BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning

Fisher Yu[1] Haofeng Chen[1] Xin Wang[1] Wenqi Xian[2*] Yingying Chen[1]
Fangchen Liu[3*] Vashisht Madhavan[4*] Trevor Darrell[1]

[1]UC Berkeley [2]Cornell University [3]UC San Diego [4]Element, Inc.

Abstract

Datasets drive vision progress, yet existing driving datasets are impoverished in terms of visual content and supported tasks to study multitask learning for autonomous driving. Researchers are usually constrained to study a small set of problems on one dataset, while real-world computer vision applications require performing tasks of various complexities. We construct BDD100K [1], the largest driving video dataset with 100K videos and 10 tasks to evaluate the exciting progress of image recognition algorithms on autonomous driving. The dataset possesses geographic, environmental, and weather diversity, which is useful for training models that are less likely to be surprised by new conditions. Based on this diverse dataset, we build a benchmark for heterogeneous multitask learning and study how to solve the tasks together. Our experiments show that special training strategies are needed for existing models to perform such heterogeneous tasks. BDD100K opens the door for future studies in this important venue.

1. Introduction

Diverse, large-scale annotated visual datasets, such as ImageNet [8] and COCO [18], have been the driving force behind recent advances in supervised learning tasks in computer vision. Typical deep learning models can require millions of training examples to achieve state-of-the-art performance for a task [16, 27, 15].

For autonomous driving applications, however, leveraging the power of deep learning is not as simple due to the lack of comprehensive datasets. Existing datasets for autonomous driving [14, 7, 23] are limited in one or more significant aspects, including the scene variation, the richness of annotations, and the geographic distribution. Additionally, models trained on existing datasets tend to overfit specific domain characteristics [25].

Real-world applications require performing a combination of perception tasks with different complexities, instead of only *homogeneous* multiple tasks with the same prediction structure [26, 37, 1, 20]. Although it may be feasible to label a large number of images with simple annotations such as drivable areas and object bounding boxes [11, 18], it remains challenging to obtain more complicated annotations such as instance segmentation [3], not to mention multi-object detection and segmentation tracking [30, 21]. As a result, even though a considerable amount of effort has been put into constructing large-scale visual datasets, research on those complicated tasks is still limited to small datasets [7, 14]. In production environments, it is also unclear how to allocate resources for various annotations to support the applications requiring *heterogeneous* tasks with various output structures.

We aim to facilitate algorithmic study on large-scale diverse visual data and multiple tasks. We build BDD100K, a new, diverse, and large-scale dataset of visual driving scenes, together with various tasks, to overcome the limitations. We have been able to collect and annotate the largest available dataset of annotated driving scenes, consisting of over 100K diverse video clips. BDD100K covers more realistic driving scenarios and captures more of the "long-tail" of appearance variation and pose configuration of categories of interest in diverse environmental domains. Our benchmarks are comprised of ten 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, as shown in Figure 1. These diverse tasks make the study of heterogeneous multitask learning possible. In our benchmarks, the models can perform a series of tasks with increasing complexities.

We conduct extensive evaluations of existing algorithms on our new benchmarks. Special attention is paid to multitask learning in homogeneous, cascaded, and heterogeneous settings. Our experiments present many new findings, made possible by the diverse set of tasks on a single dataset. Our benchmark models on heterogeneous multitask learning shed light on the challenges of designing one single model to support multiple tasks.

*Work done at UC Berkeley.

[1]The data is available at https://bdd-data.berkeley.edu

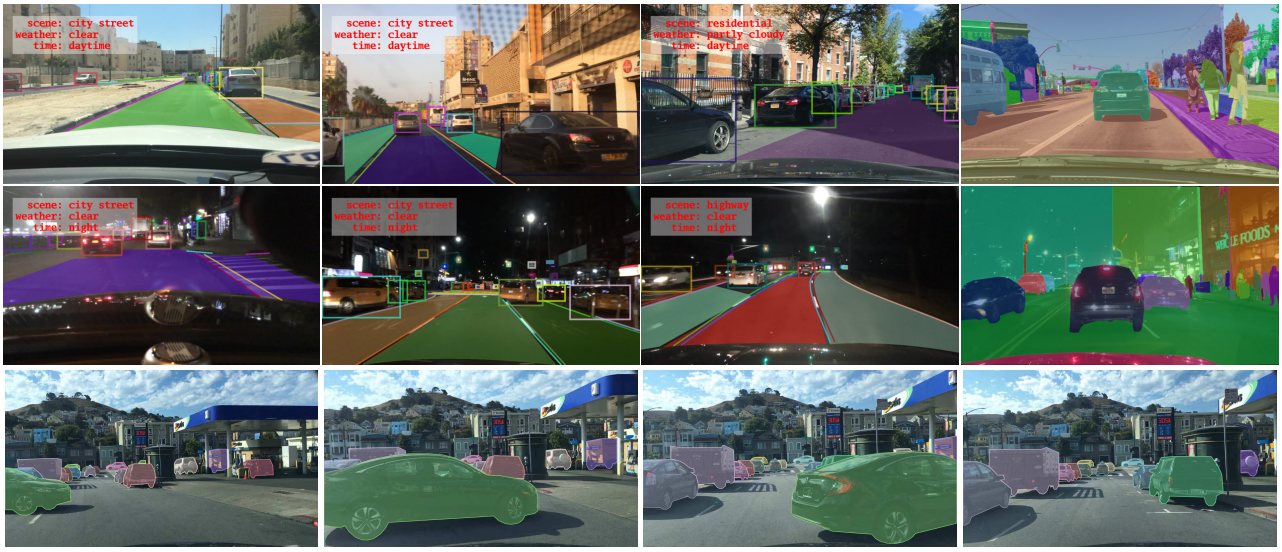


Figure 1: Overview of our dataset. Our dataset includes a diverse set of driving videos under various weather conditions, time, and scene types. The dataset also comes with a rich set of annotations: scene tagging, object bounding box, lane marking, drivable area, full-frame semantic and instance segmentation, multiple object tracking, and multiple object tracking with segmentation.

The major contributions of our paper are: 1) a comprehensive diverse 100K driving video dataset supporting tasks of multiple complexities, which can serve as an evaluation benchmark for computer vision research for autonomous driving; 2) a benchmark for heterogeneous multitask learning and baseline studies to facilitate future study.

2. Related Works

Visual datasets are necessary for numerous recognition tasks in computer vision. Especially with the advent of deep learning methods, large scale visual datasets, such as [8, 35, 39, 23], are essential for learning high-level image representations. They are general-purpose and include millions of images with image-level categorical labels. These large datasets are useful in learning representations for image recognition, but most of the complex visual understanding tasks in the real world require more fine-grained recognition such as object localization and segmentation [11]. Our proposed dataset provides these multi-granularity annotations for more in-depth visual reasoning. In addition, we provide these annotations in the context of videos, which provides an additional dimension of visual information. Although large video datasets exist [5, 2, 28], they usually are restricted to image-level labels.

Driving datasets have received increasing attention in the recent years, due to the popularity of autonomous vehicle technology. The goal is to understand the challenge of computer vision systems in the context of self-driving. Some of the datasets focus on particular objects such as pedestrians [9, 38]. Cityscapes [7] provides instance-level semantic segmentation on sampled frames of videos collected by their own vehicle. RobotCar [19] and KITTI [14] also provide data of multiple sources such as LiDAR scanned points. Because it is very difficult to collect data that covers a broad range of time and location, the data diversity of these datasets is limited. For a vehicle perception system to be robust, it needs to learn from a variety of road conditions in numerous cities. Our data was collected from the same original source as the videos in [32]. However, the primary contribution of our paper is the video annotations with benchmarks on heterogeneous tasks. Mapillary Vistas [23] provides fine-grained annotations for user uploaded data, which is much more diverse with respect to location. However, these images are one-off frames that are not placed in the context of videos with temporal structure. Like Vistas, our data is crowdsourced, however, our dataset is collected solely from drivers, with each annotated image corresponding to a video sequence, which enables interesting applications for modeling temporal dynamics.

Multitask Learning aims to improve generalization of a certain task by learning from other tasks [6, 22]. It has been widely studied in machine learning [6, 12]. The growing interests in learning the relationship between tasks gives rise to a number of multitask and transfer learning training benchmarks and challenges. Robust Vision Challenge [1] features six vision challenges, where a single model is expected to produce results on multiple vision tasks. Zamir *et al*. [37] investigate the dependency structure among twenty-six visual tasks by transfer learning. McCann *et al*. [20] present a challenge with ten natural language processing tasks, and proposes a model that solves all by formulating each task as question answering. Similar to McCann *et al*. [20], existing multitask and transfer learning setups are homogeneous in output structures. The tasks can be

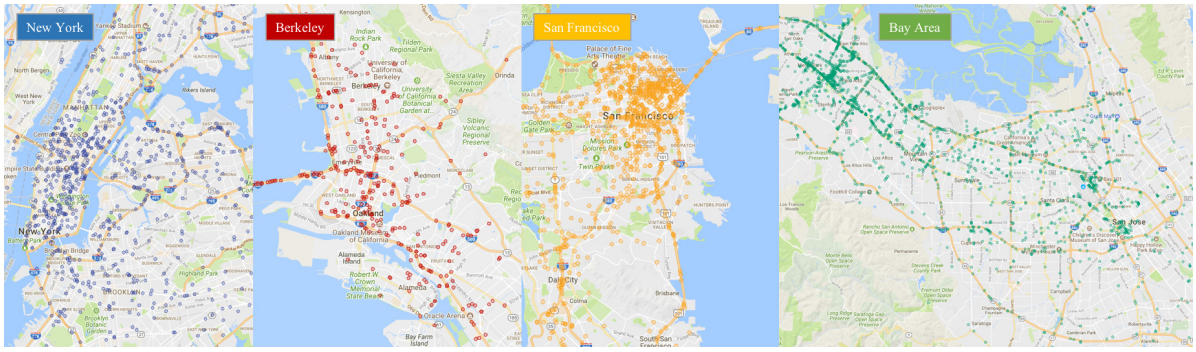
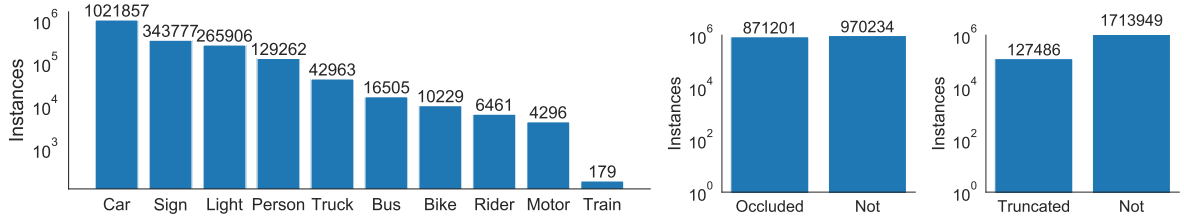


Figure 2: Geographical distribution of our data sources. Each dot represents the starting location of every video clip. Our videos are from many cities and regions in the populous areas in the US.



(a) Number of instances in each category. (b) Occlusion (c) Truncation

Figure 3: Instance statistics of our object categories. (a) Number of instances of each category, which follows a long-tail distribution. (b) Roughly half of the instances are occluded. (c) About 7% of the instances are truncated.

formulated as pixel-level or low-dimensional classification and regression. BDD100K contains multiple tasks including pixel-level, region-based, and temporally aware tasks, opening the door for heterogeneous multitask learning.

3. BDD100K

We aim to provide a large-scale diverse driving video dataset with comprehensive annotations that can expose the challenges of street-scene understanding. To achieve good diversity, we obtain our videos in a crowd-sourcing manner uploaded by tens of thousands of drivers, supported by Nexar [2]. The dataset contains not only images with high resolution (720p) and high frame rate (30fps), but also GPS/IMU recordings to preserve the driving trajectories. In total, we have 100K driving videos (40 seconds each) collected from more than 50K rides, covering New York, San Francisco Bay Area, and other regions as shown in Figure 2.

The dataset contains diverse scene types such as city streets, residential areas, and highways. Furthermore, the videos were recorded in diverse weather conditions at different times of the day. The videos are split into training (70K), validation (10K) and testing (20K) sets. The frame at the 10th second in each video is annotated for image tasks and the entire sequences are used for tracking tasks.

3.1. Image Tagging

We have collected image-level annotation on six weather conditions, six scene types, and three distinct times of day, for each image. The videos contain large portions of extreme weather conditions, such as snow and rain. They also include a diverse number of different scenes across the world. Notably, our dataset contains approximately an equal number of day-time and night-time videos. Such diversity allows us to study domain transfer and generalize our object detection model well on new test sets. Detailed distributions of images with weather, scene, and day hours tags are shown in the supplementary materials. We provide image tagging classification results using DLA-34 [36] in Figure 4. The average classification accuracy across different weather and scenes are around 50 to 60%.

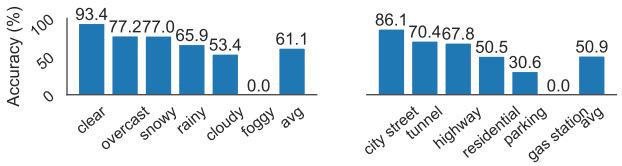


Figure 4: Image tagging classification results using DLA-34.

3.2. Object Detection

Locating objects is a fundamental task for not only autonomous driving but the general visual recognition. We provide bounding box annotations of 10 categories for each of the reference frames of 100K videos. The instance statistics is shown in Figure 3a. We provide visibility attributes including "occluded" and "truncated" in Figure 3b and Figure 3c.

[2] https://www.getnexar.com

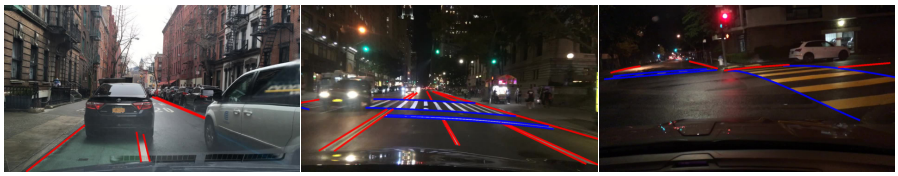


Figure 5: Examples of lane marking annotations. Red lanes are vertical and blue lanes are parallel. Left: we label all the visible lane boundaries. Middle: not all marking edges are lanes for vehicles to follow, such as pedestrian crossing. Right: parallel lanes can also be along the current driving direction.

Figure 6: Examples of drivable areas. Red regions are directly drivable and the blue ones are alternative. Although drivable areas can be confined within lane markings, they are also related to locations of other vehicles shown in the right two columns.

3.3. Lane Marking

The lane marking detection is critical for vision-based vehicle localization and trajectory planning. However, available datasets are often limited in scale and diversity. For example, the Caltech Lanes Dataset [4] only contains 1,224 images, and the Road Marking Dataset [31] has 1,443 images labeled in 11 classes of lane markings. The most recent work, VPGNet [17], consists of about 20,000 images taken during three weeks of driving in Seoul.

Our lane markings (Figure 5) are labeled with 8 main categories: road curb, crosswalk, double white, double yellow, double other color, single white, single yellow, single other color. The *other* categories are ignored during evaluation. We label the attributes of continuity (full or dashed) and direction (parallel or perpendicular). Shown in Table 1, our lane marking annotations cover a diverse set of classes. Detailed distributions of types of lane markings and drivable areas are shown in the supplementary materials.

Datasets	Training	Total	Sequences
Caltech Lanes Dataset [4]	-	1,224	4
Road Marking Dataset [31]	-	1,443	29
KITTI-ROAD [13]	289	579	-
VPGNet [17]	14,783	21,097	-
BDD100K	70,000	100,000	100,000

Table 1: Lane marking statistics. Our lane marking annotations are significantly richer and are more diverse.

3.4. Drivable Area

Lanes alone are not sufficient to decide road affordability for driving. Although most of the time, the vehicle should stay between the lanes, it is common that no clear lane marking exists. In addition, the road area is shared with all other vehicles, but a lane can not be driven on if occupied. All these conditions beyond lane markings direct our driving decisions and thus are relevant for designing autonomous driving algorithms.

Our drivable areas are divided into two different categories: *directly drivable area* and *alternatively drivable area*. The directly drivable area is what the driver is currently driving on – it is also the region where the driver has priority over other cars or the right of the way. In contrast, alternatively drivable area is a lane the driver is currently not driving on, but able to do so via changing lanes. Although the directly and alternatively drivable areas are visually indistinguishable, they are functionally different, and require the algorithms to recognize blocking objects and scene context. Some examples are shown in Figure 6. The distribution of drivable region annotations is shown in the supplementary materials. Not surprisingly, on highways or city streets, where traffic is closely regulated, drivable areas are mostly within lanes and they do not overlap with the vehicles or objects on the road. However, in residential areas, the lanes are sparse. Our annotators can find the drivable areas based on the surroundings.

3.5. Semantic Instance Segmentation

We provide fine-grained, pixel-level annotations for images from each of the 10,000 video clips randomly sampled from the whole dataset. Each pixel is given a label and a corresponding identifier denoting the instance number of that object label in the image. Since many classes (e.g., sky) are not amenable to being split into instances, only a small subset of class labels are assigned instance identifiers. The entire label set consists of 40 object classes, which are chosen to capture the diversity of objects in road scenes as well as maximizing the number of labeled pixels in each image. Besides a large number of labels, our dataset exceeds previous efforts in terms of scene diversity and complexity. The

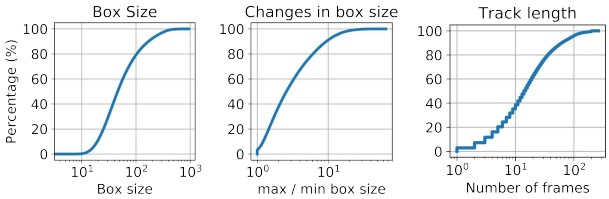


Figure 7: Cumulative distributions of the box size (left), the ratio between the max and min box size for each track (middle) and track length (right). Our dataset is more diverse in object scale.

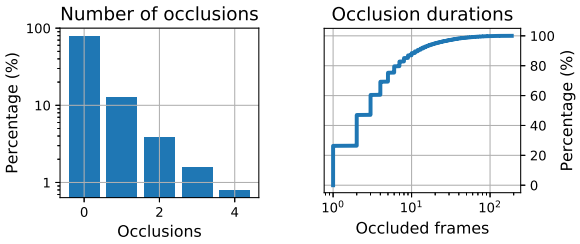


Figure 8: Number of occlusions by track (left) and number of occluded frames for each occlusion (right). Our dataset covers complicated occlusion and reappearing patterns.

whole set is split into 3 parts: 7K images for training, 1K images for validation, and 2K images for testing. The distribution of classes in the semantic instance segmentation dataset is shown in the supplementary materials.

3.6. Multiple Object Tracking

To understand the temporal association of objects within the videos, we provide a multiple object tracking (MOT) dataset including 2,000 videos with about 400K frames. Each video is approximately 40 seconds and annotated at 5 fps, resulting in approximately 200 frames per video. We observe a total number of 130.6K track identities and 3.3M bounding boxes in the training and validation set. The dataset splits are 1400 videos for training, 200 videos for validation and 400 videos for testing. Table 2 shows the comparison of BDD100K with previous MOT datasets. Our tracking benchmark provides one order-of-magnitude bigger than the previously popular tracking dataset, MOT17 [21]. A recent dataset released by Waymo [29] has fewer tracking sequences (1150 vs 2000) and fewer frames (230K vs 398K) in total, compared to ours. But Waymo data has more 2D boxes (9.9M vs 4.2M), while ours has better diversity including different weather conditions and more locations. Distributions of tracks and bounding boxes by category are shown in the supplementary materials.

Datasets	Frames	Sequences	Identities	Boxes
KITTI [14]	8K	21	917	47K
MOT17 [21]	34K	21	1,638	337K
BDD100K	318K	1,600	131K	3.3M

Table 2: MOT datasets statistics of training and validation sets. Our dataset has more sequences, frames, identities as well as more box annotations.

BDD100K MOT is diverse in object scale. Figure 7 (left) plots the cumulative distribution of box size, defined as $\sqrt{wh}$ for a bounding box with width w and height h. Figure 7 (middle) shows the cumulative distribution of the ratio between the maximum box size and the minimum box size along each track, and Figure 7 (right) shows that of the length of each track. The distributions show that the MOT dataset is not only diverse in visual scale among and within tracks, but also in the temporal range of each track.

Objects in our tracking data also present complicated occlusion and reappearing patterns are shown in Figure 8. An object may be fully occluded or move out of the frame, and then reappear later. We observe 49,418 occurrences of occlusion in the dataset, or one occurrence of occlusion every 3.51 tracks. Our dataset shows the real challenges of object re-identification for tracking in autonomous driving.

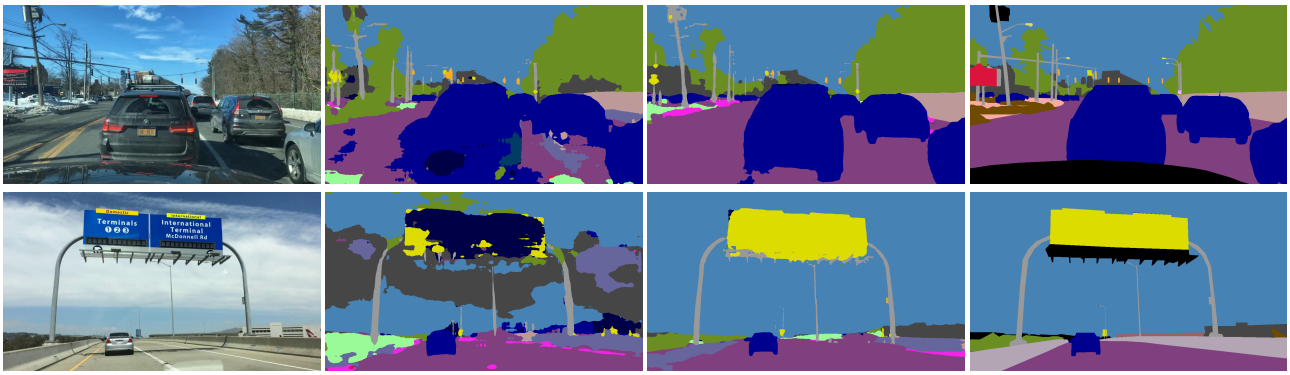
3.7. Multiple Object Tracking and Segmentation

We further provide a multiple object tracking and segmentation (MOTS) dataset with 90 videos. We split the dataset into 60 training videos, 10 validation videos, and 20 testing videos.

Datasets	Frames	Seq.	Identities	Ann.	Ann. / Fr.
KITTI MOTS [30]	8K	21	749	38K	4.78
MOTS Challenge [30]	2.9K	4	228	27K	9.40
DAVIS 2017 [24]	6.2K	90	197	-	-
YouTube VOS [33]	120K	4.5K	7.8K	197K	1.64
BDD100K MOTS	14K	70	6.3K	129K	9.20

Table 3: Comparisons with other MOTS and VOS datasets.

Table 3 shows the details of the BDD MOTS dataset and the comparison with existing multiple object tracking and segmentation (MOTS) and video object segmentation (VOS) datasets. MOTS aims to perform segmentation and tracking of multiple objects in crowded scenes. Therefore, MOTS datasets like KITTI MOTS and MOTS Challenge [30] require denser annotations per frame and therefore are smaller in size than VOS datasets. BDD100K MOTS provides a MOTS dataset that is larger than the KITTI and MOTS Challenge datasets, with the number of annotations comparable with the large-scale YouTube VOS [33] dataset. Detailed distributions of the MOTS dataset by category are shown in the supplementary materials.



(a) Our Image (b) Trained on Cityscapes (c) Trained on Ours (d) Ground Truth

Figure 9: Visual comparisons of the same model (DRN [34]) trained on different datasets. We find that there is a dramatic domain shift between Cityscapes and our new dataset. For example, due to infrastructure difference, the model trained on Cityscapes is confused by some simple categories such as sky and traffic signs.

Train \ Test	City	Non-City	Val
City-30K	29.5	26.5	28.8
Non-City-30K	24.9	24.3	24.9
Random-30K	28.7	26.6	28.3

Train \ Test	Daytime	Non-Daytime	Val
Daytime-30K	30.6	23.6	28.1
Non-Daytime-30K	25.9	25.3	25.6
Random-30K	29.5	26.0	28.3

Table 4: Domain discrepancy experiments with object detection. We take the images from one domain and report testing results in AP on the same domain or the opposite domain. We can observe significant domain discrepancies, especially between daytime and nighttime.

3.8. Imitation Learning

GPS/IMU recordings in our dataset show the human driver action given the visual input and the driving trajectories. We can use those recordings as a demonstration supervision for the imitation learning algorithms and use perplexity to measure the similarity of driving behaviors on the validation and testing set. We refer to Xu *et al*. [32] for details on the evaluation protocols. Visualizations of the driving trajectories are shown in the supplementary materials.

4. Diversity

One distinct feature of our data is diversity, besides video and scale. We can study new challenges that the diversity brings to existing algorithms and how our data complements existing datasets. We conduct two sets of experiments on object detection and semantic segmentation. In object detection experiments, we study the different domains within our dataset. While in semantic segmentation, we investigate the domains between our data and Cityscapes [7].

4.1. Object Detection

Our dataset has an advantage in diversity, compared to other popular driving datasets. We investigate the influence of domain differences on object detection. The whole dataset is partitioned into several domains based on time of day and scene types. City street and daytime are chosen as validation domains. The training sets have the same number of images (30K) in the training set. We then train Faster-RCNN [27] based on ResNet-50 on those domains and evaluate the result with COCO API [18].

We find that there is indeed a domain discrepancy between image sets from different conditions, as shown in Table 4. The difference between city and non-city is significant, but the gap between daytime and nighttime is much bigger. Although this is not completely surprising, the results indicate that more work is necessary to bridge the gap.

4.2. Semantic Segmentation

We also compare the models trained on Cityscapes and ours, to understand the difference between our new datasets and existing driving datasets. Cityscapes data is collected in German cities, while our data is mainly from the US. We observe that there is a dramatic domain shift between the two datasets for semantic segmentation models. The models perform much worse when tested on a different dataset. This suggests that even for the domain of other datasets, our new dataset is complementary, which augments existing datasets. Figure 9 shows the discrepancy visually. We can observe that the model trained on Cityscape can not recognize the traffic sign in the US.

5. Multitask Learning

BDD100K gives the opportunity to study joint solution for the heterogeneous tasks. In this section, we investigate the effects of modeling various tasks jointly with the same base model. We study how to utilize diversity and quantity

of simple labels to improve the accuracy of the complicated tasks, such as from object detection to tracking.

5.1. Homogeneous Multitask Learning

We first investigate the effects of jointly performing tasks with similar output structures. The BDD100K lane marking and drivable area datasets share the same set of 70K training images. Drivable area annotations consist of 2 foreground classes and lane marking annotations have 3 attributes (direction, continuity, and category). We formulate the detection of drivable area as segmentation and lane marking as contour detection. We evaluate drivable area segmentation by mean IoU, and lane marking by the Optimal Dataset Scale F-measure (ODS-F) for each category of the three attributes using the Structured Edge Detection Toolbox [10] with tolerance $\tau = 1$, 2, and 10 pixels. We employ morphological thinning for each score threshold during evaluation.

We employ DLA-34 [36] as the base model for the segmentation tasks. We implement the segmentation head with four 3×3 convolution blocks followed by an 1×1 convolution to produce segmentation maps in a 4x down-sampled scale, and use bilinear interpolation to upsample the output to the original scale. For lane marking, we use three segmentation heads for the three attributes. We employ the weighted cross-entropy loss with foreground weight 10 for the lane marking heads, and the gradient-based non-maximum suppression for post-processing. We construct three train sets with 10K, 20K and the full 70K images and report the evaluation results of models trained on individual tasks and both tasks in Table 5. Full evaluation results for lane marking are shown in the supplementary materials.

Training Set	Lane ODS-F ($\tau = 10$) (%)				Drivable IoU (%)		
	dir.	cont.	cat.	mean	direct	altern.	mean
Lane 10K	49.29	47.85	39.08	45.41	-	-	-
Drive 10K	-	-	-	-	73.10	55.36	64.23
Lane+Drive 10K	53.97	52.59	44.65	50.40	74.69	54.06	64.37
Lane 20K	57.36	55.85	49.88	54.36	-	-	-
Drive 20K	-	-	-	-	79.00	63.27	71.13
Lane+Drive 20K	57.19	55.64	49.50	54.11	79.39	64.06	71.73
Lane 70K	57.50	55.87	50.08	54.48	-	-	-
Drive 70K	-	-	-	-	79.40	63.33	71.37
Lane+Drive 70K	57.35	55.76	49.63	54.24	79.72	64.70	72.21

Table 5: Evaluation results of homogeneous multitask learning on lane marking and drivable area segmentation. We train lane marking, drivable area segmentation and the joint training of both on training splits with 10K, 20K, and the full 70K images.

We observe that when training with only 10K images, the mean ODS-F score of lane marking prediction improves from 45.41 to 50.40 when jointly training with the drivable area task. However, the improvement of jointly training on the drivable area detection task, from 64.23 to 64.37, is marginal compared to the individual task. As we increase the number of training images to 20K and 70K, the difference between jointly training and single-task training becomes insignificant, though the performance numbers are generally higher than those trained on 10K images.

One hypothesis for the results is that the drivable area detection task and the lane marking task share a similar prediction structure, referred as the homogeneous tasks, and therefore the additional supervision may fail to bring new information to each individual task. These results further motivate us to study multitask learning of heterogeneous tasks with diverse prediction structure and annotation types in this work.

5.2. Cascaded Multitask Learning

Certain tasks such as object tracking and instance segmentation are more time-consuming to annotate. But they can depend on predictions of simple tasks. This connection has been studied as cascaded multitask learning. For example, more accurate object detection can locate the object candidates better for tracking. A natural question is whether to spend all the annotation efforts for the complicated tasks, or to allocate some resources for the basic tasks.

Training Set	AP	AP_{50}	AP_{75}
Inst-Seg	21.8	40.5	20.5
Inst-Seg + Det	**24.5**	**45.4**	**21.6**

Table 6: Evaluation results for instance segmentation when joint training with the object detection set. Additional localization supervision can improve instance segmentation significantly.

Training Set	AP	MOTA	MOTP	IDS
MOT	28.1	55.0	84.0	**8386**
MOT + Det	**30.7**	**56.7**	**84.1**	9098

Table 7: Evaluation results for multiple object tracking cascaded with object detection. AP is the detection metric. Even though the tracking set has much more boxes, the model can still benefit from the diverse instance examples in the detection set.

Object detection and instance segmentation. The BDD instance segmentation dataset contains 7K images, whereas the detection dataset has 70K images. We first study whether adding more object detection annotations can help instance segmentation. We use Mask R-CNN [15] with ResNet-50 [16] as the backbone, and train detection and instance segmentation in a batch-level round-robin manner. As shown in Table 6, AP increases from 21.8 to 24.5 with joint training. The instance segmentation model is able to learn better object appearance features and localization from the detection set with a much richer diversity of images and object examples. Zhou *et al*. [40] explore the shape priors in the detection supervision and improve the semi-supervised instance segmentation results further.

Training Set	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mean IoU
Sem-Seg	94.3	63.0	84.9	**25.7**	**45.8**	**52.6**	56.2	54.1	**86.4**	45.1	95.3	62.4	22.1	90.2	50.5	68.3	0	35.5	**49.9**	56.9
Sem-Seg + Det	94.3	62.5	**85.2**	24.5	41.1	51.5	**63.1**	**57.9**	86.2	**47.4**	**95.5**	**64.6**	**28.1**	**90.8**	**52.9**	**70.7**	0	**43.4**	48.9	**58.3**
Sem-Seg + Lane + Driv	**94.8**	**65.8**	84.1	22.6	40.2	49.3	51.9	49.7	85.8	46.2	95.3	60.8	7.1	89.9	47.8	66.9	0	27.5	27.5	53.3

Table 8: Evaluation results for semantic segmentation. We explore segmentation joint-training with different tasks. Detection can improve the overall accuracy of segmentation, although their output structures are different. However, although Lane and Drivable area improve the segmentation of road and sidewalk, the overall accuracy drops.

MOT and object detection. BDD100K MOT has 278K training frames from 1,400 videos, whereas the detection set contains 70K images sampled from 70K videos. For the detection and MOT models, we use a modified version of Faster R-CNN [27] with a shared DLA-34 [36] backbone. The implementation details of the tracking model are shown in the supplementary materials. Table 7 shows that joint training of detection and multiple object tracking improves the single-task MOT model with detection AP increasing from 28.1 to 30.7 and MOTA from 55.0 to 56.7, with a slight increase in identity switch.

Semantic segmentation with other tasks. Following a similar manner, we fine-tune a base semantic segmentation model by jointly training semantic segmentation with detection and lane marking/drivable area as shown in Table 8. We observe that training with the additional 70K object detection dataset improves the overall mIoU from 56.9 to 58.3, with the improvement mostly attributed to the object classes that are present in the object detection dataset. When jointly training with the lane marking and drivable area sets, the IOU of the stuff classes (e.g., road and sidewalk) improves though the overall IOU across all classes decreases.

To summarize, adding more annotations to the simple tasks in the task cascade can help improve the performances of the complicated tasks that require more expensive labels.

5.3. Heterogeneous Multitask Learning

The ultimate goal of our benchmark is to study how to perform all the heterogeneous tasks together for autonomous driving. To understand the potential and difficulty, we study joint training for multiple object tracking and segmentation, a downstream task to object detection, instance segmentation, and multiple object tracking. Since the MOTS dataset requires time-consuming instance segmentation annotations at each frame, the dataset is relatively limited in video diversity, with 12K frames from 60 videos in the training set. We aim to improve the performance on the task of MOTS by leveraging the diversity from the detection set with 70K images from 70K videos, the MOT set with 278K frames from 1,400 videos, and the instance segmentation set with 7K images from 7K videos.

We report instance segmentation AP and multi-object tracking and segmentation accuracy (MOTSA), precision (MOTSP), and other metrics used by [30] in Table 9. We first fine-tune the MOTS model from pre-trained models of upstream tasks. Compared with training MOTS from scratch, fine-tuning from the pre-trained instance segmentation model improves segmentation AP and MOTSP. Fine-tuning from the pre-trained MOT model, on the other hand, reduces identity switch (IDSW). The extra training examples from the instance segmentation and MOT datasets improve the segmentation and box propagation respectively, thus improving the overall MOTSA results by a large margin. We finally fine-tune the jointly trained detection and tracking model mentioned in Table 7 by jointly training the four tasks together. We achieve an overall segmentation AP of 23.3 and MOTSA of 41.4.

Training Set	AP	MOTSA	MOTSP	FN	FP	IDSW
MOTS (S)	13.0	30.4	81.8	8352	**5116**	566
InstSeg (I) + MOTS	18.7	33.7	**81.9**	6810	5611	965
MOT (T) + MOTS	19.7	40.3	79.8	5698	5967	**390**
Det + T + I + S	**23.3**	**41.4**	81.6	**5132**	6228	472

Table 9: MOTS evaluation results. Both instance segmentation AP and MOTS evaluation metrics are reported. Instance segmentation tracking is very hard to label, but we are able to use object detection, tracking, and instance segmentation to improve segmentation tracking accuracy significantly.

6. Conclusion

In this work, we presented BDD100K, a large-scale driving video dataset with extensive annotations for heterogeneous tasks. We built a benchmark for heterogeneous multitask learning where the tasks have various prediction structures and serve different aspects of a complete driving system. Our experiments provided extensive analysis to different multitask learning scenarios: homogeneous multitask learning and cascaded multitask learning. The results presented interesting findings about allocating the annotation budgets in multitask learning. We hope our work can foster future studies on heterogeneous multitask learning and shed light on this important direction.

References

[1] Robust Vision Challenge. http://www.robustvision.net/. 1, 2

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[3] D. Acuna, H. Ling, A. Kar, and S. Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. 2018. 1

[4] M. Aly. Real time detection of lane markers in urban streets. In *Intelligent Vehicles Symposium*, pages 7–12, 2008. 4

[5] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2

[6] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 2

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1, 2, 6

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1, 2

[9] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009. 2

[10] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013. 7

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 2

[12] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004. 2

[13] J. Fritsch, T. Kuhnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, pages 1693–1700. IEEE, 2013. 4

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 5

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 7

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 7

[17] S. Lee, J. Kim, J. S. Yoon, S. Shin, O. Bailo, N. Kim, T.-H. Lee, H. S. Hong, S.-H. Han, and I. S. Kweon. VPGNet: Vanishing point guided network for lane and road marking detection and recognition. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1965–1973. IEEE, 2017. 4

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 6

[19] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The oxford robotcar dataset. *IJ Robotics Res.*, 36(1):3–15, 2017. 2

[20] B. McCann, N. S. Keskar, C. Xiong, and R. Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018. 1, 2

[21] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 5

[22] T. M. Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research , 1980. 2

[23] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[24] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[25] S.-A. Rebuffi, H. Bilen, and A. Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pages 506–516, 2017. 1

[26] S.-A. Rebuffi, H. Bilen, and A. Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pages 506–516, 2017. 1

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 6, 8

[28] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[29] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. *arXiv*, pages arXiv–1912, 2019. 5

[30] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe. Mots: Multi-object tracking

and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019. 1, 5, 8

[31] T. Wu and A. Ranganathan. A practical system for road marking detection and recognition. In *Intelligent Vehicles Symposium*, pages 25–30, 2012. 4

[32] H. Xu, Y. Gao, F. Yu, and T. Darrell. End-to-end learning of driving models from large-scale video datasets. *arXiv preprint*, 2017. 2, 6

[33] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 5

[34] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[35] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2

[36] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018. 3, 7, 8

[37] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[38] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 2

[40] Y. Zhou, X. Wang, J. Jiao, T. Darrell, and F. Yu. Learning saliency propagation for semi-supervised instance segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 7