Common Visual Data Foundation:
Enabling community-driven Research in Computer Vision

Tsung-Yi Lin
Cornell Tech, Cornell University
Common Visual Data Foundation (CVDF)

- President: Serge Belongie
- Vice President: Piotr Dollar
- Secretary: Tsung-Yi Lin
- Treasury: James Hays
Why CVDF?

- Enable community-driven research through **sharing data**

- **Why sharing data?**
  - new data enables new research direction
  - the progress of research community can be measured with the same metric
<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>AP5</th>
<th>AP10</th>
<th>AR5</th>
<th>AR10</th>
<th>AR50</th>
<th>AR100</th>
<th>AR20</th>
<th>AR500</th>
<th>AR1000</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>G_RMI</td>
<td>0.413</td>
<td>0.62</td>
<td>0.45</td>
<td>0.251</td>
<td>0.436</td>
<td>0.547</td>
<td>0.34</td>
<td>0.548</td>
<td>0.604</td>
<td>0.424</td>
<td>0.641</td>
<td>0.748</td>
<td>2016-09-18</td>
</tr>
<tr>
<td>umd_det</td>
<td>0.403</td>
<td>0.62</td>
<td>0.445</td>
<td>0.224</td>
<td>0.428</td>
<td>0.528</td>
<td>0.324</td>
<td>0.531</td>
<td>0.569</td>
<td>0.36</td>
<td>0.608</td>
<td>0.734</td>
<td>2017-06-10</td>
</tr>
<tr>
<td>MSRA</td>
<td>0.394</td>
<td>0.613</td>
<td>0.424</td>
<td>0.216</td>
<td>0.421</td>
<td>0.521</td>
<td>0.322</td>
<td>0.483</td>
<td>0.502</td>
<td>0.314</td>
<td>0.585</td>
<td>0.65</td>
<td>2016-11-24</td>
</tr>
<tr>
<td>Deformable R-FCN</td>
<td>0.373</td>
<td>0.58</td>
<td>0.407</td>
<td>0.165</td>
<td>0.402</td>
<td>0.518</td>
<td>0.317</td>
<td>0.473</td>
<td>0.483</td>
<td>0.277</td>
<td>0.541</td>
<td>0.673</td>
<td>2017-01-23</td>
</tr>
<tr>
<td>MSRA_2015</td>
<td>0.371</td>
<td>0.588</td>
<td>0.393</td>
<td>0.173</td>
<td>0.415</td>
<td>0.525</td>
<td>0.319</td>
<td>0.475</td>
<td>0.489</td>
<td>0.257</td>
<td>0.532</td>
<td>0.679</td>
<td>2015-12-26</td>
</tr>
<tr>
<td>Trimeps</td>
<td>0.359</td>
<td>0.59</td>
<td>0.383</td>
<td>0.158</td>
<td>0.407</td>
<td>0.505</td>
<td>0.312</td>
<td>0.470</td>
<td>0.407</td>
<td>0.260</td>
<td>0.557</td>
<td>0.683</td>
<td>2016-09-16</td>
</tr>
<tr>
<td>FPN (single model)</td>
<td>0.338</td>
<td>0.585</td>
<td>0.385</td>
<td>0.175</td>
<td>0.387</td>
<td>0.478</td>
<td>0.309</td>
<td>0.463</td>
<td>0.479</td>
<td>0.264</td>
<td>0.524</td>
<td>0.63</td>
<td>2016-11-04</td>
</tr>
<tr>
<td>Imagine-Lan</td>
<td>0.333</td>
<td>0.579</td>
<td>0.399</td>
<td>0.147</td>
<td>0.376</td>
<td>0.519</td>
<td>0.316</td>
<td>0.484</td>
<td>0.509</td>
<td>0.278</td>
<td>0.563</td>
<td>0.735</td>
<td>2016-12-24</td>
</tr>
<tr>
<td>Wal</td>
<td>0.345</td>
<td>0.559</td>
<td>0.369</td>
<td>0.165</td>
<td>0.389</td>
<td>0.478</td>
<td>0.303</td>
<td>0.47</td>
<td>0.493</td>
<td>0.256</td>
<td>0.565</td>
<td>0.658</td>
<td>2016-09-19</td>
</tr>
<tr>
<td>FAIRCNN</td>
<td>0.332</td>
<td>0.519</td>
<td>0.363</td>
<td>0.126</td>
<td>0.372</td>
<td>0.478</td>
<td>0.295</td>
<td>0.463</td>
<td>0.483</td>
<td>0.234</td>
<td>0.56</td>
<td>0.652</td>
<td>2015-11-20</td>
</tr>
<tr>
<td>CMU_AQ-VGG16</td>
<td>0.323</td>
<td>0.528</td>
<td>0.341</td>
<td>0.144</td>
<td>0.355</td>
<td>0.452</td>
<td>0.295</td>
<td>0.462</td>
<td>0.471</td>
<td>0.246</td>
<td>0.522</td>
<td>0.65</td>
<td>2016-09-19</td>
</tr>
</tbody>
</table>
First deep learning algorithm was applied
Mask R-CNN (SOTA Algorithm)

Image Classification

Box Detection

Instance Segmentation
Aren’t We Just Doing Fine?

- What we are doing well
  - the community is very open to share data when publishing!
- What are we not doing so well?
Academia

• Funding
• Long term viability

Industry

• Image copyright and license
• Take-down policy
What Are The Pain Points?

- **Where** to store pixels?

- What happens after students graduate?

- What if I need funding to expand my dataset?

- Could I get sued because of image copyright infringement?

- How should I design take-down policy?
Funding Model

- 501(c)(3) non-profit organization
- Funding from multiple sources
Case Study

- Common Objects in Context (COCO) Dataset
- COCO-stuff (University of Edinburgh)
- Open Images Dataset (Google)
COCO

• Goal: build the next object detection dataset

• Results:
  • annual competition
  • Interdisciplinary dataset
External Annotations on COCO and Related Datasets

SPEECH-COCO
SPEECH-COCO augments the COCO dataset with speech captions generated using TTS synthesis. The corpus contains more than 60,000 spoken captions allowing researchers of language acquisition, unsupervised term discovery, keyword spotting, or semantic embedding using speech and vision.

COCO-Stuff
COCO-Stuff augments the COCO dataset with pixel-level stuff annotations for 10,000 images. The JJ stuff classes are carefully selected to have a similar level of granularity to the thing classes in COCO, allowing the study of stuff and things in context.

VISUAL GENOME
Visual Genome is a dataset and knowledge base to connect structured image concepts to language.

RefCOCO
RefCOCO dataset was collected using Ref-It-Game. Each expression aims to unambiguously indicate a particular person or object in an image.

COCO Attributes
COCO Attributes has over 1.7M attribute annotations for People, Animals, and Objects from the COCO training dataset.

G-Ref
Google referring expression dataset (G-Ref) is a dataset focused on unambiguous object, text, and image descriptions (i.e. referring expressions) that allow us to uniquely identify a single object or region within an image.

VerSe
VerSe annotates COCO images with CoreNotes senses for 90 verbs (actions) which have ambiguous visual relations. Along with the sense information, we provide visual links for CoreNotes senses of 130 visual verbs.

COCO-Text
COCO-Text is for both text detection and recognition. The dataset annotates scene text with transcriptions along with attributes such as legibility, print style, thin or thick letter, and level.

FM-IQA
The First Multilingual Image Question Answering (FM-IQA) dataset contains over 120,000 images and 230,000 freestyle Chinese question-answer pairs and their English translations.

VQA
VQA is a new dataset containing open ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

VISUAL MADLIBLES
Visual Madlibs is a new dataset consisting of 240,000 natural language descriptions selected using automatically produced fill-in-the-blank prompts. This dataset can be used for targeted generation of multiple-choice testing materials.

COCO-a
COCO-a annotates human actions and interaction with objects (not people) with 130 visual actions. Verbs with an ambiguous visual connotation, along with information such as emotional state and relative distance and position with the object.

SALICON
The SALICON dataset offers a large set of saliency annotations on the COCO dataset. This data complements the task-specific annotations to advance the ultimate goal of visual understanding.
COCO

- How do we achieve that?
  - data sharing
  - challenging data with real world applications
  - generic annotation format to be adopted by other research groups
  - committed group members
COCO-stuff

- Initiated by the team at University of Edinburgh
- 10k images are annotated when first published
COCO-stuff

- + MightyAI

- COCO-stuff segmentation for COCO challenge this year

- $50k funding to collect annotations for 200k images
Open Images Dataset

- Large scale dataset: 8M images, only urls were released
- CVDF helps hosting the actual pixels
- CVDF is responsible for copyright issue of hosted images
How We Would Like to Work With You

• Hosting pixels/data
• Providing funding
• Advice on data collection and annotations
• Advice on organizing competitions