Benchmarking for IU

Image Understanding
VU 186.846, SS2018

Author: Tomáš Ondruch
Student ID: 11740257

18.04.2018

Outline

1. Introduction to Benchmarking in IU
   - → definition & basic idea
   - → CVOnline: Image Databases
   - → CV research projects
2. State-of-the-art highlights
   - → Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images
   - → Boosting Object Proposals: From Pascal to COCO
   - → The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes
   - → How Good Is My Test Data? Introducing Safety Analysis for Computer Vision
3. Summary & Conclusion

Benchmarking

Basic idea

- a technique of strategic management (from early 1980s)
- “Benchmarking is the process of **comparing** a company's performance to the performance of other companies.” [1]
- nowadays benchmarking software, benchmarking in healthcare / education / …

Benchmarking in IU

- Comparison of
  - CV algorithms’ performance
  - benchmark datasets
  - evaluation metrics, annotation [2]

IU & Datasets

- for algorithm learning, validation and evaluation
- images / video sequences

CVOnline: Image Databases

http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm

- collated list of datasets for CV research purposes
- November 2016: 670 datasets
- various topics:
  - traffic scenes
  - facial expression datasets
  - fingerprints
  - mice activity [3], ear recognition [4],…

Index by Topic

1. Action Databases
2. Attribute recognition
3. Autonomous Driving
4. Biological/Medical
5. Camera calibration
6. Face and Eye/Iris Databases
7. Fingerprints
8. General Images
9. General RGBD and depth datasets
10. General Videos
11. Hand, Hand Grasp, Hand Action and Gesture Databases
12. Image, Video and Shape Database Retrieval
13. Object Databases
14. People (static and dynamic), human body pose
15. People Detection and Tracking Databases (See also Surveillance)
16. Remote Sensing
17. Scenes or Places, Scene Segmentation or Classification
18. Segmentation
19. Simultaneous Localization and Mapping
20. Surveillance and Tracking (See also People)
21. Textures
22. Urban Datasets
23. Vision and Natural Language
24. Other Collection Pages
25. Miscellaneous Topics

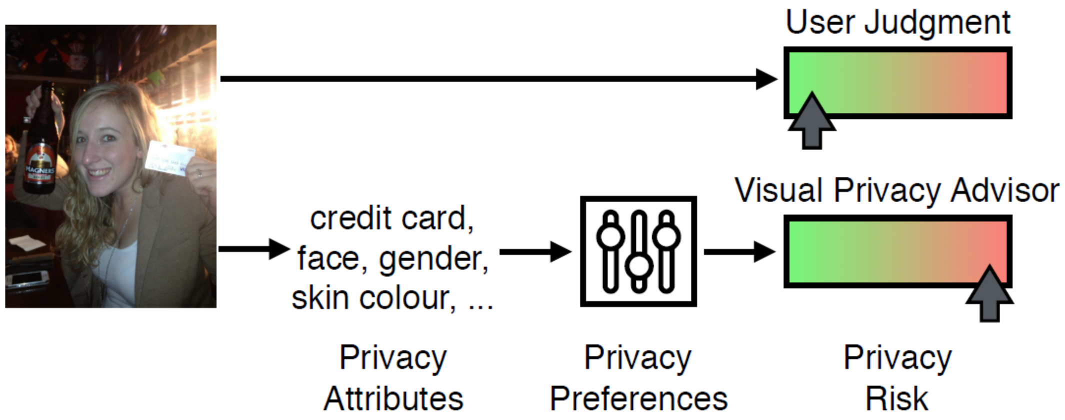
CVOnline: List of topics

Image datasets & related projects

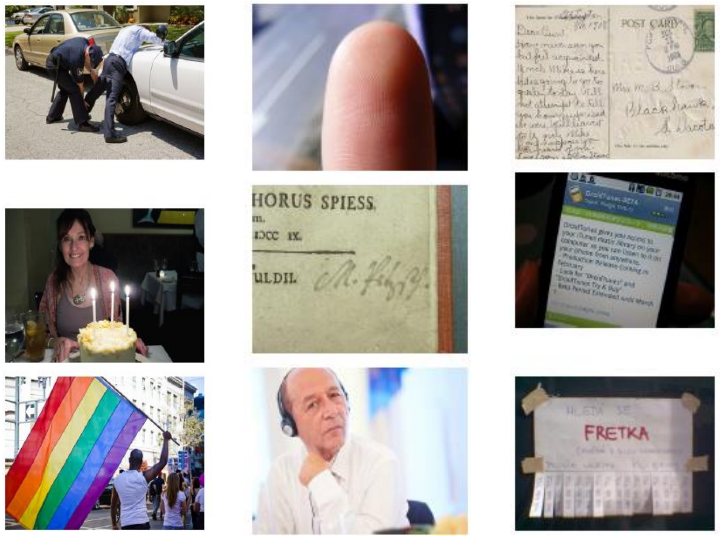
- focus on very specific problem
- advancement presented in paper is typically of similar kind
- usual sequence of steps:
    1. Introduction to problem
    2. Review of related work & identification of imperfection
    3. Presentation of paper’s contribution
        a) novel dataset (with innovative features, improvements)
        b) metrics definition
        c) algorithm testing, evaluation and comparison
    4. Conclusion (+ call to action, outlook)

Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images [5]

- user-specific privacy feedback from image content
- privacy risk prediction
- VISPR Dataset
  - 68 image attributes (gender, passport, medical history, tattoo,…) – novel issue
  - 22k manually annotated images



Visual Privacy Advisor Model



Sample Images from VISPR Dataset

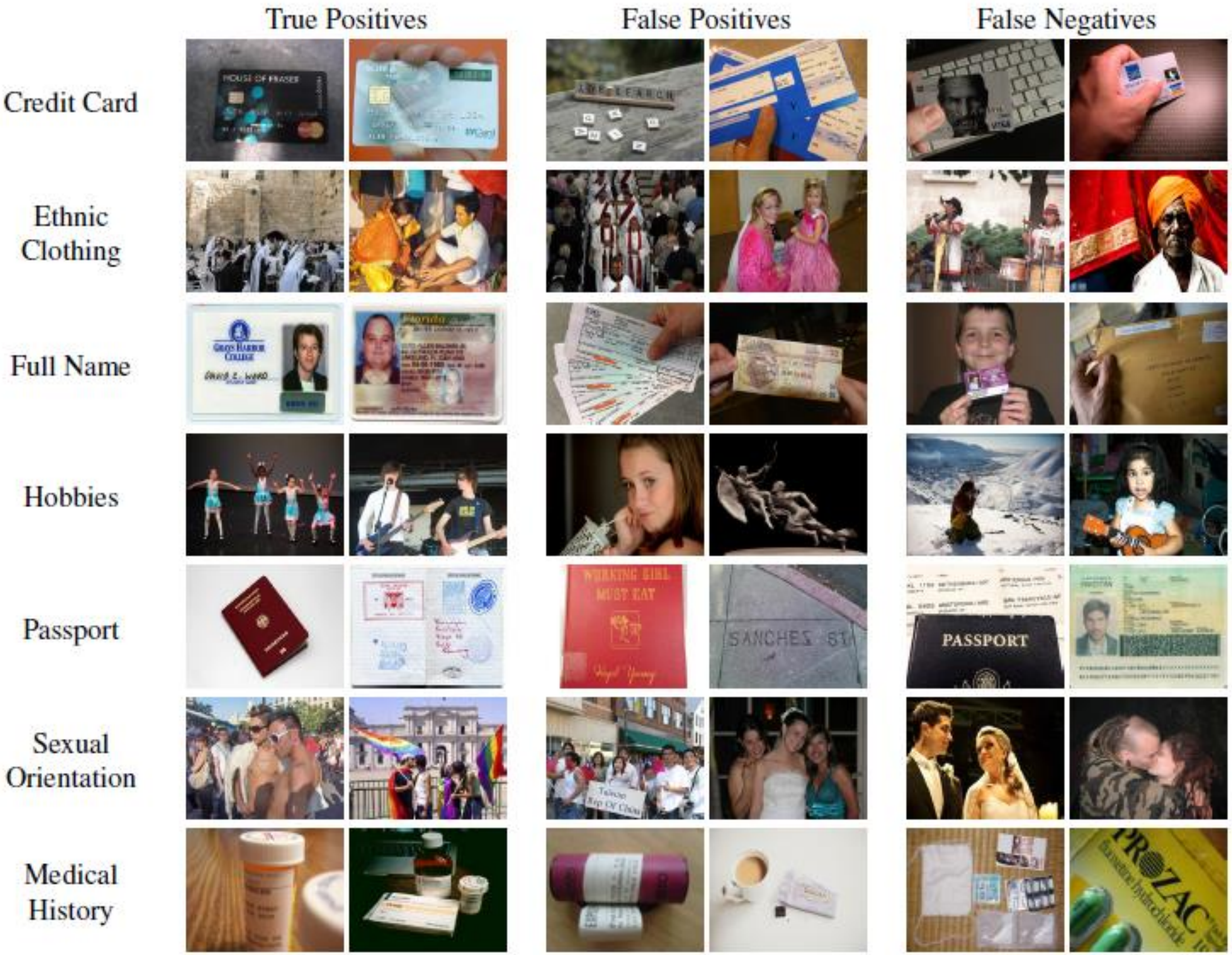
Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images [5]

- User study in two steps
  1) questions on privacy preferences
  2) visual privacy judgement
- Recognition model
  1. Multilabel (user-independent) classification problem
     $\rightarrow$ Challenging problem
  2. Metrics: Average Precision (AP), C-MAP
  3. Methods: deep CNNs (CaffeNet, GoogleNet, ResNet-50) supported by SVM
- Privacy risk (user-dependent) prediction
  - two PRE Methods: AP-PR, PR-CNN
  - qualitative and quantitative evaluation
- Final comparison of PRE Methods vs. Users' Visual Risk Assessment

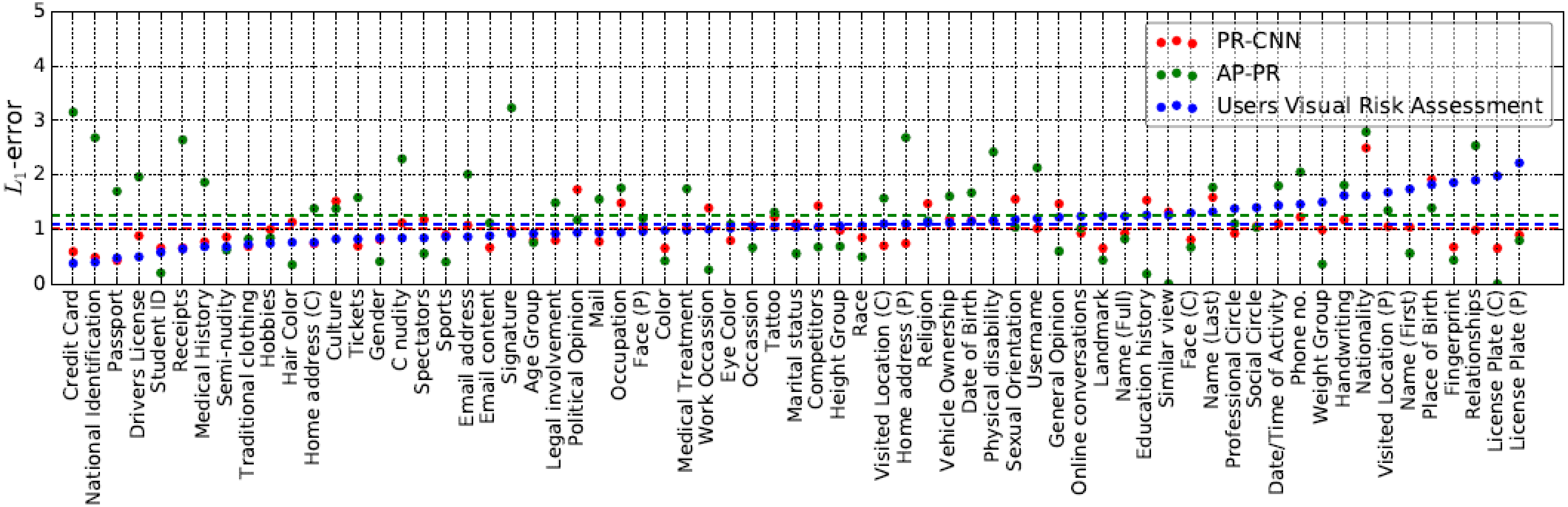
Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images [5]

Results:

1) Good association of privacy attributes to distinctive visual cues (clothing, nudity, text,…) vs. occasional misinterpretation (identification of driving licence, first name/ last name)
2) PR-CNN better for high-risk images, AP-PR better for low-risk images
3) On average, PR-CNN outperforms human judgement



Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images [5] – Qualitative Results



Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images [5]
L1 Errors over attributes

Datasets – further remarks

- training and evaluation of various algorithms
- benchmarking & comparing different approaches

- designed to challenge algorithms with issues they currently struggle with
  $\rightarrow$ methods' improvement + introduction of new methods $\rightarrow$ **<u>dataset's limited lifespan</u>**

- Solution:
  1. Dataset's maintenance and update by community
  2. Introduction of a new dataset

Boosting Object Proposals: From Pascal to COCO [6]

- study of transition from Pascal Dataset (SegVOC12) and Semantic Boundary Dataset (SBD) to Microsoft Common Objects in Context (COCO)
- field of CV: object segmentation and annotation

Pascal SBD COCO

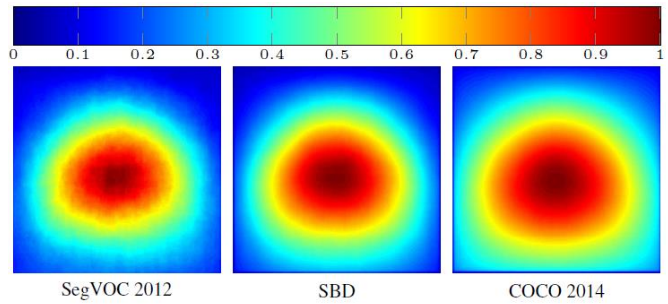
	Number of Categories	Number of Images	Number of Instances
SegVOC12	20	2 913	6 934
Train+Val		*1 464+1 449*	*3 507+3 427*
SBD	20	11 355	26 843
Train+Val		*8 498+2 857*	*20 172+6 671*
COCO14	80	123 287	886 284
Train+Val		*82 783+40 504*	*597 869+288 415*

Size of the databases

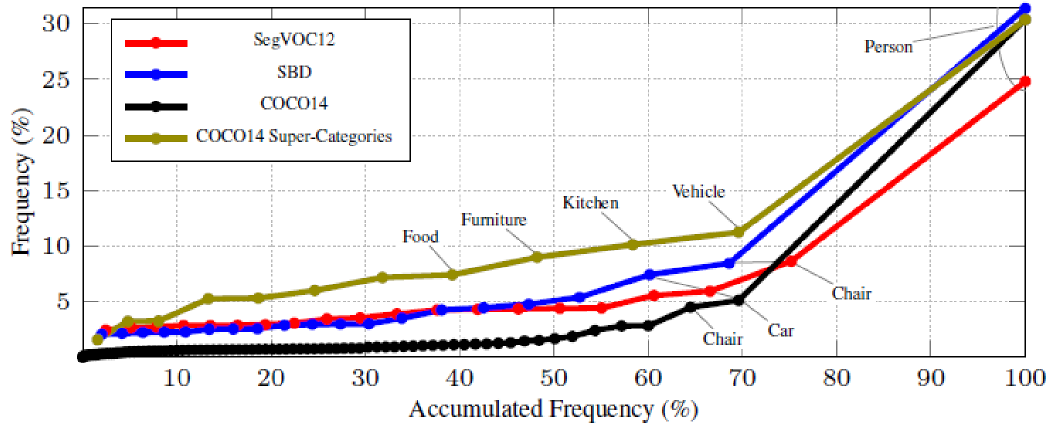
= **benchmark update**

Boosting Object Proposals: From Pascal to COCO [6]

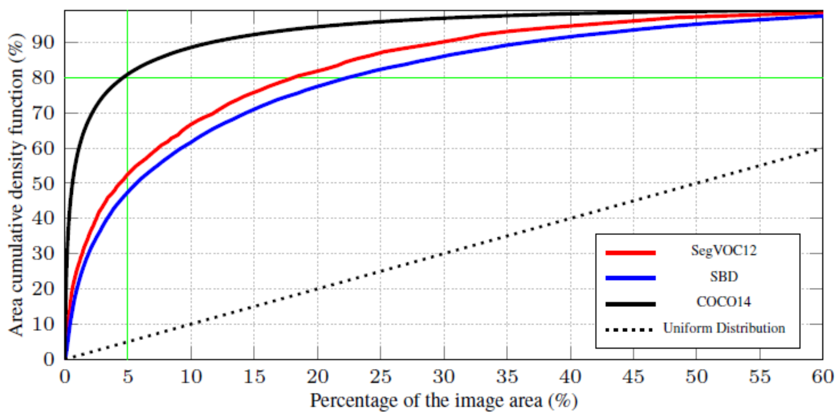
- in-depth comparison of datasets (Pascal, SBD, COCO)
  - size
  - category balance
  - annotated instances localization
  - annotated instances areas



Annotated instances localization



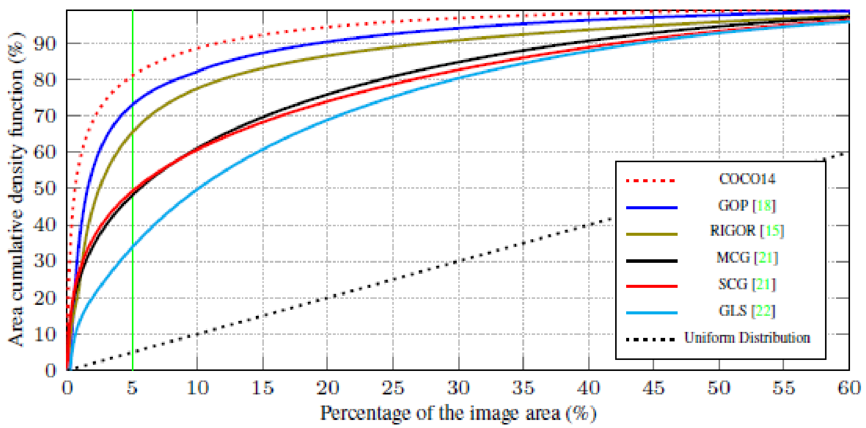
Category balance



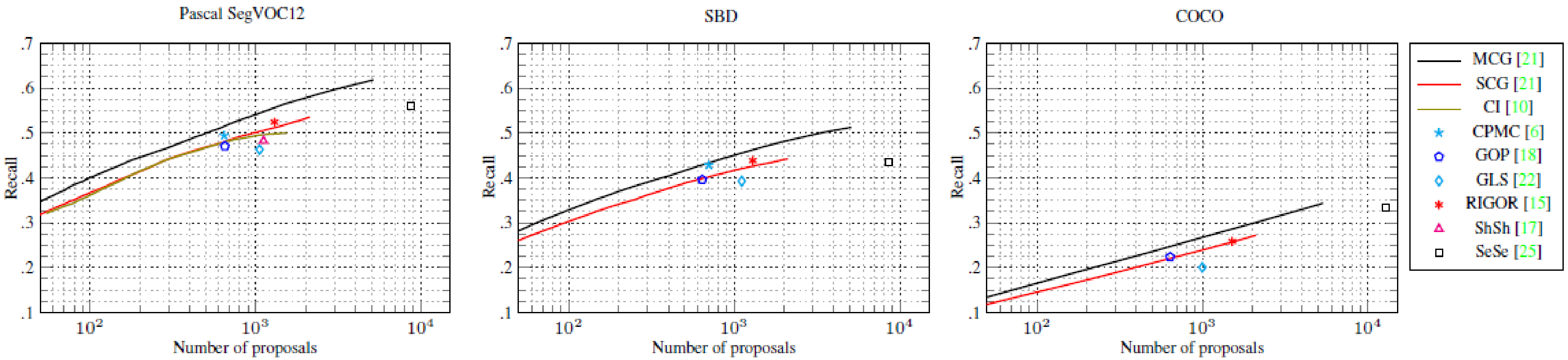
Annotated instances areas

Boosting Object Proposals: From Pascal to COCO [6]

- Analysis of SoA object proposal techniques on COCO
- MCG, GOP, SCG, RIGOR, SeSe, GLS
    - timing
    - average recall (AR) score
    - per-category evaluation
    - area and localization of the proposals
    - quality vs. object areas
    - superpixel computation COCO



Area of the proposals



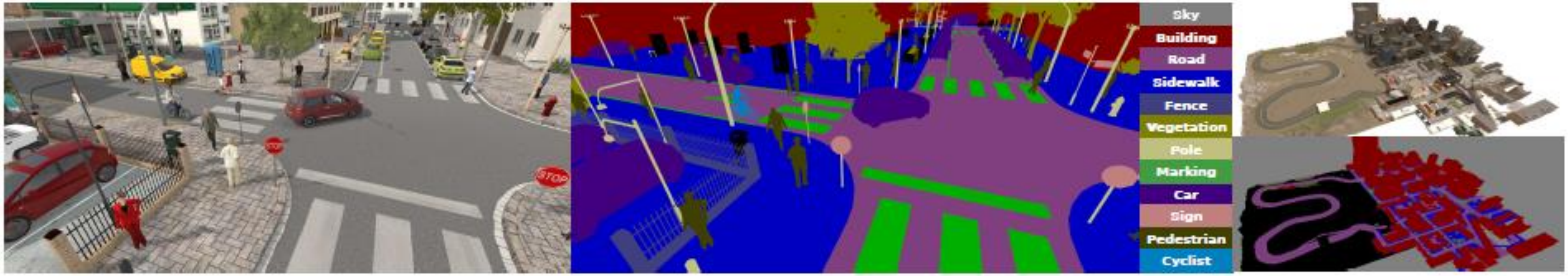
← AR score

Boosting Object Proposals: From Pascal to COCO [6]

- Results:
  1. All datasets biased towards small objects centered in the image
  2. Lower AR-score and superpixel computation on COCO $\rightarrow$ more challenging dataset
  3. All object proposal techniques biased towards small objects
  4. MCG, GOP: the smaller the object, the lower the quality of segmentation proposal
  5. Combination of techniques yield boosted performance

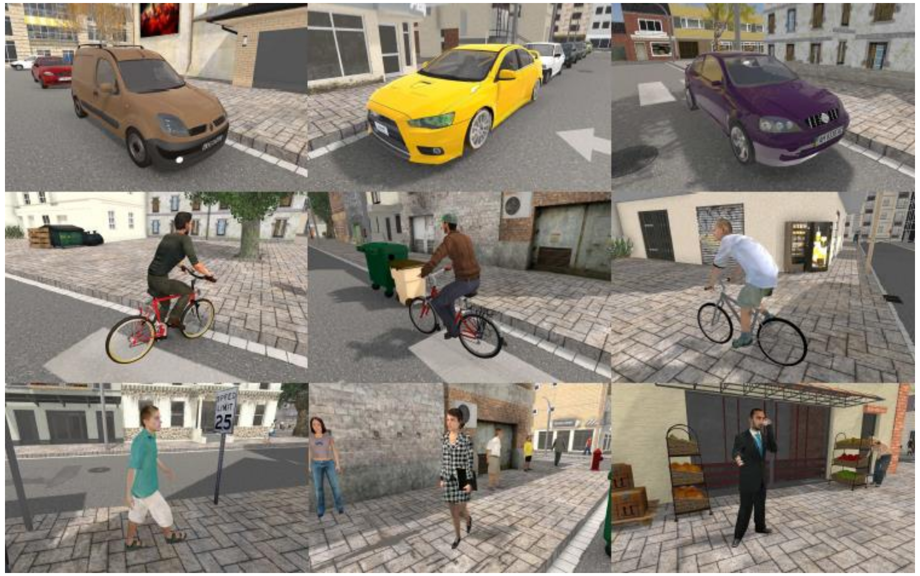
The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes [7]

- problem: tiresome process of annotating images for DCNNs' training
- presented solution: generating virtual worlds' realistic images with pixel-level annotations
- field of CV: vision-based semantic segmentation (in autonomous driving)

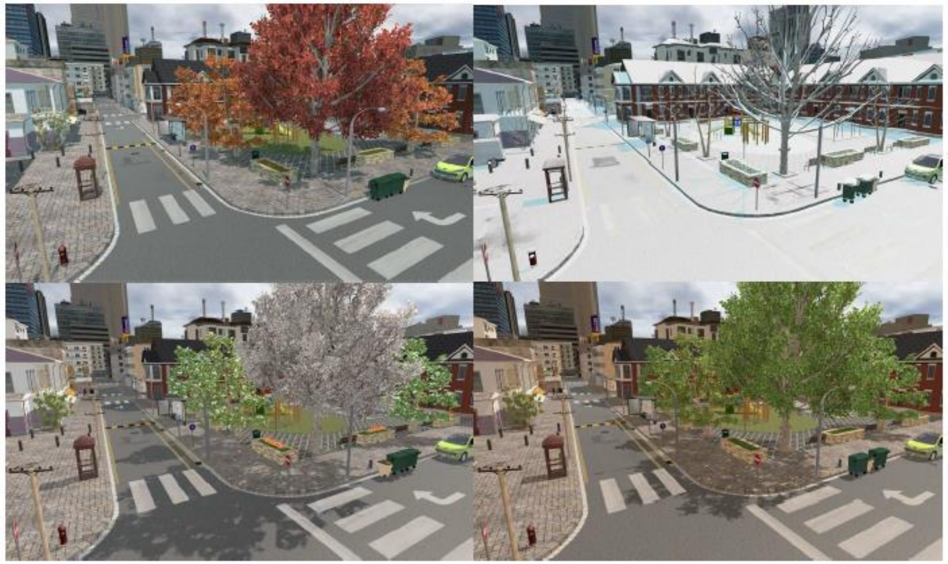


The SYNTHIA Dataset: A sample frame (Left) with its semantic labels (center) and a general view of the city (right)

The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes [7]



SYNTHIA – Examples of dynamic objects



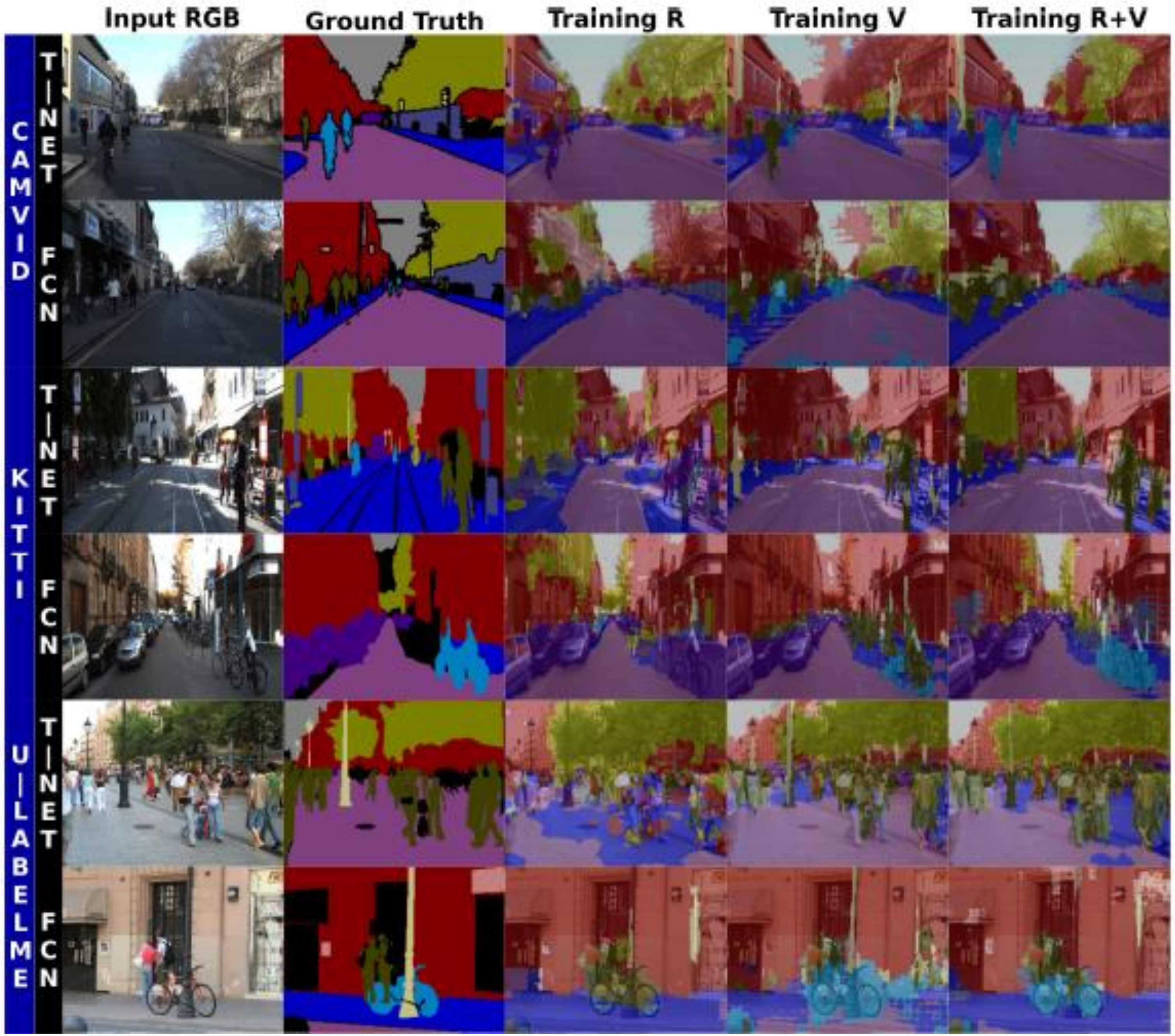
SYNTHIA – Visualisation of four seasons

The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes [7]

- SYNTHIA Dataset
  - rendered from virtual city created with the Unity development platform
  - precise pixel-level semantic annotations
  - 13 classes (sky, building, traffic signs, vegetation,…)
  - multiple view-points
  - two sets
    1. SYNTHIA-Rand – 13 400 images, from randomly moving camera
    2. SYNTHIA-Seqs – 4 video sequences, each of 50 000 frames
- real image datasets (driving scene sets) – CamVid, KITTI, Urban LabelMe,…
- tested methods for semantic segmentation
  1. T-Net – deep CNN, easy to train
  2. FCN – state-of-the-art in the field

The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes [7]

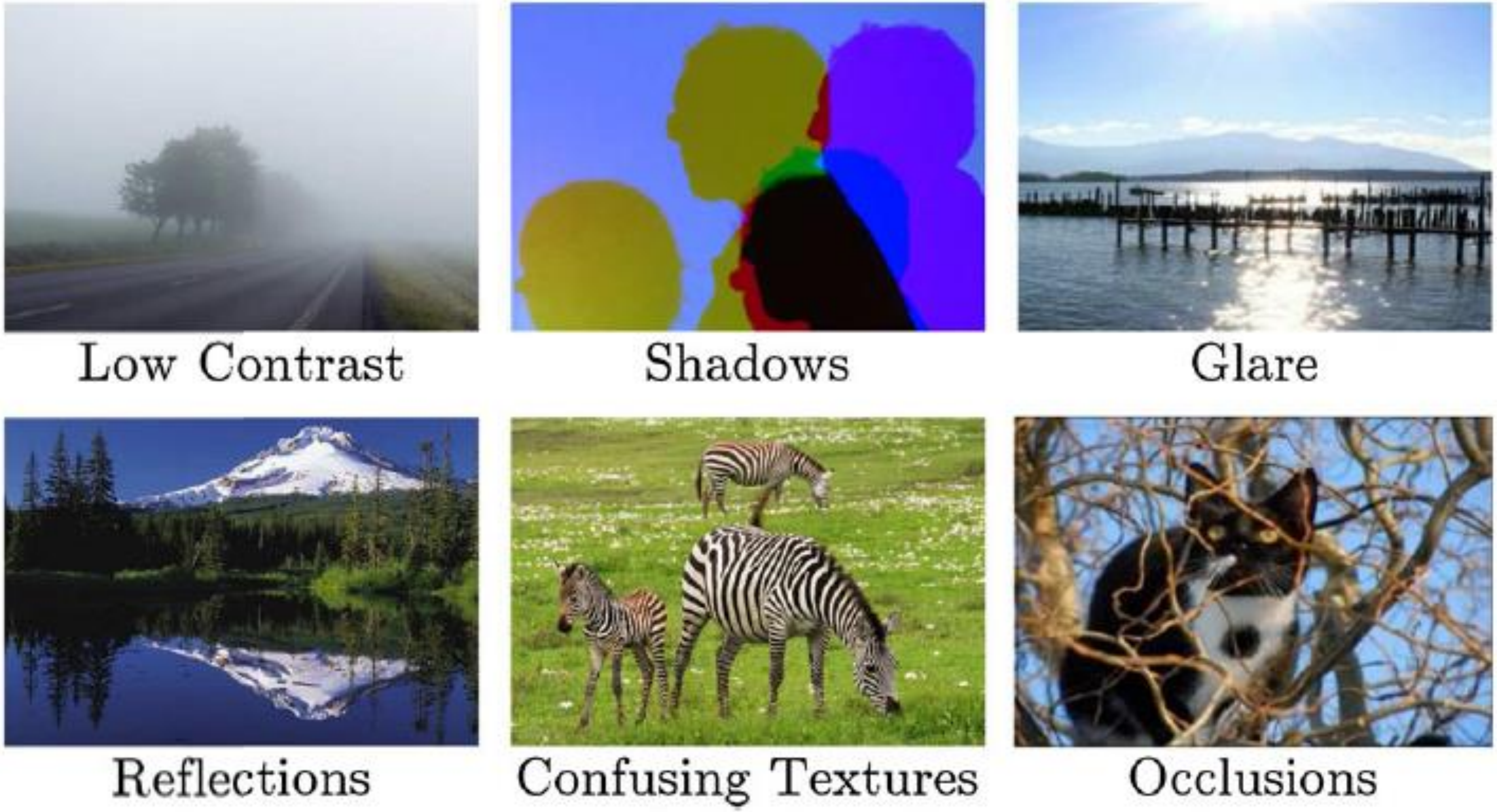
- Sketch of experimental evaluation
  1. Training the T-Net and FCN architectures on
     a) Synthetic data only
     b) Real image datasets only
     c) Real image datasets combined with data from SYNTHIA
  2. Evaluation of total and per-class accuracy for each architecture
  3. Quantitative & qualitative comparison of performance
- Results
  1. Training on synthetic data: Good results in recognizing roads, buildings, cars and pedestrians
  2. Per-class accuracy boosted for mixed training sets
  3. Mixed data produces smooth and very accurate results in segmentation



Qualitative results for different testing datasets and architectures (T-Net and FCN)

How Good Is My Test Data? Introducing Safety Analysis for Computer Vision [8]

- problem setting: Algorithms scoring high in public benchmarks perform rather poor in real world scenarios
- datasets rarely have to undergo independent evaluation
- Crucial questions from the field of CV validation:
  1. What should be part of the test dataset to ensure that the required level of robustness is achieved?
  2. How can redundancies be reduced (to save time and remove bias due to repeated elements)?
- Visual hazards = elements and relations known to be difficult for a CV algorithm (like optical illusions for humans)
- Answers to questions above:
  1. Ensure completeness of test datasets by including all relevant hazards from the list.
  2. Reduce redundancies by excluding test data that only contains hazards that are already identified.



Examples for potential visual hazards for CV algorithms

How Good Is My Test Data? Introducing Safety Analysis for Computer Vision [8]

<u>Solution: application of the HAZOP risk assessment method to the CV domain</u>

- HAZOP = hazard and operability analysis
- initially in chemical industry, aircraft industry,…
- systematic process to identify potential risks of a system

Contribution and results

1. Generic model of information flow analyzed with HAZOP
2. CV-HAZOP checklist of 947 visual hazards created
   i. evaluation of the quality and thoroughness of test datasets
   ii. lead to improvement in evaluation of robustness of CV algorithms
3. Hazard list applicable further on already existing datasets
4. Statistical significance test: identified hazards reduce output quality

HID	Location/parameter	Guide word	Meaning	Consequence	Example
125	Light source/intensity	More	Light source shines stronger than expected	Too much light in scene	Overexposure of lit objects
481	Object/reflectance	As well as	Obj. has both shiny and dull surface	Diffuse reflection with highlight/glare	Object recognition distorted by glares
445	Object/texture	No	Object has no texture	Object appears uniform	No reliable correspondences can be found
706	Objects/reflectance	Close	Reflecting Obj. is closer to Observer than expected	Reflections are larger than expected	Mirrored scene taken for real
584	Objects/positions	Spatial periodic	Objects are located regularly	Same kind of objects appear in a geometrically regular pattern	Individual objects are confused
1059	Optomechanics/aperture	Where else	Inter-lens reflections project outline of aperture	Ghosting appears in the image	Aperture projection is mis-interpreted as an object
1123	Electronics/exposure	Less	Shorter exposure time than expected	Less light captured by sensor	Details uncorrelated due to underexposure

Examples from CV-HAZOP checklist

Examples for each entry in table above

Summary

- CV research is boosting, rapid progress in recent years
- Various specific tasks in CV ⇒ need for appropriate benchmarks
- Emergence of innovative approaches in image collection / annotation / segmentation / …
- Need for dataset updates / benchmark switch
- Benchmark dataset’s quality ⇒ CV algorithm evaluation quality

References

[1] "Benchmarking Definition | Benchmarking Techniques • The Strategic CFO." The Strategic CFO. March 01, 2018. Accessed April 17, 2018. https://strategiccfo.com/benchmarking/.

[2] Papadopoulos, D. P., J. R. R. Uijlings, F. Keller, and V. Ferrari. 2017. "Training Object Class Detectors with Click Supervision.". doi:10.1109/CVPR.2017.27.

[3] Salem, G. H., J. U. Dennis, J. Krynitsky, M. Garmendia-Cedillos, K. Swaroop, J. D. Malley, S. Pajevic, et al. 2015. "SCORHE: A Novel and Practical Approach to Video Monitoring of Laboratory Mice Housed in Vivarium Cage Racks." Behavior Research Methods 47 (1): 235-250. doi:10.3758/s13428-014-0451-5.

[4] Emeršič, Ž., V. Štruc, and P. Peer. 2017. "Ear Recognition: More than a Survey." Neurocomputing 255: 26-39. doi:10.1016/j.neucom.2016.08.139.

[5] Orekondy, T., B. Schiele, and M. Fritz. 2017. "Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images.". doi:10.1109/ICCV.2017.398.

[6] Pont-Tuset, J. and L. V. Gool. 2015. "Boosting Object Proposals: From Pascal to COCO.". doi:10.1109/ICCV.2015.181.

[7] Ros, G., L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. 2016. "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes."

[8] Zendel, Oliver, Markus Murschitz, Martin Humenberger and Wolfgang Herzner. “How Good Is My Test Data? Introducing Safety Analysis for Computer Vision.” International Journal of Computer Vision 125 (2017): 95-109.

All presented projects and related papers are also available on *CVOnline: Image Databases* website
http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm