The following publication Z. Deng et al., "TrafficCAM: A Versatile Dataset for Traffic Flow Segmentation," in IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 2, pp. 2747-2759, Feb. 2025 is available at https://doi.org/10.1109/TITS.2024.3510551.

TrafficCAM: A Versatile Dataset for Traffic Flow Segmentation

Zhongying Deng, Yanqi Cheng, Lihao Liu, Shujun Wang, Rihuan Ke, Carola-Bibiane Schönlieb, Angelica I Aviles-Rivero

***Abstract*—Traffic flow analysis is revolutionising traffic management. By leveraging traffic flow data, traffic control bureaus could provide drivers with real-time alerts, advising the fastest routes and therefore optimising transportation logistics and reducing congestion. The existing traffic flow datasets have two major limitations. They feature a limited number of classes, usually limited to one type of vehicle, and the scarcity of unlabelled data. In this paper, we introduce a new benchmark traffic flow image dataset called TrafficCAM. Our dataset distinguishes itself by two major highlights. Firstly, TrafficCAM provides both pixel-level and instance-level semantic labelling along with a large range of types of vehicles and pedestrians. It is composed of a large and diverse set of video sequences recorded in streets from eight Indian cities with stationary cameras. Secondly, TrafficCAM aims to establish a new benchmark for developing fully-supervised tasks, and importantly, semi-supervised learning techniques. It is the first dataset that provides a vast amount of unlabelled data, helping to better capture traffic flow qualification under a low-cost annotation requirement. More precisely, our dataset has 4,364 image frames with semantic and instance annotations along with 58,689 unlabelled image frames. We validate our new dataset through a large and comprehensive range of experiments on several state-of-the-art approaches under four different settings: fully-supervised semantic and instance segmentation, and semi-supervised semantic and instance segmentation tasks. Our benchmark dataset and official toolkit are released at https://math-ml-x.github.io/TrafficCAM/.**

***Index Terms*—TrafficCAM dataset, traffic flow analysis, semantic segmentation, instance segmentation, semi-supervised learning.**

I. Introduction

WITH the rapid population growth, the increasing number of vehicles, and human repetitive travel patterns [1], traffic congestion in city networks has become challenging to analyse [2], [3]. The overcrowding of road traffic increases the risk of traffic accidents and air pollution which undermines environmental sustainability. For that reason, smart traffic solutions are a great focus of interest. Using smart traffic solutions, traffic control bureau could provide drivers with real-time alerts advising the fastest routes, and therefore optimising transportation logistics and reducing congestion. However, deep learning-based research to develop those smart solutions requires extensive traffic data analysis to find patterns.

There is a large number of datasets for city/street scene understanding with moving and low position cameras including Cityscapes [4], Mapillary [5] and IDD [6]. However, their purpose greatly differs from ours. Contrary to these datasets that seek to provide an understanding of urban-level streets, with classes such as buildings, vegetation, and street objects, our TrafficCAM dataset seeks to focus on the traffic flow analysis. That is, our setting is data coming from traffic cameras instead of a moving street-level camera. Moreover, our dataset provides more classes for different types of vehicles.

Machine learning techniques have proven to be very successful for pattern recognition [7], [8] based on the requirement of large amounts of data. In the field of traffic data, this however is a limiting factor. Currently available datasets are either very small [9], [10] or use moving cameras [4], [11], which makes traffic flow analysis nearly impossible. *Traffic flow analysis from images and videos boils down to segmenting vehicles and people from surroundings.* We thus introduce the new, large, fixed camera traffic dataset named TrafficCAM. “CAM” stands for “camera” as our dataset is captured by traffic cameras. Moreover, “CAM” also aims to represent the University of Cambridge, where this work has been done. TrafficCAM not only covers various traffic scenes but also contains sufficient challenging samples to provide a solid basis for traffic flow segmentation (see Figure 1). Our contributions are:

☞ We introduce the largest benchmark for traffic segmentation covering a wide range of cities in India. TrafficCAM dataset is specifically designed for traffic-related object segmentation with both pixel and instance annotations. Unlike existing datasets for flow analysis, we provide the first dataset with a large variety of vehicles.

☞ Our TrafficCAM dataset, and different from other datasets, is composed of a vast amount of unlabelled frames. This key feature opens the door to further developments of robust and generalisable models that learn with limited annotated data and a large number of unlabelled data.

☞ We validate the usability of our dataset through a set of extensive experiments over the existing state-of-the-art methods, and for four different settings supervised learning to semi-supervised learning for both semantic and instance segmentation.

Our dataset can be used for traffic flow analysis, machine learning, or transportation management, such as providing drivers with real-time alerts or optimal routes and enhancing transportation logistics, ultimately reducing congestion.

ZD, YC, LL, CBS, and AIAR are with DAMTP, University of Cambridge, Cambridge, CB3 0WA, UK (e-mail: {zd294, yc443, ll610, cbs31, ai323}@cam.ac.uk).

SW is with Department of Biomedical Engineering, Hong Kong Polytechnic University, Hong Kong, China (e-mail: shu-jun.wang@polyu.edu.hk).

RK is with School of Mathematics, University of Bristol, Bristol, BS8 1UG, UK (e-mail: rihuan.ke@bristol.ac.uk).

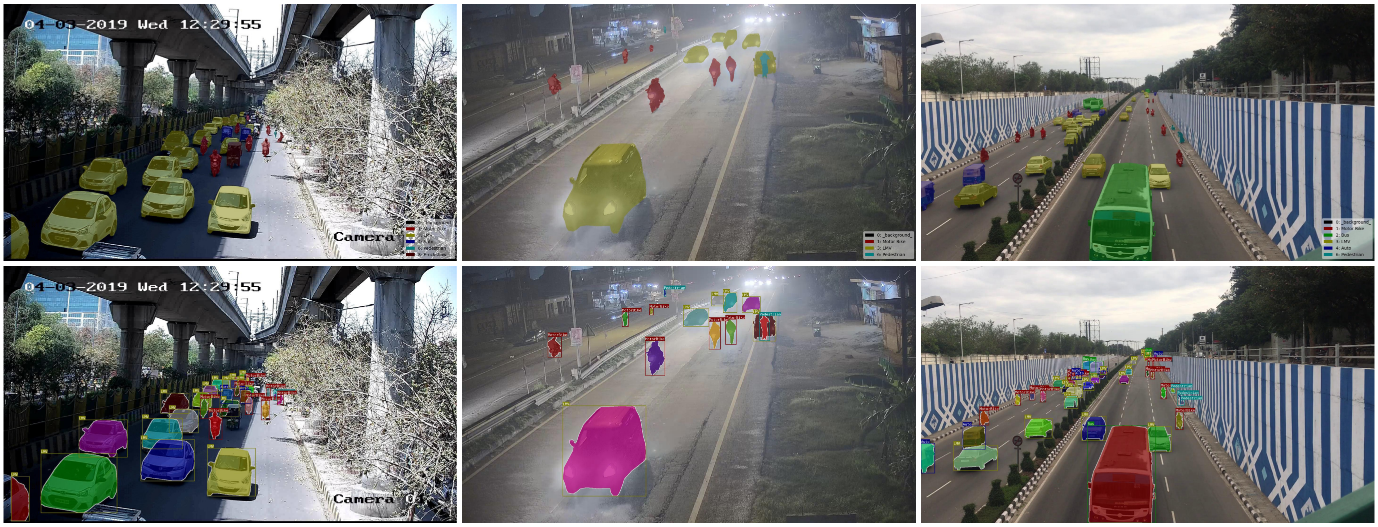


Fig. 1. Samples from our TrafficCAM datasets, which covers various traffic scenes.

TABLE I
COMPARISON OF OUR DATASET AND EXISTING TRAFFIC SURVEILLANCE DATASETS. VIDEO DURATION COLUMN SHOWS THE NUMBER OF VIDEOS AND THE TOTAL VIDEO DURATION. ‡: RELEASED IN 2023 BUT ACQUIRED IN 2019.

DATASET	Year	Videos/Duration	#Frames	#Classes	Resolution	Annotation Type	Task	Extra Frames w/o Annotations	Locations
MIT Traffic [9]	2008	1/90 mins	-	1	720×480	BBox Level	Detection	-	USA
UrbanTracker [12]	2014	5/-	8,141	3	800×600 1024×576 1280×720	BBox Level	Tracking	-	France & Canada
Ko-PER [13]	2014	6.28 mins	-	1	656×494	BBox Level	Tracking	-	Germany
CityCam [14]	2017	-	60,000	10	352x240	BBox Level	Counting & Detection	-	-
AAU RainSnow [15]	2018	22/109 mins	2,200	3	640×480	Pixel level	Semantic Segmentation	-	Danmark
MIO-TCD [16]	2018	-	786,702	11	From 342×228 to 720×480	BBox Level & Image level	Classification & Detection	-	Canada & USA
STREETS [17]	2019	-	3,000	-	-	Vehicle Number	Counting	Over 4M	USA
CityFlow [10]	2019	40/195 mins	-	1	960p	BBox Level	Tracking	-	USA
TUMTraf V2X [18]	2024		5,000	8	1920×1200	BBox Level	Detection & Tracking	-	Germany
TrafficCAM (Ours)	2023‡	-	4,364	10	352×288 1056×864 1289×720 1920×1080	Pixel level BBox level	Semantic & Instance Segmentation	58,689	India

II. RELATED WORK

Traffic data has been an inherent part of research in traffic management for nearly twenty years. In general, datasets in this area can be classified into two main categories – fixed camera traffic data and moving camera traffic data. Whilst the latter is captured e.g., from within moving cars, fixed camera traffic datasets make use of static cameras recording scenes within the same surroundings.

Moving camera datasets are popular in the community, where the goal is to record scenes with a non-static camera which can be attached to a car. Within this category several datasets have been introduced. The probably most famous dataset in this category is Cityscapes [4] which has been introduced in 2016. Further datasets in this area include the Honda Research Institute Driving Dataset (HDD) [19], the Dataset for Object deTection in Aerial images (DOTA) [11], the IDD dataset [6], and BDD100K [20]. We underline that this family of datasets greatly differs from ours and our purpose. Their goal is to capture not only traffic flow but general urban level objects. For example, buildings, vegetation, animals and non-traffic vehicles such as boats. Whilst our purpose is to tackle inherent factors for traffic flow including different types of vehicles and pedestrians. Our TrafficCAM is closer in purpose to existing fixed camera traffic datasets. In the following, we review the existing traffic datasets following this philosophy.

A. Traffic Datasets with Fixed Camera

Fixed camera traffic datasets use stationary cameras to film scenes within the same surroundings. Most of the time, the cameras are installed several meters above street level, e.g., next to traffic lights. These types of datasets can be broadly classified into categories of single- and multi- devices datasets. Single device datasets make use of a single camera, whilst multi device datasets use either multiple cameras along with other sensors like laser scanners or thermal infrared cameras.

For traffic flow analysis, there exist two fixed single camera datasets. To the best of our knowledge, the oldest available dataset is the Highway Traffic Videos [12] dataset. It was published in 2004 and consists of two days of video footage of a CCTV camera in Seattle, USA. Another dataset is the MIT Traffic [9] which has been introduced in 2009 provides a total of 90 minutes of video clips, and it is split into 20 sequences.

More recent datasets make use of multiple devices, either by recording videos of multiple cameras or a camera along with another sensory device. Multi camera traffic datasets include Urban Tracker [12], CityCam [14], MIO-TCD [16], and STREETS [17], and CityFlow [10]. Urban Tracker uses five cameras to film different surroundings from different angles. It has been introduced in 2014 and not only provides videos for traffic flow analysis of motorised vehicles but also includes pedestrians. CityCam collects 60 thousand frames from 212 web cameras for the counting and detection tasks. It has more classes than Urban Tracker, covering different types of weather. MIO-TCD is captured using thousands of traffic cameras. It has two subsets: 648,959 frames for the classification subset and 137,743 frames for the localization subset. STREETS leverages one hundred publicly available web cameras to count vehicle numbers in the suburbs of Chicago, USA. It provides over 4 million frames without annotations. For the CityFlow dataset, three hours of video of ten intersections were filmed by 40 cameras. It was published in 2019.

In addition to only multiple cameras, the Ko-PER dataset [13], AAU RainSnow Traffic Surveillance dataset [15], and TUMTraf V2X Cooperative Perception dataset [18] incorporate additional sensory data. Ko-PER uses eight monochromal cameras and 14 laser scanners to film a single intersection. The AAU RainSnow dataset however uses seven RGB cameras and seven thermal infrared cameras to film 22 five-minute videos of seven intersections in Denmark. TUMTraf V2X dataset uses onboard and roadside cameras and LiDAR to capture 3D traffic participants in Germany.

B. Existing Datasets & Comparison to Ours

TrafficCAM is a unique dataset within its category. In contrast to existing ones [9], [10], [12], [13], [15], TrafficCAM provides 10 classes whilst the majority of datasets solely focus on a single class. This feature provides an opportunity to get a more detailed *uninterrupted flow* analysis [21], [22], where more insightful vehicle-vehicle types analysis can be achieved. This is translated to have more complex yet informative models for better traffic policies.

Secondly, compared with STREETS [17], MIO-TCD [16], CityCam [14] and the latest TUMTraf V2X [18], our TrafficCAM provides pixel-level annotation, facilitating the segmentation task. Additionally, TrafficCAM is captured in India rather than North America or Europe. The different locations further lead to new features of our TrafficCAM: 1) dense traffic instances due to the huge population in India, and 2) new classes (e.g., E-rickshaw and Auto) which are rarely observed in North America or Europe.

Lastly, TrafficCAM is the first dataset to highlight a large amount of unlabelled data for segmentation, rather than vehicle number counting, e.g., STREETS [17]. Existing datasets such as MIO-TCD [16] and TUMTraf V2X [18] are limited to feature annotated data, however, labelling is expensive and time-consuming. By introducing a vast amount of unlabelled data we open the door to more robust and generalisable models that lean with limited annotations. Our key dataset highlights are in Table I.

III. TrafficCAM: A New Dataset for Segmentation

The TrafficCAM dataset is collected to capture the traffic flow from complex city networks in India. Although designing such a large-scale dataset requires a lot of efforts in image collection and annotations, our dataset has the potential to promote the traffic solution with respect to smart cities. In what follows, we detail the characteristics of our TrafficCAM dataset according to data specifications, classes and annotations, and statistical analysis.

A. Data Specifications

The TrafficCAM dataset is provided by Kritikal Solutions and they confirm that the dateset is openly available for members of the public. Our dataset is collected from the stationary traffic cameras from eight cities in India from a top and side frontal view We detail the city locations in Figure 2. This is a good indication that the image frames in the TrafficCAM dataset are from diverse regions, including the geographic north, south, inner, west, and east parts, which is a distinctive feature of the proposed benchmark. The videos in our TrafficCAM dataset are captured with various types of cameras, so that our data is not limited to one resolution. The resolution of our collected data ranges from 352×288 to 1920×1080. After the collection, we then trim all the videos to short videos, and extracted frames from each short video. Specifically, we extracted the first frame at the beginning of every second (note that a second usually has many frames) and saved them as images in JPG format. These JPG images were then sent to a professional company for annotation (Kritikal Solutions). More information can be found in Table I.

It is worth noting that diverse sources and viewpoints can have better practical relevance with a wide range of urban environments, traffic patterns, and conditions. Incorporating these properties, a dataset can improve models' robustness to different lighting, weather, and seasonal conditions, and to handle diverse road layouts and traffic behaviors. Additionally, the dataset supports advanced tasks like multi-view fusion, making it a strong benchmark for scalable, transferable traffic

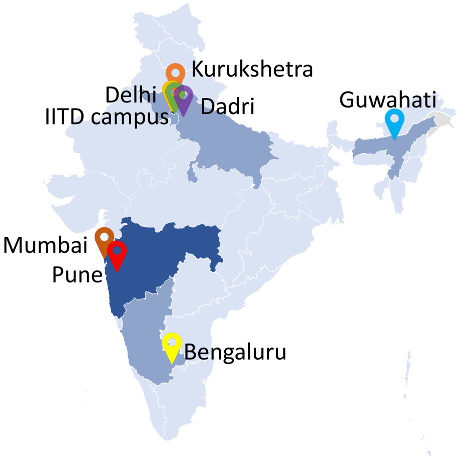


Fig. 2. Geographic distribution of TrafficCAM collection.

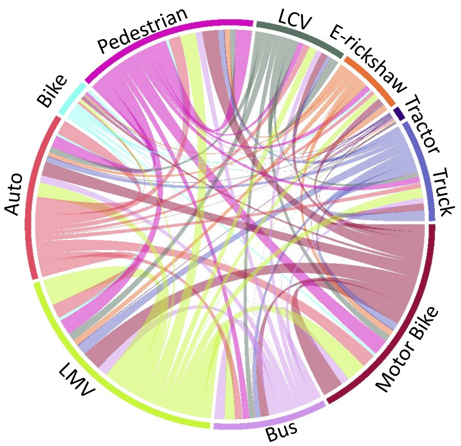


Fig. 3. Correlation analysis.

monitoring solutions. Therefore, these properties contribute to the overall quality and utility of our dataset, which is important for traffic analysis and model development.

The TrafficCAM dataset comprises 2,102 videos, where we extract a total of 63,053 frames. Specifically, the frames from 78 videos have fully pixel-level annotations, while 2,024 videos only have the first frame annotation (where 218 videos are captured under foggy weather, 123 under the night condition). Therefore, our TrafficCAM dataset has a large number of frames without annotations (58,689 frames) from the above partially annotated videos, which opens the door to the development of techniques relying on limited annotations as semi-supervised learning.

We split our TrafficCAM dataset into training, validation, and test sets. The dataset is split at the video level, which guarantees that the frames from the same video only appear in one split. We chose not to split the data randomly due to the special annotation character of our TrafficCAM dataset. Specifically, we set all the videos with only first-frame annotation and 8 fully annotated videos (30 frames per video) as the training split, while the remaining videos with fully annotations are randomly split into validation and testing. We left out 12 videos for future purposes like optical flow learning. Our split criteria finally leads to 2,264 annotated frames, and 58,689 unannotated frames for training, 210 annotated frames for validation, and 1,530 annotated frames for testing. We also remark that our dataset covers several conditions such as night and day acquisition, and other conditions such as foggy scenarios.

B. Classes and Annotations Details

For each annotated frame in TrafficCAM dataset, we provide pixel annotations for ten classes in both semantic and instance levels. Due to the attribute of our data to analyse traffic flow data, *we only provide annotations on the moving traffic vehicles and pedestrians on the main road.* More annotation examples and unannotated data can be found in Figure 8.

The dataset primarily affects the image content and the traffic categories. Our TrafficCAM dataset consists of 10 categories, which can be organised into 2 groups, vehicle, and pedestrian. The vehicle group contains 9 different classes, namely Motor Bike, Bus, LMV, Auto, Bike, LCV, E-rickshaw, Tractor, and Truck. The example images of the some vehicle categories are shown in Figure 4. The class overview is shown in Figure 5. We highlight that this is the first traffic flow dataset to provide such large variety of classes.

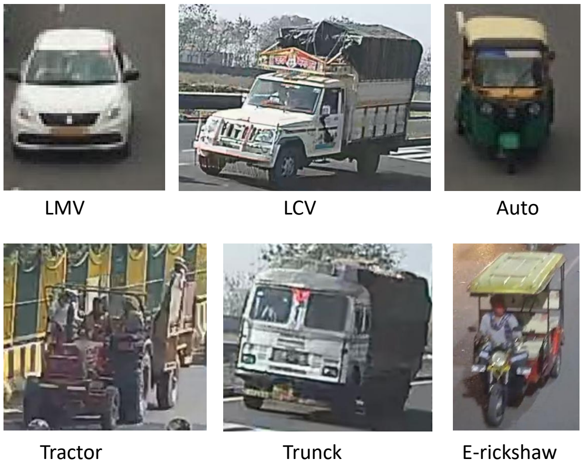


Fig. 4. Images of LMV, LCV, Auto, Tractor, Truck, and E-rickshaw. LMV means Light Motor Vehicle, referring to light and small motor vehicles, e.g., cars and jeeps. LCV is the abbreviation for Light Commercial Vehicle which is defined as any commercial vehicle with a gross weight of less than 3.5 tonnes. It includes car-derived vans, panel vans, box vans, etc. Auto stands for auto rickshaw which is a motorized version of the pulled rickshaw, usually having three wheels. Similar to Auto, electric rickshaws (E-rickshaws) are small three-wheeled vehicles powered by an electric battery. A Tractor usually has large back wheels and thick tyres. It is designed to haul a trailer or machinery for agriculture or construction. The trunk can be used to transport freight or carry heavy payloads. It features body-on-frame construction, with a cabin that is independent of the payload portion of the vehicle. The other categories of Motorbikes, Bike, Bus, and Pedestrian widely exist in many other traffic datasets, so we omit the clarification on them for brevity.

The criteria for selecting categories for the annotations process are driven by the following reasons. 1) Our dataset mainly focuses on the traffic information instead of scene understanding like Cityscapes [4]. Therefore, vehicles and people play a more important role. 2) We provide a large variety of vehicles captured from a large range of cities in India, where unique and complex vehicles and traffic situations can be captured. This feature introduces a unique opportunity to develop more robust techniques. The most common types of cars on Indian streets have been included under this category.

Our TrafficCAM dataset provides both bounding box-level and pixel-level annotation. Pixel-level segmentation outlines object boundaries in the scene, complementing bounding boxes. This allows for more accurate object behavioral analysis, enabling the detection of subtle patterns and anomalies on the road. Additionally, it enhances the visibility of vehicles in interaction areas and the road's far end, where bounding boxes overlap. These advantages of pixel-level segmentation can help better identify areas of congestion, potential bottlenecks, lane encroachment, illegal parking, and design better pedestrian safety measures.

The annotations comprise layered polygons from in-house LabelMe [23] to guarantee the highest quality levels. Specifically, we first manually annotate the dataset only for our object of interest, e.g., cars and motorbikes. The annotations are done by the professional company named KritiKal Solutions. Labels are in the form of bounding boxes and segmentation masks, but do not contain any information about the identities of individuals appearing in the images. The company was asked to annotate from back to front to avoid object boundaries being labeled repeatedly. The objects with less than 10 pixels were not annotated because human eyes can hardly recognize these ultra-small objects, making the annotation highly untrustworthy. After the annotations are completed by the professional company, a quality control group checks all the annotations and finds out low-quality ones which will be sent back to the company for annotation refinement. Specifically, the quality control group comprises two Ph.D. students and a senior research fellow in this field. These two students check all the annotations one by one and record all the missed annotations or wrong annotations. They are asked to take down the file names of these low-quality annotations and write instructions on how to improve the quality. Then all the low-quality annotations are reported to the senior research fellow for double check. If there are disagreements on specific annotations, a discussion will be conducted until the consensus is achieved. After that, these low-quality annotations and the instructions for improving their quality are sent back to the company for further refinement. Finally, the refined annotations are returned to the quality control group for a new round of checking. This process is conducted several times until no low-quality annotations can be found by the quality control group, which ensures a high annotation quality.

Frames without Annotations. Another key highlight in our dataset is that we provide a massive amount of unlabelled samples making our TrafficCAM dataset substantially different from other existing datasets. The main purpose to provide such a large unlabelled set is two-fold. Firstly, annotations are time-consuming and expensive. Secondly, our TrafficCAM dataset provides an opportunity for developing robust and generalisable techniques with limited annotated data and a large number of unlabelled data. We therefore provide a strong semi-supervised benchmark with the ratio of unlabelled frames to labelled frames reaching 30:1. In the experiment section, we provide comparison results not only under the fully supervised learning setting but also under semi-supervised learning schemes. *To the best of our knowledge, TrafficCAM is the largest traffic object segmentation benchmark under both supervised and semi-supervised settings.*

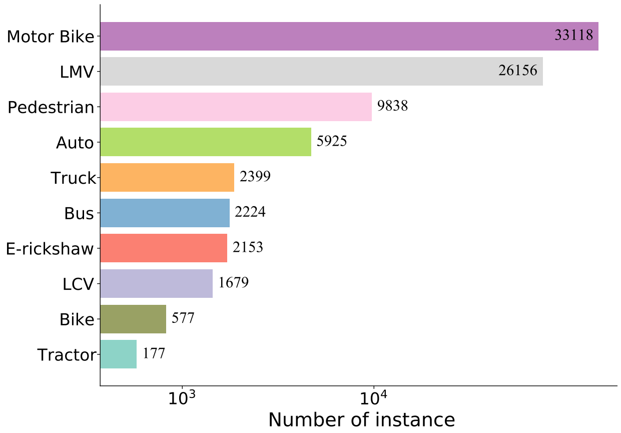


Fig. 5. Illustration of the number of instances per category.

C. Statistical Analysis

We illustrate the number of images per category in Figure 5. As we can see, our dataset is a typical long-tailed dataset. The Motor Bike and LMV occupy around 90% percentage of all frames, while the frame numbers of Bike and Tractor are no more than 600. Such imbalance natural character makes our TrafficCAM dataset more challenging yet of a great interest for model development.

We report in Figure 6 the frequency of images with a fixed number of annotated objects in the TrafficCAM dataset. TrafficCAM is the largest dataset focused on traffic stationary views. We find that our dataset covers a good variety of category complexity. There are more than 600 images with the number of instances being more than 40 yielding to challenging cases. From Figure 3 we may also observe the diversity of classes in our dataset frames.

We further provide the distribution of object sizes in Figure 7. We can see that most annotated objects are of small sizes. For example, about 51.6K objects are smaller than 1024 pixels (32×32), among which 23K objects are smaller than 256 pixels (16×16). It is also worth noting that the smallest objects are at 10 pixels. These small-size objects make our TrafficCAM a challenging dataset for segmentation. This phenomenon is mainly caused by the fixed traffic cameras installed several meters above street level. From Figure 7, we can also observe that the object size in our dataset is sufficiently diverse, covering small sizes ($\leq$16×16 pixels) to large ones (*e.g.*, $\geq$160×160 pixels).

Compared with other existing traffic surveillance dataset, our TrafficCAM dataset contains the most video and frame numbers, as shown in Table I. The resolution of our dataset ranges in a big difference, which shows the diversity of the

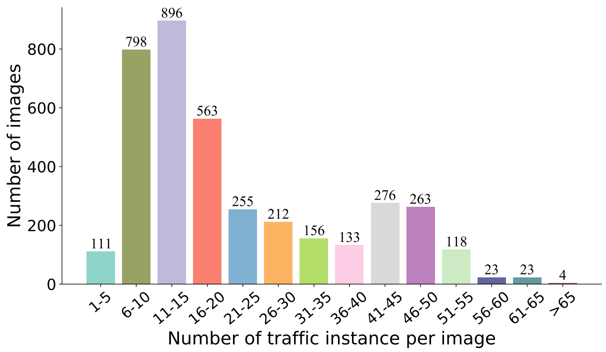


Fig. 6. Dataset statistics regarding complexity: the frequency of images with a fixed number of annotated object instances per image.

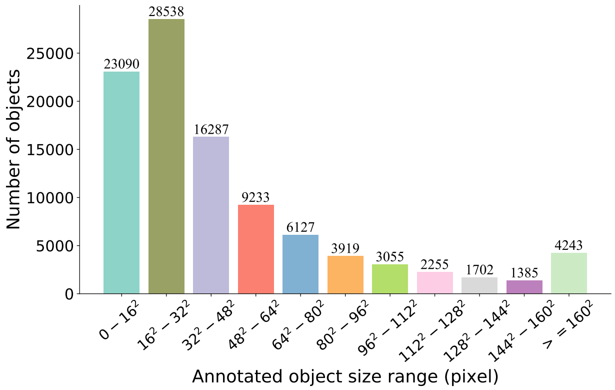


Fig. 7. Distribution of annotated object sizes.

TrafficCAM dataset. *Furthermore, our dataset is the only one that provides both pixel and instance level annotations under traffic object segmentation scenario.*

D. Temporal Information

TrafficCAM provides frame indexes of each video's snippet and leaves out 12 videos for optical flow learning (in addition to still images for segmentation). With the optical flow, we can obtain coarse temporal information for further flow analysis, like tracking. However, we emphasise semantic and instance segmentation as carefully designed schemes together with our collaborators from epidemiology and transport experts, where temporality is not as informative for the evaluation of traffic infrastructure interventions. For example, for planning or evaluating a new cycle infrastructure, pedestrianised streets or other developments that support a healthier and more active lifestyle. We underline that TrafficCAM is unique in this sense and will make a high impact not only in the ML community but also in real-world design for epidemiology and transport policies.

E. Data Privacy and Ethical Issues

Our TrafficCAM has obtained ethical approval from the University of Cambridge covering any issues with ethical implications including data protection regulations, privacy, and copyright.

Throughout the collection, annotation, and analysis processes, we seek to tackle the privacy issues as follows. 1) In our dataset, videos are collected in public areas (e.g., the main roads, and streets) of Indian cities. In this collection process, the privacy issue can be partly addressed as we are not targeting a specific group of people, and the individuals appear in the video in a random fashion, which is almost not recognisable due to the resolution of the data. In addition, these videos have gained permission from the data source KritiKal Solutions to be used for our analysis. 2) During the annotation process, we ask KritiKal Solutions to manually annotate the dataset only for our object of interest e.g., cars and motorbikes. Labels are in the form of bounding boxes and segmentation masks, but do not contain any information about the identities of individuals appearing in the images. 3) For the analysis process, the main participants are people appearing in the video data, but our project extracts only statistical/quantity information rather than personal information from the data. Due to the complication of informing the participants and based on the general public interests of our project, we apply for a UK General Data Protection Regulation (GDPR) exemption of consent from the participants which will be introduced in the next paragraph. Throughout the whole process, our dataset does not include any other type of private information except images. To ensure anonymity, the names, addresses, and other information of the people appearing in these images are not provided in our dataset.

Following our institutional protocol, our dataset is qualified for a UK GDPR exemption for the following reasons (according to GDPR Articles 5(1)(b) and (e), 14(1)-(4), 15(1)-(3), 16, 18(1), 19, 20(1) and 21(1)):

- The access to the video data is for scientific and statistical purposes. The objective of this research is to develop novel image analysis and machine learning methods to understand health, traffic, and urban policies in India's cities. We extract useful statistical information rather than individual information from the data.
- The project studies the general health and urban policies in the cities. As the focus is not on the behaviour of individuals, the project will not cause any substantial damage or substantial distress to an individual. No measure or decisions will be made upon particular individuals as a result of the project, since our algorithms are unable to reveal personal information.
- Research results are in the form of methodologies and image analysis toolboxes. No individual information will be published.
- The research in the project is for the public interest. We process the data in order to analyse the general public health and help to improve urban policies.

IV. Experiments

We validate our TrafficCAM dataset through a large and comprehensive range of experiments on several state-of-the-art approaches under four different settings: fully-supervised semantic and instance segmentation, and semi-supervised semantic and instance segmentation tasks.

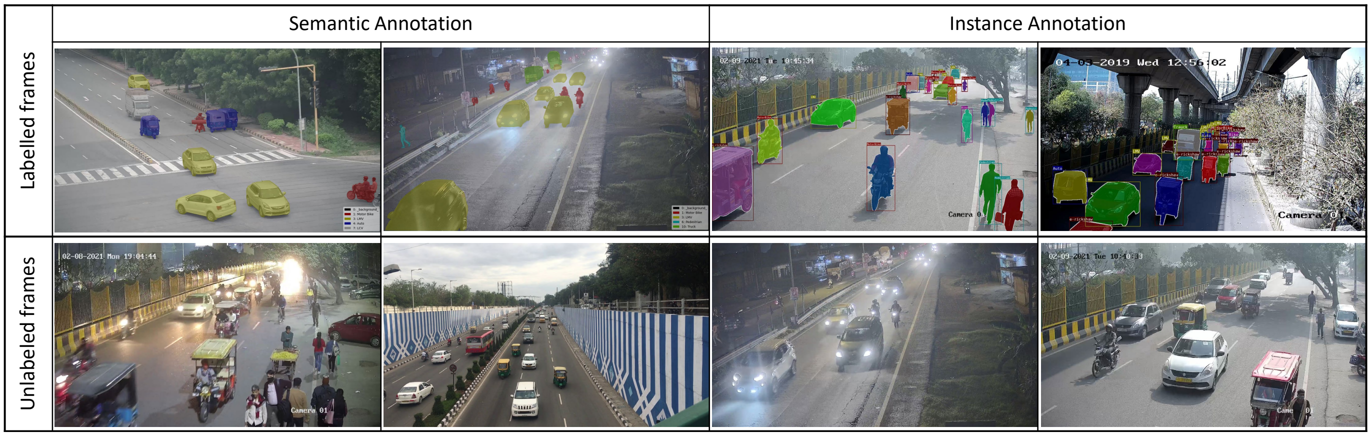


Fig. 8. Illustration of our TrafficCAM dataset. The first row shows labelled frames, where the first two columns are with semantic annotation and the latter two with instance annotation frames. The second row consists of unlabelled image frames.

Specifically, we highlight the semi-supervised tasks which aim to improve the model's performance using limited labelled data and enormous unlabelled set. This task is meaningful because it can significantly reduce the annotation cost – annotating images is labor-extensive and time-consuming, especially when pixel-level annotations are required for segmentation tasks. Our dataset accommodates this scenario as it contains both labelled and massive unlabelled data. Therefore, we further investigate different baseline methods on our dataset for semi-supervised semantic and instance segmentation.

A. Evaluation Metrics

Semantic segmentation task: We use Intersection over Union scores (IoU) to evaluate semantic segmentation tasks on our benchmark. IoU is defined as $\frac{TP}{TP+FP+FN}$, where TP stands for the number of true positive pixels, FP for false positive, and FN for false negative. We compute the IoU for each class and report the mean IoU (mIoU) as the final result.

Instance segmentation task: The evaluation metrics used for instance segmentation is region-level average precision (AP) for each class, the same as [24]. Specifically, the AP is calculated by first computing the IoU of a single instance in a region, then applying different thresholds, *e.g.*, from 0.5 to 0.95 with step 0.05, to the IoU to obtain the scores for predictions, and finally averaging these scores. The class-wise AP is then used for computing mAP, which is the mean of the class-wise AP over all the classes. We also report the mAP$^{50\%}$ for the threshold=50%.

B. Fully Supervised Semantic Segmentation

For this task defined on TrafficCAM, we perform pixel-level classification for semantic labels, without considering multiple instances. For example, if multiple pedestrians are in the same image, their pixels will be classified into the same label 'pedestrian'.

Experimental settings. We adopt MMSegmentation [25] to run the experiments on our TrafficCAM. During training, we adopt random flipping and scale as the data augmentation, with a scale ratio in $[0.5, 2]$. We then crop the image to the same size for model training. We train the model using SGD optimiser with learning rate initially set as 0.01 with polynomial decay at a power of 0.9; momentum at 0.9 and weight decay at 0.0005. All the baselines are trained with a batch size of 16 for $20K$ iterations, and tested using the model with the best validation mIoU result.

During inference, we adopt test-time augmentations to flip and resize the testing images to multiple scales, the same as training. These multi-scale predictions are interpolated in a bi-linear way so that they are of the same size as the input images. The predictions of different scales are then averaged as the final prediction.

State-of-the-art methods. We consider two different categories of the most widely-used baselines: ResNet-based and Transformer-based methods. The most classical methods are usually based on ResNet, including FCN [26], PSPNet [27], DeepLabV3+ [28]. FCN is the first work to exploit deep CNNs for semantic segmentation. It motivates the follow-up methods such as PSPNet and DeepLabV3+. The latter two methods try to improve FCN by constructing multi-scale representations. As the field evolved, attention mechanisms and transformers began to take center stage. Therefore, the latest methods are usually based on transformers. Transformers have achieved state-of-the-art performance on semantic segmentation task partly because the attention mechanism can effectively capture long-range dependencies for such task. Some representative works are SegFormer [29] and SETR [30], both employing vision transformers (ViT) [31] as backbone.

Experimental results. We conduct experiments with different backbones and crop sizes to better evaluate the SOTA methods on TrafficCAM. We show the results in Table II and find that SegFormer with MiT-B5 backbone and 1024×1024 crop size achieves the best performance, *i.e.*, 71.21%. In addition, we have the following observations:

For these ResNet-based methods, DeepLabV3+ with ResNet-101 backbone and 512×1024 crop size achieves the best performance, *i.e.*, 66.49%. Another observation is that deeper ResNet always achieves better mIoU regardless of different methods, *e.g.*, FCN with ResNet-101 increases the mIoU of ResNet-50 by 7.85% while DeepLabV3+ benefits from a deeper backbone by about 10%. However, deeper ResNet can lead to slower inference speed, e.g., decreasing from ∼6 FPS to ∼5. Crop sizes of 769×769 usually bring a performance gain at the expense of slower inference speeds, compared with 512×1024. For instance, it improves the performance of PSPNet from 63.61 to 66.21. The exception is that 769×769 results in a subtle decrease for DeepLabV3+.

For the transformer-based methods, we adopt a fixed crop size with 1:1 aspect ratio to facilitate patch embedding. For SegFormer, it proposes MiT backbone which is based on ViT but tailored for semantic segmentation tasks. It is clear that larger backbone models usually contribute to better mIoU, *e.g.*, 64.01% of MiT-B5 over 60.93% of MiT-B4 for the crop size of 512×512, and that larger crop size favorably increases performance, *e.g.*, for MiT-B5 backbone, 71.21 obtained by 1024×1024 beating 64.01 of 512×512. Despite effectiveness, the efficiency of MiT-B5 with 1024×1024 crop size is poor which results in 2.0 FPS inference speed. This is about 4 times slower than the most efficient MiT-B0 with 512×512 crop size (8.8 FPS). SETR adopts ViT-Large as its encoder and further designs different decoders, namely, the naive decoder (Naive), the progressive upsampling one (PUP), and the multi-level feature aggregation (MLA). We find that the naive version achieves the best performance on TrafficCAM.

Comparing the ResNet-based methods with the transformer-based ones, we observe that the ResNet-101-based FCN, PSPNet, and DeepLabV3+ generally obtain better (or at least comparable) performance than the SETR, but cannot beat the SegFormer with MiT-B5. The SegFormer achieves the best performance of 71.21%, outperforming the best ResNet-based method of DeepLabV3+ by 4.72%. Regrading the inference speed, SegFormer with MiT-B0 as the backbone and 512×512 as the crop size obtains the fastest inference speed, i.e., 8.8 FPS. Overall, smaller crop size and lighter backbone bring better inference speed.

Overall, all these methods are unsatisfactory on our TrafficCAM, showing that TrafficCAM is a challenging dataset for semantic segmentation. We owe the challenge to the diverse scenes and highly imbalanced class distributions.

TABLE II
RESULTS FOR FULLY SUPERVISED SEMANTIC SEGMENTATION TASKS. FPS DENOTES FRAMES PER SECOND, SHOWING THE INFERENCE SPEED ON A SINGLE NVIDIA A100-SXM-80GB GPU. THE BEST RESULT IS IN BOLD.

	Methods	Backbone	Crop Size	FPS	mIoU
ResNet	FCN [26]	R50	512 × 1024	6.2	55.61
		R101	512 × 1024	5.4	63.46
		R101	769 × 769	2.9	65.77
	PSPNet [27]	R50	512 × 1024	6.2	60.58
		R101	512 × 1024	5.3	63.61
		R101	769 × 769	2.8	66.21
	DeepLabV3+ [28]	R50	512 × 1024	6.4	56.50
		R101	512 × 1024	5.1	66.49
		R101	769 × 769	2.6	66.26
Transformers	SegFormer [29]	MiT-B0	512 × 512	**8.8**	52.61
		MiT-B0	1024 × 1024	6.0	53.50
		MiT-B4	512 × 512	5.8	60.93
		MiT-B4	1024 × 1024	2.4	70.77
		MiT-B5	512 × 512	6.4	64.01
		MiT-B5	1024 × 1024	2.0	**71.21**
	SETR [30]	ViT-L_MLA	512 × 512	6.1	54.57
		ViT-L_MLA	768 × 768	1.8	60.87
		ViT-L_Naive	512 × 512	5.7	57.91
		ViT-L_Naive	768 × 768	1.1	61.70
		ViT-L_PUP	512 × 512	5.8	57.15
		ViT-L_PUP	768 × 768	1.0	63.11

C. Fully Supervised Instance Segmentation

Instance segmentation aims to detect and segment each object instance. In this task, even if these instances are from the same class, we will need to assign them a separate label.

Experimental settings. The fully supervised instance segmentation baselines are implemented in MMDetection [32]. We adopt similar settings as in Section IV-B. The differences are that 1) we keep the crop size to 1333 × 800 for all the different instance segmentation methods; 2) we train the model using AdamW [33] optimizer with learning rate initially set as 0.0002, weight decay at 0.0005, and polynomial decay at power of 0.9. All the methods are trained for $20K$ iterations.

State-of-the-art methods. According to whether region proposals are used, the SOTA methods can be divided into one-stage (without) and two-stage (with region proposals) methods. We evaluate six SOTA methods on our TrafficCAM, covering both one-stage and two-stage methods. They are: Mask R-CNN [34], Cascaded Mask R-CNN [35], InstaBoost [36], SOLOv2 [37], QueryInst [38], and Mask2Former [39].

Mask R-CNN is a well-known instance segmentation method. It essentially exploits Faster R-CNN [40] for instance detection and then uses FCN to segment each detected instance. Cascaded Mask R-CNN further improves it by using a sequence of detectors to balance the positive and negative samples. These two are two-stage methods which use region proposals for instance segmentation. InstaBoost uses the existing instance mask annotations for random jittering to objects so that the training set can be augmented. SOLOV2 introduces dynamic parameters to the mask head of object segmenter so that the mask head is conditioned on the location, where the location information usually contributes to better performance. Notably, SOLOv2 is a one-stage method without region proposal. QueryInst also employs a dynamic design, but differs in treating instances of interest as learnable queries. Mask2Former constrains cross-attention within predicted mask regions to extract localized features. It takes Swin transformer [41] as its backbone.

Experimental results. From Table IV, we can observe that 1) Mask2Former equipped with Swin-S [41] and InstaBoost with ResNet-101 outperform all the other competitors, both ranking first in terms of mAP. Mask2Former also obtains the best mAP$^{50\%}$ of 71.8%, which illustrates the powerful representation ability of attention mechanism. 2) All these method benefits more from larger backbones than the small ones (*e.g.*, ResNet-101 always beats ResNet-50), which is similar to conclusion drawn from semantic segmentation tasks. 3) The one-stage method, SOLOv2, is inferior to these two-stage methods like Mask R-CNN and Cascade Mask R-CNN. This conclusion is also justified in many other datasets. 4) The mAP of each SOTA method are relatively low, *e.g.*, less than 50%, which manifests the challenge of our TrafficCAM.

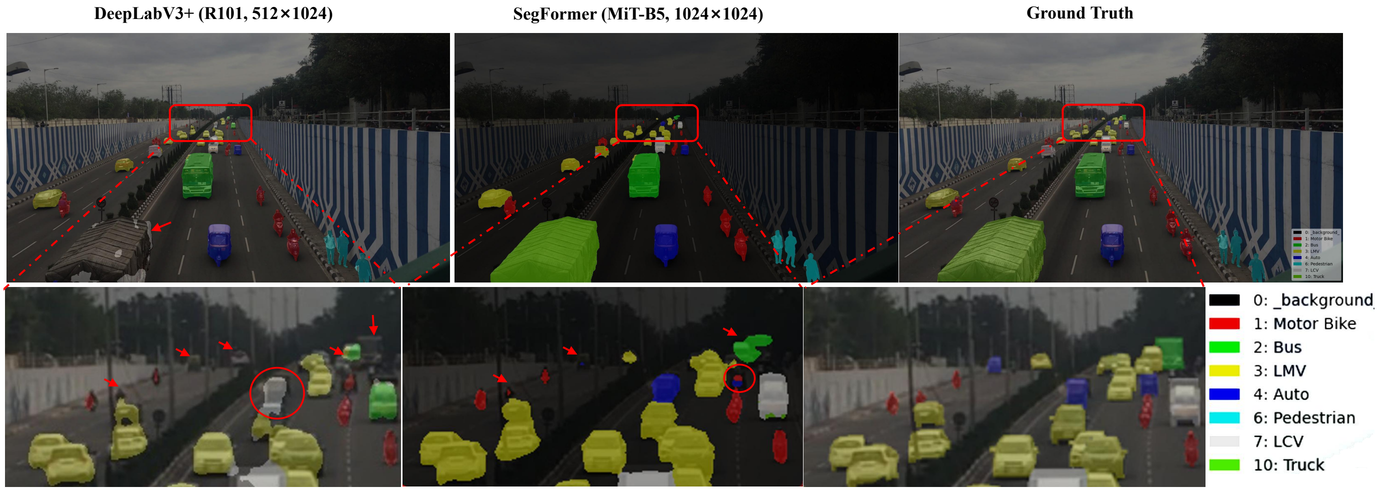


Fig. 9. The failure cases of DeepLabV3+ and SegFormer for fully supervised semantic segmentation task. The bottom row zooms in on the rectangle area in the top row. The red arrows show the missing objects or inconsistent segmentation results while the red circles depict the wrong predictions.

To better understand the challenges and difficulties of our datasets, we show in Figure 9 the failure cases of DeepLabV3+ and SegFormer, being the best-performing ResNet and Transformer methods. DeepLabV3+ cannot correctly classify the Truck that is very close to the traffic camera (see the red arrow in the top row). In addition, it fails to capture the LMV, Auto, and Pedestrian in the far distance (the red arrows in the bottom row) due to the ultra small scale. The missing objects or inconsistent segmentation caused by scale changes can also be observed in SegFormer. Both DeepLabV3+ and SegFormer assign wrong class labels to Auto (see the red circle regions), probably due to the occlusion of two different objects or the small scale of the objects. These failure cases highlight the challenges of our dataset.

Transfer learning results. Considering the unsatisfactory on our TrafficCAM, we further consider fine-tuning models, pre-trained on existing segmentation datasets, to investigate the impact of transfer learning strategy on our TrafficCAM. Specifically, we fine-tune the best-performing ResNet and Transformer-based methods, namely DeepLabV3+ and SegFormer, on our dataset. MMSegmentation provides the Cityscapes pre-trained models of these two methods where Cityscapes is a popular dataset. Thus, we fine-tune the Cityscapes pre-trained model on our TrafficCAM and present their results in Table III. The results show that the pre-trained models based on existing segmentation datasets can improve the performance: the Cityscapes pre-trained models can increase the mIoU of DeepLabV3+ by 14.57 (81.06 versus 66.49), and SegFormer by 6.72 (from 71.21 to 77.93). However, there is still room for further improvement.

TABLE III
THE PERFORMANCE OF FINE-TUNING MODELS PRE-TRAINED ON EXISTING SEGMENTATION DATASETS FOR FULLY SUPERVISED SEMANTIC SEGMENTATION TASKS.

Methods	Backbone (Pre-trained)	Crop Size	mIoU
DeepLabV3+ [28]	R101 (ImageNet) R101 (Cityscapes)	512×1024 512×1024	66.49 81.06
SegFormer [29]	MiT-B5 (ImageNet) MiT-B5 (Cityscapes)	1024×1024 1024×1024	71.21 82.32

TABLE IV
RESULTS FOR FULLY SUPERVISED INSTANCE SEGMENTATION TASK. THE BEST RESULTS ARE IN BOLD.

Methods	Backbone	mAP	$\mathrm{mAP}^{50\%}$
Mask R-CNN [34]	R50	43.7	64.1
	R101	45.1	65.3
Cascade Mask R-CNN [35]	R50	44.5	64.9
	R101	45.3	65.3
InstaBoost [36]	R50	46.1	68.0
	R101	**47.8**	69.5
SOLOv2 [37]	R50	28.2	45.0
	R101	30.2	48.0
QueryInst [38]	R50	45.9	69.8
	R101	46.6	71.3
Mask2Former [39]	Swin-T	47.5	70.2
	Swin-S	**47.8**	**71.8**

D. Semi-Supervised Semantic Segmentation

Experimental settings. For the semi-supervised task, we consider different ratios of labelled samples to all the training set: 1/27, 1/54, 1/108, 1/216 where we use 2,264 (all of our annotated frames in the training set), 1,137, 572 and 288 frames as labelled samples respectively. We then ignore the labels of the remaining labelled samples and use them as unlabelled data. The label-ignored samples together with 58,689 unlabelled ones comprise the unlabelled training set. Throughout the experiments, the total number of labelled and unlabelled training samples is fixed to 60,953. The experimental settings of SOTA methods are largely method-dependent. We thus use their released code and follow their settings used for Pascal VOC [46] dataset.

State-of-the-art methods. We implement the following semi-supervised semantic segmentation methods: CutMix [42], ClassMix [43], Three-stage Self-training [44], and ST++ [45]. The former two are the famous semi-supervised se-

TABLE V
RESULTS (MIOU) FOR SEMI-SUPERVISED SEMANTIC SEGMENTATION TASKS. THE BEST RESULTS OF EACH LABEL COUNT ARE IN BOLD.

Methods		Backbone	1/27	1/54	1/108	1/216
CutMix [42]		R101-DeepLabV2	54.94	54.05	52.58	49.73
ClassMix [43]			56.10	**54.79**	52.92	49.81
Three stage [44]	Stage1		50.56	50.72	46.72	47.71
	Stage2		55.04	53.37	51.18	51.09
	Stage3		**56.72**	54.31	**53.71**	**53.17**
ST++ [45]		R50-DeepLabV3+	55.34	50.17	46.83	48.92

mantic segmentation methods proposing augmentation strategies, while the latter two are the latest ones based on pseudo-labels. CutMix augments an unlabelled image by pasting a randomly copied patch of another sample into it. It is initially proposed for the semi-supervised classification task, but was further extended to segmentation. ClassMix further improves CutMix for segmentation tasks by using the predicted semantic masks for mixing. Three-stage Self-training leverages a self-training strategy which utilises pseudo-labels for unlabelled data to train a better model. A three-stage training scheme is then proposed to refine the pseudo-labels. Also following this paradigm, ST++ further exploits the prediction discrepancy of multiple checkpoints to measure the quality of pseudo-labels, with high-quality pseudo-labels used for model training.

Experimental results. Table V shows the comparison of these methods on our TrafficCAM dataset. We observe that ClassMix is the champion for 1/54 label count while Three Stage wins the other label count settings. The trend for the varied label count is that the mIoU of all these four baselines decreases when the label count reduces. This is reasonable and consistent with what most of the semi-supervised works [44], [45] find, which infers that the boost in performance brought by the increase in unlabelled frames cannot counteract the effect of reducing the labelled frames. We also notice that Three Stage and CutMix is less sensitive to varied label count than ST++. For example, when the label count changes from 1/27 to 1/54, the mIoU of CutMix only decreases by 0.89% while ST++ drops by 5.17%.

We further analyze the results for each specific method. Regarding the augmentation-based methods, CutMix is beaten by ClassMix for all the label count settings. This is not surprising as CutMix is not proposed for segmentation tasks while ClassMix improves it for segmentation. ClassMix is also highly competitive with best-performing Three Stage method when there are sufficient labels, like 1/27, 1/54, and 1/108. But it is less effective to handle the label-scarce setting of 1/216. The first stage training in Three stage [44] method is also the supervised only (SupOnly) baseline. Its performance can be improved remarkably (up to 6% gain) by using the unlabelled data for the second and third stage training, in all the label count ratios. The improvement highlights the importance of the vast amount of unlabelled samples in our TrafficCAM. This is also one of the major contributions of our TrafficCAM.

Though ST++ [45] performs close to the other baselines at the label count of 1/27, its mIoU decreases considerably when the number of labels drops. The significant gap between ST++ and the others may be because the backbone used for ST++ is ResNet-50, whereas all the other baselines employ a deeper backbone, ResNet-101. Here, ResNet-50 is used for ST++ to save the training cost –if ResNet-101 is used, ST++ takes significantly longer training time (*e.g.*, 12× GPU hours) than all the other baselines.

It is also noticeable that the semi-supervised results are significantly worse than the baselines in Table II implemented with MMSegmentation. The reason can be three-fold. First, these semi-supervised methods do not use test-time augmentation such as flip and multi-scale during inference while MMSegmentation does. Second, the crop size for these methods can be smaller than that in MMSegmentation, *e.g.*, Three Stage [44] uses a crop size of 321, smaller than 769 in MMSegmentation. Last, the semi-supervised methods use CNN backbone, which can be less effective than the transformers used by SETR or SegFormer.

E. Semi-Supervised Instance Segmentation

For this task, we use the same label count setting as in Section IV-D. Since there are limited works focusing on this task, we only choose the most recent one for evaluation, *i.e.*, the pseudo-label-based method [47]. This work aims to tackle the noisy boundaries in pseudo-labels by introducing noise-tolerant mask and boundary-preserving map. For the experimental setting, we adopt the ResNet-50 based Mask R-CNN as the baseline to perform the training in [47]. We use the official code released by [47], which is also based on MMDetection [32].

Experimental results. Table VI shows the comparison of [47] with supervised-only baselines. The semi-supervised method [47] consistently surpasses the supervised only (SupOnly) baselines by a clear margin, *e.g.*, >2% mAP and 3% mAP$^{50\%}$ over SupOnly. The better performance attributes to the large amount of unlabelled data in our TrafficCAM dataset. Furthermore, the mAP drops with the decreasing number of labels. The deterioration requires the design of advanced semi-supervised methods to tackle the extreme label-scarce setting on our TrafficCAM.

Since [47] adopts the same MMDetection [32] for the experiments as in fully supervised cases (see Table 3 of the main paper), it is interesting to compare [47] with the fully-supervised results under the same framework (and similar settings). For a fair comparison, we only consider the label count of 1/27 (2,264 fully labelled training images), where the number of labelled data is the same as the fully-supervised

TABLE VI
RESULTS (MAP) FOR SEMI-SUPERVISED INSTANCE SEGMENTATION TASKS. THE BEST RESULTS OF EACH LABEL COUNT ARE IN BOLD.

Setting	1/27		1/54		1/108		1/216	
	mAP	mAP$^{50\%}$	mAP	mAP$^{50\%}$	mAP	mAP$^{50\%}$	mAP	mAP$^{50\%}$
SupOnly	44.3	65.0	41.4	62.0	35.0	52.7	31.7	50.4
Semi-sup. [47]	**47.4**	**69.1**	**43.9**	**66.4**	**37.4**	**58.8**	**33.7**	**53.9**

TABLE VII
RESULTS FOR FULLY SUPERVISED SEMANTIC SEGMENTATION TASK WITH DEEPLABV3+ BASED ON RESNET-101 AS BACKBONE AT CROP SIZE 512 × 1024, TRAINED ON 1967 FRAMES.

Conditions	Test Frames	mIoU
Normal	1524	65.13
Fog	218	61.70
Night	123	39.67

case. We can see that the semi-supervised method [47] outperforms the Mask R-CNN baseline by 0.7 mAP (47.4 of [47] vs. 43.7 in Table IV). Notably, it achieves comparable performance to the transformer-based Mask2Former (47.4 vs. 47.8 mAP). We then observe that effective exploitation of massive unlabelled data can significantly contribute to performance gain, given that a large scale unlabelled set are available in our TrafficCAM.

F. Experiments on Different Acquisition Conditions

Our TrafficCAM covers several conditions, such as night and day acquisition, and foggy scenarios. In this section, we further experiment to study these conditions.

We use all the 218 foggy and 123 night videos as two different test sets and the remaining normal conditions as the training set. Following the fully supervised semantic segmentation settings, we adopt the DeepLabV3+ with ResNet-101 and 512×1024 crop size to perform such study. The DeepLabV3+ obtains 65.13 when tested on normal (day) conditions. But its performance drops to 61.7 and 39.67 when tested on foggy and night conditions, respectively. The drop suggests the challenges of these conditions.

V. CONCLUSION

We introduce TrafficCAM, the largest traffic flow benchmark for traffic segmentation covering a wide range of cities in India. The key features of our dataset are two-fold. Firstly, the large variety in classes. Secondly, TrafficCAM is the first dataset to provide a vast amount of unlabelled data. We seek to push forward robust and generalisable models that can learn under limited annotations. We validate our TrafficCAM dataset on SOTA techniques over four challenging settings: fully-supervised semantic and instance segmentation, and semi-supervised semantic and instance segmentation tasks. The experimental results show that even the best performing methods can only achieve unsatisfactory performance on our TrafficCAM dataset, with the best mIoU of 66.49% for semantic segmentation and mAP of 47.8% for instance segmentation. It hence demonstrates the complexity of TrafficCAM, opening the door to develop more advanced segmentation methods.

ACKNOWLEDGMENTS

Zhongying Deng, Angelica I Aviles-Rivero, and Carola-Bibiane Schönlieb acknowledge support from the EPSRC grant EP/T003553/1. Yanqi Cheng and Angelica I Aviles-Rivero greatly acknowledge support from a C2D3 Early Career Research Seed Fund and CMIH EP/T017961/1, University of Cambridge. Lihao Liu gratefully acknowledges the financial support from a GSK scholarship and a Girton College Graduate Research Fellowship at the University of Cambridge. Angelica I Aviles-Rivero acknowledges support from CMIH and CCIMI, University of Cambridge, and ESPRC capital award. Carola-Bibiane Schönlieb acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. The authors also greatly acknowledge KritiKal and Christina Runkel for their insightful discussion.

REFERENCES

[1] L. Li, F. Zhou, and X. Bai, "Infrared pedestrian segmentation through background likelihood and object-biased saliency," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 9, pp. 2826–2844, 2017.

[2] P. Lopez-Garcia, E. Onieva, E. Osaba, A. D. Masegosa, and A. Perallos, "A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 557–569, 2015.

[3] X. Li, T. Hao, X. Jin, B. Huang, and J. Liang, "Fine traffic congestion detection with hierarchical description," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 439–24 453, 2022.

[4] 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*, 2016, pp. 3213–3223.

[5] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.

[6] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1743–1751.

[7] Y. Li, J. Cai, Q. Zhou, and H. Lu, "Joint semantic-instance segmentation method for intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[8] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3448–3460, 2022.

[9] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 31, no. 3, pp. 539–555, 2008.

[10] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8797–8806.

[11] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[12] J.-P. Jodoin, G.-A. Bilodeau, and N. Saunier, "Urban tracker: Multiple object tracking in urban mixed traffic," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014, pp. 885–892.

[13] E. Strigel, D. Meissner, F. Seeliger, B. Wilking, and K. Dietmayer, "The ko-per intersection laserscanner and video dataset," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014, pp. 1900–1901.

[14] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura, "Understanding traffic density from large-scale web camera data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5898–5907.

[15] C. H. Bahnsen and T. B. Moeslund, "Aau rainsnow traffic surveillance dataset," 2018. [Online]. Available: https://www.kaggle.com/dsv/105294

[16] Z. Luo, F. Branchaud-Charron, C. Lemaire, J. Konrad, S. Li, A. Mishra, A. Achkar, J. Eichel, and P.-M. Jodoin, "Mio-tcd: A new benchmark dataset for vehicle classification and localization," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5129–5141, 2018.

[17] C. Snyder and M. Do, "Data for streets: A novel camera network dataset for traffic flow," *University of Illinois at Urbana-Champaign: Urbana/Champaign, IL, USA*, 2019.

[18] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf v2x cooperative perception dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 668–22 677.

[19] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7699–7707.

[20] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[21] B. Bae, Y. Liu, L. D. Han, and H. Bozdogan, "Spatio-temporal traffic queue detection for uninterrupted flows," *Transportation Research Part B: Methodological*, vol. 129, pp. 20–34, 2019.

[22] L. Li and X. M. Chen, "Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey," *Transportation Research Part C: Emerging Technologies*, vol. 76, pp. 170–188, 2017.

[23] A. Torralba, B. C. Russell, and J. Yuen, "Labelme: Online image annotation and applications," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1467–1484, 2010.

[24] 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*. Springer, 2014, pp. 740–755.

[25] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[29] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[30] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[34] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[35] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2019. [Online]. Available: http://dx.doi.org/10.1109/tpami.2019.2956516

[36] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu, "Instaboost: Boosting instance segmentation via probability map guided copy-pasting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 682–691.

[37] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[38] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6910–6919.

[39] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *arXiv*, 2021.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[42] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv preprint arXiv:1906.01916*, 2019.

[43] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1369–1378.

[44] R. Ke, A. Aviles-Rivero, S. Pandey, S. Reddy, and C.-B. Schönlieb, "A three-stage self-training framework for semi-supervised semantic segmentation," *arXiv preprint arXiv:2012.00827*, 2020.

[45] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4268–4277.

[46] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

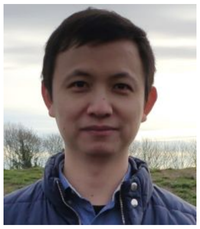
[47] Z. Wang, Y. Li, and S. Wang, "Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 826–16 835.

Zhongying Deng is a post-doctoral research associate at Cambridge Image Analysis Group, University of Cambridge. He received the Ph.D. degree from the University of Surrey in 2022 and a master's degree from the University of Chinese Academy of Sciences (UCAS) and Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), in 2019. His research interests include domain adaptation, semi-supervised learning, semantic segmentation, and medical image analysis. He serves as a reviewer for IEEE T-PAMI, IEEE T-IP, IEEE T-CSVT, IJCV, Pattern Recognition, Neurocomputing, CVPR, ICCV, ECCV, AAAI, and MICCAI. He also organised the 1st and 2nd International Workshop on Foundation Models at MICCAI.

Yanqi Cheng is currently a Ph.D. student at the Department of Applied Mathematics and Theoretical Physics, University of Cambridge. Yanqi works at the intersection of applied mathematics and machine learning for large-scale real-world problems. In particular, Yanqi focuses on designing novel knowledge-driven and data-driven models to solve inverse problems, with areas of interest including optimisation, machine learning, and mathematical modelling.

Lihao Liu received the Ph.D. degree from Department of Applied Mathematics, University of Cambridge. His research interests include Medical Image Analysis and Video Understanding. Specifically, he focus on using Machine Learning techniques to solve unsupervised medical image registration and segmentation tasks, as well as surgical video processing and video content understanding tasks.

Shujun Wang is currently an Assistant Professor in the Department of Biomedical Engineering at the Hong Kong Polytechnic University. Previously, she was a Research Associate at the University of Cambridge. Her research is in the interdisciplinary field of artificial intelligence (AI) and healthcare. She is dedicated to designing AI-driven computational methods to enable reliable decision-making models for precision medicine, covering from disease diagnosis to prognosis, and from medical image computing to multi-modal biomedical data integration. Her current and future research will facilitate personalized prognosis and treatment with multi-modal biomedical data computing from both imaging and non-imaging information and reliable machine learning-based disease diagnosis algorithms.



Rihuan Ke is a Lecturer at the School of Mathematics, University of Bristol. He was previously a research associate at the Department of Applied Mathematics and Theoretical Physics, University of Cambridge. His research interests are mathematical and machine learning methods for image analysis, inverse problems, and their applications.

Carola-Bibiane Schönlieb received the B.S. degree from the Institute for Mathematics, University of Salzburg, Salzburg, Austria, in 2004, and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2009. From 2004 to 2005, she held a teaching position in Salzburg. She held a post-doctoral position at the University of Göttingen, Göttingen, Germany, for one year. She became a Lecturer at the Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge, in 2010, where he was promoted to a Reader in 2015 and a Professor in 2018. Since 2011, she has been a fellow of the Jesus College, University of Cambridge, where she is currently a Professor of applied mathematics with DAMTP, the Head of the Cambridge Image Analysis Group, the Director of the Cantab Capital Institute for Mathematics of Information, and the Director of the Engineering and Physical Sciences Research Council Center for Mathematical and Statistical Analysis of Multimodal Clinical Imaging. Her research interests include variational methods, partial differential equations, machine learning for image analysis, image processing, and inverse imaging problems.

Angelica I Aviles-Rivero is currently an Assistant Professor at the Yau Mathematical Sciences Center at Tsinghua University. Previously, she was a Senior Research Associate at DAMTP and DPMMS, University of Cambridge. She works to bring a unique blend of applied mathematics, computational mathematics, and machine learning expertise, particularly in developing hybrid analysis techniques for large-scale real-world challenges. Her work bridges novel variational methods with modern machine learning. She regularly reviews for several journals (e.g., SIAM Imaging Science, IJCV) and conferences (e.g., MICCAI, NeurIPS, ICML, CVPR). She has also served as an Area Chair for some conferences. Her research has been highlighted, including receiving an Outstanding Paper Award at ICML 2020. She was elected as an officer for the SIAM SIAG/IS secretary position for the term 2023.