CRCV HSAP 2020
WEEK 11 PRESENTATION

EMILY PARK

November 17rd – December 1st

OVERVIEW

- New data collection on ground/aerial videos
- Research paper reading

DATA COLLECTION FOR KRISHNA

- **Ground Videos:** 150 new *(~411 total ground)*
- **Aerial Videos:** 150 *(~409 total aerial)*
- **Total:** 300 new videos *(~820 total videos collected)*

+13 cities: (1) Adelaide, (2) Agra, (3) Ahmedabad, (4) Aleppo, (5) Anman, (6) Ankara, (7) Donetsk, (8) Dubai, (9) Durban, (10) Edmonton, (11) Genova, (12) Glasgow, (13) Guatemala City

RESEARCH PAPER READING "BDD100K"

- Benchmarks are comprised of **10 tasks**: image tagging, lane detection, drivable area segmentation, road object detection, semantic segmentation, instance segmentation, multi-object detection tracking, multi-object segmentation tracking, domain adaptation, and imitation learning
- Conducted extensive evaluations of existing algorithms on the new benchmarks
- BDD 100K
  - Achieves good diversity by obtaining videos in a crowd-sourcing manner uploaded by 10,000+ drivers,
    - Contains high resolution images (720p) and high frame rate (30 fps), and also GPS/IMU recordings to preserve the driving trajectories
    - City streets, residential areas, and highways
    - The videos are split into training (70K), validation (10K), and testing (20K) sets.

RESEARCH PAPER READING CONT.

- **Image Tagging**
  - Collected image-level annotation on 6 weather conditions, 6 scene types, and 3 distinct times of day for each image
    - Domain transfer = train in one area (on cats, one city, etc.) - trying to collect all possible domains
    - Dataset contains approx. an equal number of day-time and night-time videos
- **Object Detection**
- Bounding box annotations of 10 categories of the reference frames of 100K videos
  - We provide visibility attributes including "occluded" and "truncated"
- **Lane Marking**
- Lane marking detection = critical for vision-based vehicle localization and trajectory planning
- Our lane marking are labeled with 8 main categories: road curb, crosswalk, double white, double yellow, double other color, single white, single yellow, single other color. We label these attributes of continuity (full or dashed) and direction (parallel or perpendicular)

RESEARCH PAPER READING CONT.

- **Drivable Area**
  - **1.) Directly drivable area** = what the driver is currently on & the region where the driver has priority over other cars (or the right of the way)
  - **2.) Alternatively drivable areas** = visually distinguishable, functionally different, and require algorithms to recognize blocking objects and scene context
- **Semantic Instance Segmentation**
  - Fine-grained, pixel-level annotations for images from each of the 10,000 video clips randomly samples from the whole dataset; Each pixel is given a label and a corresponding identifier of the instance # of that object label in the image
- **Multiple Object Tracking (MOT)**
  - MOT dataset includes 2,000 videos with about 400K frames
    - Annotated at 5 frps = ~200 frames per video; 130.gK track identities 3.3M bounding boxes in the training and validation set
    - An object may be fully occluded or moved out of the frame, and then reappear later

RESEARCH PAPER READING CONT.

- **Imitation Learning**
  - GPS/IMU recordings show human driver action given the visual input and the driving trajectories
  - Use these recordings as a demonstration supervision for the imitation learning algorithms and use perplexity (look at the reference 33 to define perplexity and put it in your presentation) to measure the similarity of driving behaviors on the validation and testing set
- **Diversity**
  - Conducts 2 sets of experiments on object detection and semantic segmentation
    - 1.) Object detection = study the different domains within the dataset
    - 2.) Semantic segmentation = investigate the domains between our data
- **Object Detection**
  - We then train Faster-RCNN (used for object detection - click on reference 28, get a pic of the model architecture that come up, and do a single slide that says Faster-RCNN and shows the architecture - I didn't have chance to read the paper thoroughly because it was in the reference) based on ResNet-50 (used for image classification - has residuals inside the model) on those domains and evaluate the result with COCO API Annotated at 5 frps = ~200 frames per video
  - The difference between city and non-city is significant, but the gap between daytime and nighttime is much bigger

THANK YOU!

ANY QUESTIONS?