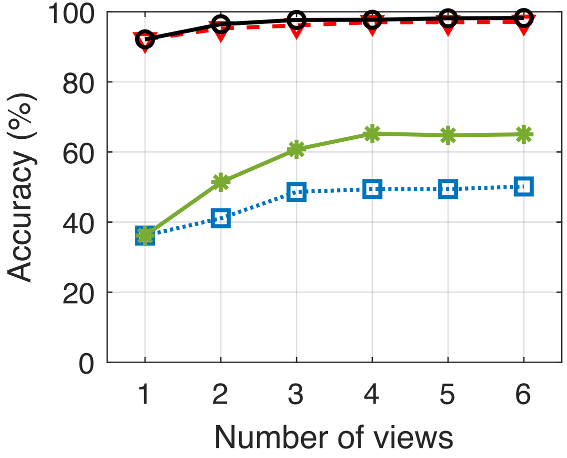
Evaluation: image recognition accuracy

2. Performance of multi-view collaboration:

- Multi-view single-distortion



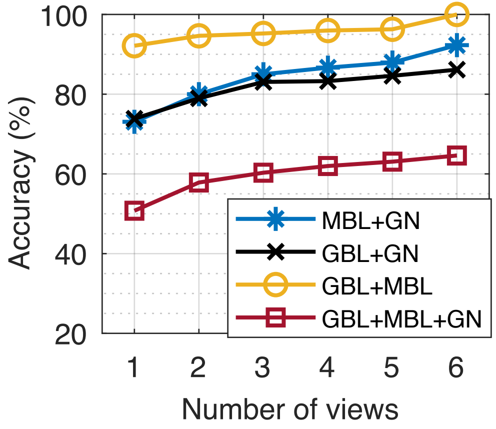
(a) MIRO



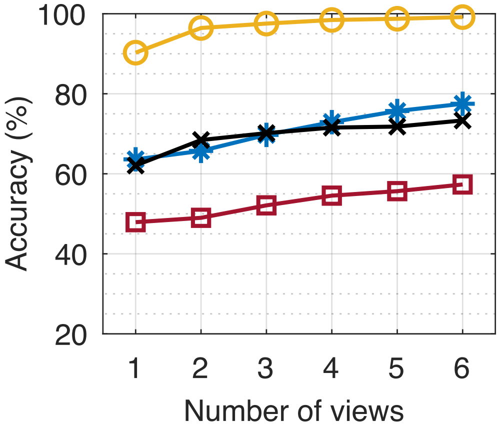
(a) MVMDD

Evaluation: image recognition accuracy

2. Performance of multi-view collaboration:
   - Multi-view multi-distortion



(a) MVMDD

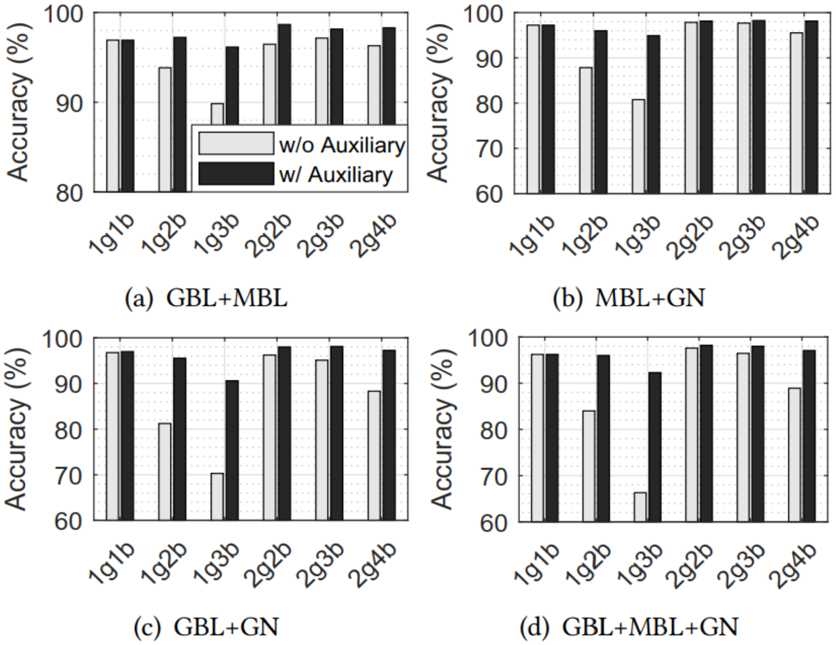


(b) MIRO

Evaluation: image recognition accuracy

3. Advantages of AMEL

- Good view: the image is a pristine image.
- Bad view: the image contains multiple distortions with high distortion levels.



CollabAR accuracy on the MVMDD dataset with and without the auxiliary feature

Evaluation: system profiling

1. Computational latency

Processing Component	Processing Unit	Hardware Platform			
		Edge server	Nokia 7.1	Pixel 2 XL	Xiaomi 9
AMEL (AlexNet)	CPU	48.0	255.4	189.9	124.2
	GPU	6.8	331.0	127.5	71.2
AMEL (MobileNetV2)	CPU	28.1	108.4	81.9	33.7
	GPU	4.7	46.9	26.0	33.1
Distortion classifier	CPU	4.4	29.3	22.33	13.1
	GPU	3.4	20.3	8.7	8.4
Lowest overall latency	CPU	32.5	137.7	104.2	46.8
	GPU	$\boxed{8.1}$	67.2	34.7	41.5

2. Total end to end latency

Nokia 7.1	Pixel 2 XL	Xiaomi 9
32.7ms	21.5ms	17.8ms

Open source

We have open sourced the MVMDD dataset and the code of CollabAR!

- MVMDD: https://github.com/CollabAR-Source/MVMDD.
- The code of CollabAR: https://github.com/CollabAR-Source/CollabAR-Code.

Acknowledgements

- Lord Foundation of North Carolina
- NSF awards CSR-1903136, CNS-1908051

Thank you & Questions?