Attribute Dissection of Urban Road Scenes for Efficient Dataset Integration

Jiman Kim, Chanjong Park
Samsung Research,
Samsung Electronics
{jiman14.kim, cj710.park}@samsung.com

Abstract

Semantic scene segmentation or scene parsing is very useful for high-level scene recognition. In order to improve the performance of scene segmentation, the quantity and quality of the datasets used for deep network's learning are important. In other words, we need to consider various external environments and various variations of the predefined objects in terms of image characteristics. In recent years, many datasets for semantic scene segmentation focused on autonomous driving have been released. However, since only quantitative analysis of each dataset is provided, it is difficult to establish an efficient learning strategy considering the image characteristics of objects. We present definitions of three frame attributes and five object attributes, and analyze their statistical distributions to provide qualitative information about datasets that are to be merged. We also propose an integrated dataset configuration that can exploit the advantages of each data set for deep network learning after class matching. As a result, we can build new integrated datasets that are optimized for the scene complexity and object properties of the environment by considering the statistical characteristics of each dataset.

1 Introduction

Scene understanding requires information such as the objects that are present in the scene, their characteristics, and the relationships among them. Semantic segmentation provides information on the location and type of objects by dividing an image into regions that include predefined objects. To apply scene segmentation functions to autonomous vehicles, many road scene-centered datasets have been released. Representative datasets labeled at the pixel level are CamVid [Brostow *et al.*, 2008; 2009], Cityscapes [Cordts *et al.*, 2015; 2016], SYNTHIA [Ros *et al.*, 2016], GTA-V [Richter *et al.*, 2016], Mapillary [Neuhold *et al.*, 2017]. Papers that explain each dataset provide statistical information on the image data collection environment, area, device, the amount of images, and the relative proportions of objects. Papers [Perazzi *et al.*, 2016] that compare the scene parsing accuracy of several state-of-the-art algorithms focus on their advantages and disadvantages, rather than on the characteristics of the image data. However, the datasets include different categories defined and different image characteristics of the instances in them, so efficient learning of the deep network requires detailed analysis of various attributes of the data. For example, some datasets may contain many small objects, some datasets may contain densely distributed objects, and other datasets objects may show deformed objects. By using these image characteristics, a network can be developed that has excellent specialization for a specific environment and object, or that has excellent generality in a general environment by combining characteristics.

In this paper, we analyze the CamVid, Cityscapes, SYNTHIA, GTA-V, and Mapillary datasets quantitatively, based on two criteria. First, we analyze image-centric criteria such as the average number of categories in the image, the average number of objects, and the average proportion of the image that is a road region. Second, we analyze object-centric criteria, such as the average spatial density of objects in the image, their average size, average shape, average color, and average position in the image. This analysis provides we good insight into ways to train deep networks. We also propose a new set of integrated classes that can be used commonly among datasets, and a method to construct an integrated dataset. The integrated dataset contributes to improve the generality of the deep network by including various road environments and object characteristics. This paper has the following structure. Section 2 introduces papers related to published datasets. Section 3 summarizes each dataset and proposes image- and object-centric attributes. Section 4 proposes a new integrated dataset by performing class alliance. Section 5 provides a detailed comparative analysis of the proposed attributes, and suggests insights for constructing integrated datasets. Section 6 summarizes all findings and contributions of this paper.

2 Related Work

Road scene-centric public datasets for pixel-level semantic segmentation have been released (Table 1, Fig. 1) with papers that explain them. CamVid [Brostow *et al.*, 2008; 2009] was the first dataset that had semantic labels of object class for each pixel. The images were images acquired

from the perspective of a driving automobile; they are divided into 32 semantic classes with manually-annotated labels. To reduce the effort of the person who must label objects, the authors proposed joint tracking of keypoints and regions; this method propagates the label information to the 100 subsequent frames. The set includes the camera's 3D pose in each frame and is has a software tool that users can use to label their additional images. The publicly available datasets [Martin *et al.*, 2001; Fei-Fei *et al.*, 2006; Bileschi, ; Shotton *et al.*, 2006; Smeaton *et al.*, 2006; Griffin *et al.*, ; Yao *et al.*, 2007; Russell *et al.*,] before CamVid have polygon-level labels not pixel-level and they were obtained from fixed CCTV-style cameras. The paper provides statistical information on the percentage of the objects in the image for each sequence and the number of occurrences.

Cityscapes [Cordts *et al.*, 2015; 2016] is a large-scale dataset that includes complex real-world urban scenes. Cityscapes has image data that are labelled at the pixel level and instance level. The images were acquired from 50 cities to include a variety of road environments. The authors provided the results of statistical analysis between datasets by grouping 30 classes into eight categories. The results describe the number and relative ratios of annotated pixels of each class, annotation density, the distribution of the number of instances related with traffic per an image, and distribution of the number of vehicle according to the distance.

SYNTHIA [Ros *et al.*, 2016] is a dataset of synthetic images obtained from a virtual world (Unity development platform) [Technologies,]. The images were captured from multiple view-points by using two multi-camera with four monocular cameras that are mounted on a virtual car. The images include different seasons, weather and illumination conditions. The captured images were annotated with 11 predefined classes. In experiments, the authors showed that combining a real dataset and SYNTHIA dataset dramatically increases the accuracy of semantic segmentation.

Grand Theft Auto V (GTA-V) [Richter *et al.*, 2016] consists of images captured from a computer game. The authors proposed a method to quickly generate semantic label maps. Each image is automatically divided into patches, then merged using MTS (mesh, texture, shader). For each patch, a semantic class is manually assigned. Within a brief space of time, these methods yield far more pixel labeling than previous datasets. When virtual images generated by the proposed method were added to real-world images, segmentation accuracy was greatly improved even though a large number of real-world images are replaced by virtual images. The related paper provided statistical information on the number of labeled pixels, annotation density, and the time and speed of labeling.

Mapillary [Neuhold *et al.*, 2017] is the dataset that contains the most real-world images, and the largest number (66) of categories to consider. The images were captured by differently-experienced photographers on various imaging devices. The considered cities are Europe, North and South America, Asia, Africa and Oceania and the scenes include urban, countryside, and off-road scenes. Manual annotation was performed using polygons by specialized image annotators. Statistical analyses performed by the authors include image resolution, focal length, number of images taken with the devices used for image acquisition, region where the image was acquired, number of instances per class, number of objects per image, number of traffic regulation objects per image, and number of traffic participants per image.

These papers mainly analyzed how often each class appeared in each image. They also focused on the number and proportions of major classes that are closely related to traffic. If the detail and organization of the information on the frame and object side can be obtained, they would improve the learning efficiency of deep networks. Therefore, in this work, we perform detailed characterization of each dataset to derive insight. Also, to enable simultaneous use of two or more datasets with different class numbers and types, we define a common usable class and propose a way to efficiently combine datasets.

3 Attribute Analysis

Scene segmentation or scene parsing at the pixel level to extract the boundaries of many kinds of objects solves object detection and localization simultaneously. To achieve high accuracy of pixel-level segmentation, large-scale datasets are required; they must include a variety of shape and appearance variations of static objects (backgrounds) and dynamic object (foreground). Therefore, construction of data sets that focus on road scenes has increased resolution and number of images, and to an increased variety of environments.

Trends of Dataset. The constructed and released datasets for the same goal are described in Table 1. Higher resolution and larger amount of images are two common trends in constructing road scene-centric dataset. The increase in the resolution of the collected images is closely related to pixel-level accuracy. In addition, virtual environment tools have been used to collect a large number of images in a short time. In particular, the volume of real images in the Mapillary dataset was increased sharply by a community-led service to share street-level photographs. Diversification of the environments that the images represent has yielded datasets from different regions and environments, and recently-constructed datasets include increasing diversity of regions and of environmental conditions. Real images are much more difficult to obtain than virtual images, and the continental, regional, and environmental conditions in which the images are acquired has become very diverse in the Cityscapes and Mapillary datasets. Each dataset has different properties (Table 1). The CamVid dataset was the dataset that focused on road scenes; it contains many lane-clear highway images. The Cityscapes dataset includes images that are specific to European urban scenes, The SYNTHIA dataset has many virtual images with multiple seasons. GTA-V dataset's virtual images are extremely realistic, and its effects are richly controllable. The Mapillary dataset contains the largest number of images collected in the broadest variety of regions.

Attribute Definition. We defined two types of criteria to specifically analyze the attributes from an image frame (still-shot) perspective, with the exception of the collection method and environment from five representative datasets for road scene segmentation. One is the metrics for each image frame,

Name	Year	Class	Resolution	Image (Training/Validation/Test)		Description
				RGB	GT	
CamVid	2008, 2009	32	960 × 720 Unified Size	701 (367/101/233)	701 (367/101/233)	Real Image, Normal Light/Weather
Cityscapes (fine)	2015, 2016	30	2,048 × 1,024 Unified Size	5,000 (2,975/500/1,525)	3,475 (2,975/500/-)	Real Image (50 Cities), Normal Light/Weather
SYNTHIA (cityscapes)	2016	23	1,280 × 760 Unified Size	9,400 (9,400)	9,400 (9,400)	Virtual Image, Dynamic Light/Weather
GTA-V	2016	34	1914 × 1052 Unified Size	24,966 (24,966)	24,966 (24,966)	Virtual Image, Dynamic Light/Weather
Mapillary	2017	66	3,420 × 2,480 Averaged Size	25,000 (18,000/2,000/5,000)	20,000 (18,000/2,000/-)	Real Image (6 Continents), Dynamic Light/Weather
Integration	2018	30	2,048 × 1,024 Unified Size	65,067 (41,961/6,038/17,068)	58,542 (41,961/6,038/10,543)	Real/Virtual Image, Dynamic Light/Weather

Table 1: Quantitative summary of various datasets for semantic scene segmentation. The number of classes of each dataset includes 'void' class. GTA-V and Mapillary contain images of different sizes. The images of SYNTHIA and GTA-V are not divided for training, validation, and test. Recently released datasets include more classes and higher resolution images. Also, many virtual tools that simulate road scene environment were released to increase the number of virtual images in various conditions, because collecting of real images has high cost. Integrated dataset is based on the Cityscapes dataset (Table 3) but has many more images.

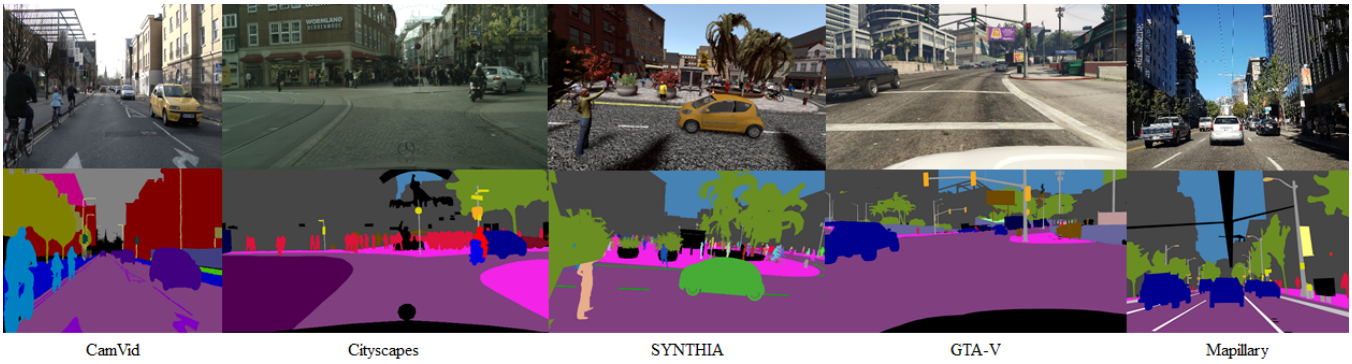

Figure 1: Example images of five datasets. First row: randomly-selected original images (RGB) from each dataset; second row: ground-truth images that correspond to each original image. CamVid and Mapillary datasets provide ground-truth color values for each class; Cityscapes, SYNTHIA, and GTA-V datasets also provide label images with different integer index values assigned to pixels of each class. Each dataset includes different types of urban road scenes, and various types and sizes of objects.

and the other is the object metrics (Table 2). For each metric, we computed the mean value and its distribution. Analyzed information of image complexity and object diversity can be utilized to construct new datasets with different goals. Metrics to explain scene complexity from an image frame perspective are class diversity, object density, and road diversity. Class diversity means a distribution of the number of all classes appearing per frame, and the diversity of objects in a scene can be determined. Object density means a distribution of the total number of all objects appearing per frame, and explains how many object concentrate in a scene. Road diversity means a distribution of the relative ratio of road area and building area. We can estimate the scene as highway or city center from the road diversity. Metrics to explain object's extrinsic variability of each class from an object perspective are class density, object size variability, object shape variability [Collins *et al.*, 2001], object intensity (one channel color) variability, and geometrical position variability. Class density means a distribution of the number of objects of a specific class per frame, and represents the number of objects of the class that exist in a scene. Object's size/shape/intensity(one channel color) variability means distributions of object's external appearances of a specific class, and it shows how appearances of object varies in scenes. Geometrical position variability means a distribution of object positions in scenes; it explains which positions are major regions of interest in scenes.

4 Dataset Integration

Class Alliance for Scene Integration. Each country has different object attributes, road surface properties, rules of the road, traffic patterns (traffic signs and signals), and climate conditions. If the image characteristic used for learning and testing of deep neural networks are different, this diversity is a major cause of degradation of the accuracy of semantic scene segmentation. Quick construction of a dataset that includes all varieties of road scenes is a real challenge, but it is the most reasonable way to efficiently integrate the released datasets collected in different regions. Models created with

Attributes		Definition	Explanation
Frame Attributes	Class Diversity	$\frac{1}{N}\sum \#(Classes)$	how diverse objects exist in a scene
	Object Density	$\frac{1}{N}\sum \#(Objects)$	how many object concentrate in a scene
	Road Diversity	$\frac{1}{N}\sum max(\frac{Area_R\text{-}Area_B}{Area_R}, 0)$	how diverse road scenes exist in a scene
Object Attributes	Class Density	$\frac{1}{N}\sum \#(Objects_i)$	how many objects of the class exist in a scene
	Object Size Variability	$\frac{1}{N}\sum Size_i$	how the size of object varies in scenes
	Object Shape Variability	$\frac{1}{N}\sum Dispersedness_i$	how the shape of object varies in scenes
	Object Intensity Variability	$\frac{1}{N}\sum Intensity_i$	how the intensity of object varies in scenes
	Geometrical Position Variability	$\frac{1}{N}\sum {X^c Y^c}_i$	where is the object most likely to appear

Table 2: Summary of frame/object attributes. We propose three frame attributes and five object attributes. Frame attributes are used to analyze scene complexity of each dataset. Object attributes are used to understand the extrinsic variability of objects of each class. N: total number of image frames of each dataset; : number of corresponding objects. $\text{Area}_R and Area_B$: $areaofroadandbuildingineachimageframe, respectively. If \text{Area}_R < Area_B$, then road diversity is set to 0.

this integrated dataset that fully represents the diversity of road scenes can be very good choice of initial model needed to create a model optimized for a specific environment.

To build an integrated dataset, we refer consider 30 classes and 8 groups of Cityscapes dataset. The author of the Cityscapes dataset selected 30 classes in the road scene and grouped them semantically by referring to WordNet [Miller, 1995]. Cityscape is a reasonable basis for performing matching between classes of datasets because it has about the average number of classes among the five datasets considered here. We performed a semantic comparison between the classes considered by each dataset, from 11 to 66, with the classes defined in Cityscapes dataset (Table 3). Usually, fewer than 30 classes from each of the other datasets correspond to the superclass of Cityscapes's classes, but 1:1 matching with one of the most suitable Cityscapes's class was accomplished without any division. If the number of classes is ¿ 30, they are usually subclasses of the 30 classes in Cityscapes, so we have matched the classes in other datasets to the semantically-higher class of the Cityscapes classes. In this way, the images in each dataset can be unified to construct a large-scale dataset of N images, which represent ¿ M urban environments. For the integrated dataset based on common classes, we performed image-based and object-based analysis (Section 3), and observed how the characteristics changed (Section 4)

Sampling Methods for Image Integration. To build an integrated dataset, the images from each dataset must be mixed appropriately after the classes are unified. In this paper, we propose five image-sampling methods to combine images from different datasets. The second and third are aimed at balancing the numbers of images between datasets, and the fourth through sixth are aimed at building an integrated dataset that is optimized for a specific purpose.

- Naive Integration: The simplest method to integrate datasets is to merge all the images in datasets into a unified image size. This method retains the original image data of each dataset, but naturally, the dataset characteristics with large image quantities become dominant.
- Randomized Undersampling: Undersampling is one of the most commonly used methods to match the number of images among classes or among datasets [Buda *et al.*, 2017; Haixiang *et al.*, 2016; Drummond and Holte, 2003]. It is a way to randomly select a number of images from each dataset that is equal to the number of images in the smallest datasets. The integrated dataset consists of $\min(N_m) \times M$ images, where N_m means the number of images of mth dataset and M is the number of datasets. Undersampling is an intuitive and easy-to-use sampling method, but it has the drawback of not being able to exploit the large amount of residual images.
- Randomized Oversampling: Oversampling is another frequently-used method [Buda *et al.*, 2017; Haixiang *et al.*, 2016; Janowezyk and Madabhush, ; Jaccard *et al.*, 2017]. It is a way to randomly select images, and allows duplicates in each dataset that has the largest number of images. The integrated dataset consists of $\max(N_m) \times M$ images, where N_m is the number of images in the mth dataset and M is the number of datasets. Overfitting may occur in some cases [Chawla *et al.*, ; Wang *et al.*,], but variations exist to reduce this problem [Chawla *et al.*, ; Han *et al.*, ; Shen *et al.*, 2016]. Oversampling is the most common method to get the largest number of images for training.
- Diversity Oriented Sampling: This method means that the larger the number of average classes contained in the image of the dataset, the more images are reflected in the integrated dataset. The integrated dataset consists of $\sum_{m=1}^{M}(w_m^{CD} \times \max(N_m))$ images, where $w_m^{CD} = CD_m / \sum_{m=1}^{M} CD_m$ is the weight of the mth dataset, CD_m is the average class density of the mth dataset, and M is the number of datasets. The maximum number of images that can be sampled is limited to $\max(N_m)$. This sampling method enables construction of an integrated dataset that is optimized on the variety of static/dynamic backgrounds in a target environment. This method can be used to construct an integrated dataset that best adapts to the static/dynamic background variety of the target environment. As a variation of diversity-oriented sampling, an integrated dataset may by constructed by selecting only images including classes that are more than the average number desired by the user.
- Density Oriented Sampling: An integrated dataset can be built that is optimized for the object density of

Cityscapes: Base	CamVid	SYNTHIA	GTA-V	Mapillary
01. Road	Road, Road Shoulder, Lane Markings Drivable	Road, Lanemarking	Road	Road, Pothole, Lane, Service Lane, General Lane Marking
02. Sidewalk	Sidewalk	Sidewalk	Sidewalk	Sidewalk, Pedestrian Area, Curb, Curb Cut
03. Parking	Parking Block	Parking Slot	-	Parking
04. Rail Track	-	-	-	Rail Track
05. Person	Child, Pedestrian	Pedestrian	Person	Person
06. Rider	Bicyclist	Rider	Rider	Bicyclist, Motorcyclist, Other Rider
07. Car	Car	Car	Car	Car
08. Truck	SUV/Pickup Truck	Truck	Truck	Truck
09. Bus	Truck/Bus	Bus	Bus	Bus
10. On Rails	Train	Train	Train	On Rails
11. Motorcycle	Motorcycle/Scooter	Motorcycle	Motorcycle	Motorcycle
12. Bicycle	-	Bicycle	Bicycle	Bicycle
13. Caravan	-	-	-	Caravan
14. Trailer	-	-	Trailer	Trailer
15. Building	Building	Building	Building	Building
16. Wall	Wall	Wall	Wall	Wall
17. Fence	Fence	Fence	Fence	Fence
18. Guardrail	-	-	Guardrail	Guardrail, Barrier
19. Bridge	Bridge	-	Bridge	Bridge
20. Tunnel	Tunnel	-	Tunnel	Tunnel
21. Pole	Column/Pole	Pole	Pole	Pole, Utility Pole, Street Light, Traffic Sign Frame
22. Pole Group	-	-	-	-
23. Traffic Sign	Sign/Symbol	Traffic Sign	Traffic Sign	Traffic Sign Front
24. Traffic Light	Traffic Light	Traffic Light	Traffic Light	Traffic Light
25. Vegetation	Tree, Vegetation Misc	Vegetation	Vegetation	Vegetation
26. Terrain	-	Terrain	Terrain	Terrain, Sand
27. Sky	Sky	Sky	Sky	Sky
28. Ground	Non-Drivable	-	-	Crosswalk Plan, Crosswalk Zebra, Water
29. Dynamic	Animal, Cart/Luggage/Pram, Other Moving	-	-	Bird, Animal, Trash Can, Boat, Wheeled Slow, Other Vehicle
30. Static	Archway, Misc Text, Traffic Cone, Void	Road-Work, Void	Ego Vehicle, Static, Void	Ego Vehicle, Car Mount, Mountain, Snow, Banner, Billboard, CCTV Camera, Traffic Sign Back, Catch Basin, Manhole, Fire Hydrant, Bench, Bike Rack, Junction Box, Mailbox, Phone Booth, Unlabeled

Table 3: Class matching table of Cityscapes dataset (8 categories, 30 classes) and other datasets: Object (01-24) and Nature (25-30). We assigned the classes of four datasets (CamVid, SYNTHIA, GTA-V, Mapillary) to 30 classes by referring the class definition of the Cityscapes dataset. All datasets share most of Cityscapes's classes. Especially, important classes (road, human, vehicle, traffic sign/light) are essentially included in all datasets. GTA-V does not have class information, so we manually checked class names (we could not find 8 classes). Mapillary dataset divides the 'static' class into many sub-classes.

the target environment. An integrated dataset that closely represents the image of a dense dataset consists of $\sum_{m=1}^{M}(w_m^{OD} \times \max(N_m))$ images, where $w_m^{OD} = OD_m / \sum_{m=1}^{M} OD_m$ is the weight of mth dataset, OD_m is the average object density of mth dataset, and M is the number of datasets. A modified method constructs an integrated dataset by selecting only images that correspond to more than an average density that user desired.

- Target Oriented Sampling: If the goal is to extract a specific target object accurately, the integrated dataset must have images that contain as many of the target objects as possible in one scene. In addition, construction of a training set with uniform distribution on each object attribute enables generation of a model that is insensitive to attributes of the target object. In addition, a model can be constructed that is insensitive to changes in object attributes, if the training dataset is built by selecting images so that each attribute has an even distribution, that is, as many variations as possible.

5 Experiments

Six datasets were used for analysis of frame and object attributes. These datasets were CamVid (701 images), Cityscapes (3,475 images), SYNTHIA (9,400 images), GTA-V (24,966 images), Mapillary (20,000 images), and the integrated dataset proposed in Section 4 (58,542 images). Frame attributes were evaluated for each image frame, regardless of class, and object attributes were evaluated individually for each class in each dataset.

The three frame attributes indicate the number and variety of objects that are present in the image frames of each dataset. First, the attribute values were individually calculated from the image frames (Table 2) and the distribution of each attribute was expressed as a histogram (Fig. 2, Fig. 3). To compare the distribution's variance, we normalized it to [0, 1] for each histogram by dividing each bin by the maximum value of bins. We use the naïve integration method to construct an integrated dataset. Image size or resolution do not significantly affect frame attributes, but image size can affect object attributes. However, the variance of attribute's distribution, and the image resolution itself are characteristics of the image-acquisition devices used in each dataset, so we displayed the original distributions without performing image size normalization, then compared their variances by considering the absolute size range. We used the target-oriented sampling method to construct another integrated dataset and computed the object attributes based on Cityscapes dataset's image size.

5.1 Relative Analysis of Frame Attributes

Class Diversity. The Mapillary dataset contains the largest number of classes per frame on average, but this result occurs because the number of classes defined in the Mapillary dataset is much higher than for any other dataset. If we unify the number of classes as the minimum, and calculate the rela-

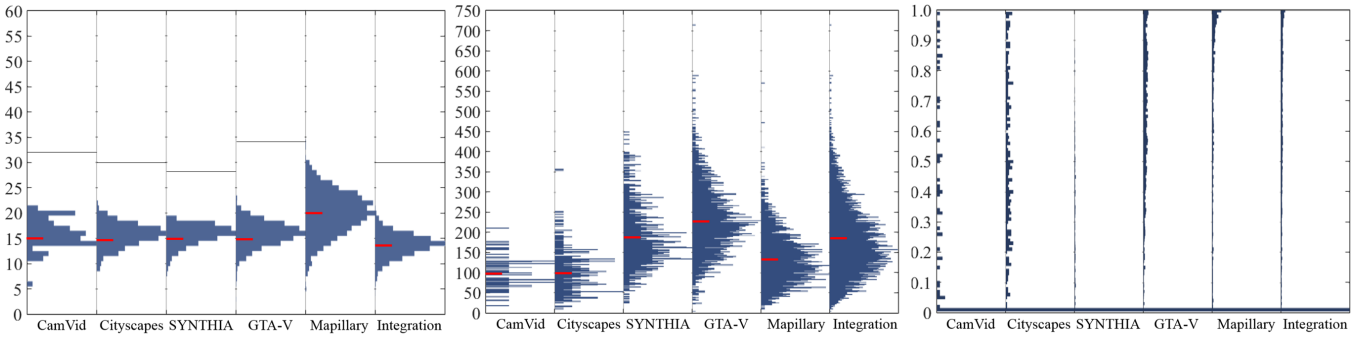

Figure 2: Histogram of frame attributes. (a) Distribution of class diversity for the six data sets, including the integrated dataset. Horizontal black line: number of classes of each dataset; Red line: position of the mean value. Considering that the number of classes in a Mapillary dataset is twice as many as other datasets, they all exhibit a similar variance. (b) Distribution of object density. Images in the virtual image dataset generally contain more, and more-varied objects on than the other datasets. (c) Distribution of road diversity. Most datasets contain images of a road area that is smaller than the building area on average; i.e., many urban scenes with buildings rather than highways or countryside. Images in the Cityscapes dataset have the most diverse road area.

tive ratio, the SYNTHIA dataset includes the largest number of classes per frame on average, and the remaining data sets include an average of 15 classes per frame. The variance of the GTA-V dataset is the largest, which means that the classes present in one frame are the most diverse from the smallest to the largest.

Object Density. The vertical range of object density was larger than expected; the reason is that the segmentation label is also assigned to all small segments which are only part of an object. Average Object Density varies slightly among datasets. The GTA-V dataset contains an average of 230 object segments. On average, datasets that contain virtual images contain more objects in a scene than datasets that contain real images. Thus, we can utilize a virtual-image dataset to increase the complexity of the scene.

Road Diversity. When the road diversity is calculated, it is set to 0 when no road segment is present, or the building area is larger than the road area. Most of the images have road diversity = 0 (Fig. 2(c)); i.e., many road scenes include numerous buildings, or do not have an area that is labeled as road. This result indicates that all datasets contain many images that had been captured in urban environment rather than on the highway. Except for the zero bin, the Cityscapes dataset evenly covers the roadscapes of various areas.

Integrated Dataset. In class diversity, our integrated dataset shows the most typical normal distribution, in which the mean value is in the middle of the number of classes. Most experiments assume that the normal distribution is the most common. In class density, the integrated dataset is close to the normal distribution after GTA-V, and the value of each point in the distribution is high because the number of images is much larger in an integrated dataset than in each of the component datasets. This observation means that the integrated dataset that we proposed is more advantageous than the component datasets to learn models for scene segmentation. The road diversity of the integrated dataset represents the common characteristics of the other datasets. In summary, the properties of image complexity of the integrated dataset is not biased to one side, but shows about the average characteristics of the five component datasets. Depending on the complexity of the field in which the dataset is to be applied, the weight of the dataset that has the corresponding complexity can be increased to create a new integrated dataset that is optimized for a specific research field. For example, if the scene includes a complex environment where a large number of objects appear, the weights can be increased for virtual image datasets such as SYNTHIA and GTA-V.

5.2 Attribute Analysis of Important Objects

To analyze the object attributes, we selected four objects that are important in the driving situation: persons and cars as objects that generate the most serious damage in a collision; and traffic lights and traffic signs that provide the most essential information for driving.

Class Density. The density distributions of persons and traffic lights were even in SYNTHIA and GTA-V, and density distributions of car and traffic sign were similar in most datasets. Class Density has a higher average density value in virtual image datasets than real image datasets, as is true of object density of frame attributes.

Object Size Variability. Cityscapes and Mapillary datasets include variously-sized instances of people, vehicles, traffic lights, signs. It is useful to use the two datasets for segmentation that is less sensitive to the scale change of the object.

Object Shape Variability. Shape complexities of the important objects do not change much, regardless of dataset. Cityscapes dataset and Mapillary dataset have large variances in size, but small variance of shape. This result means that the morphological characteristics of each object do not depend on the size or scale of the image. For extremely small or large instances, the detail of appearance can vary widely, and most datasets include histogram bins for such cases. Sometimes, relatively large traffic lights and traffic signs appear in virtual image datasets.

Object Intensity Variability. Instead of considering each of the RGB values, we consider the intensity value by convert-

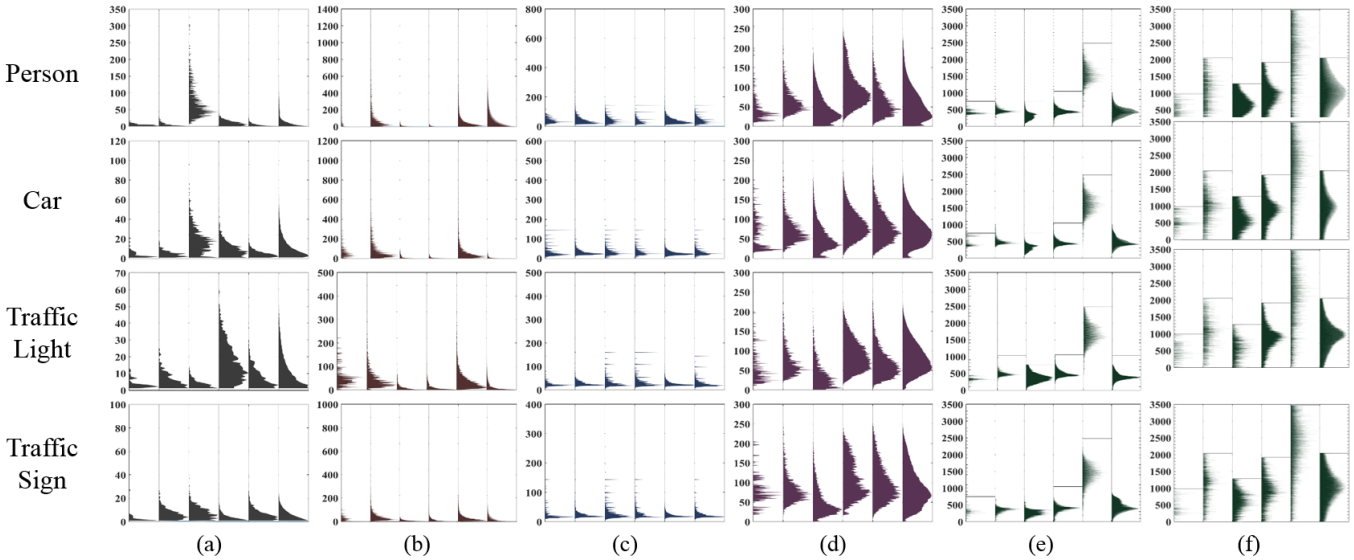

Figure 3: Histograms of object attributes for important objects. (a) Distributions of class density. Each object has a diversity of densities in a each dataset. (b) Distributions of object size variability. The Cityscapes dataset and the Mapillary dataset contain objects of the most diverse scales. (c) Distributions of object shape variability. The variability of shape of object in all datasets is not large; i.e., few images contain extremely large or small objects. (d) Distributions of object intensity variability. A more recent data set shows a richer color for each important object. (e) Distributions of geometrical position variability. Row, horizontal line: range of image height. (f) Distributions of geometrical position variability. Col: horizontal line: range of image width. All important objects exist in various width ranges in most datasets. Analysis of the integrated data set is described in Section 5.

ing all images to gray images. The average intensity values are calculated in each object region and represented as a histogram. For all important objects, the SYNTHIA, GTA-V, and Mapillary datasets contain instances of much more color than the CamVid and Cityscapes datasets. The difference occurs because SYNTHIA, GTA-V, and Mapillary datasets were constructed more recently than Camvid and Cityscape, and therefore had more images and more environmental conditions. SYNTHIA, GTA-V dataset's tool can change various attributes of objects and backgrounds, and Mapillary dataset was photographed on six continents, so colors vary widely.

Geometrical Position Variability: Row. The last two columns of Fig. 3 show distributions that represent the row and column (col) in which each object appears in the image. The horizontal lines of histograms represent the image resolution range (height, width) of each dataset. Persons and traffic signs are mainly located at the middle height of the image, whereas cars and traffic lights are mainly located in the upper part of the image. The SYNTHIA dataset contains more objects at various heights than do other datasets.

Geometrical Position Variability: Column. In all datasets, most objects exist in various locations from left to right of the image. In particular, the Cityscapes dataset and the Mapillary dataset include many cases in which objects are uniformly present in all column ranges, but the range of rows in which important objects exist is limited, but the col range is relatively various. A dataset with an even distribution of the locations of objects implies a diversity of situations or scenarios.

Integrated Dataset. An integrated dataset distribution that is within the range of characteristics of the component datasets. This characteristic is true for the four object attributes (density, size, shape, intensity) in the integrated dataset, so it is much more useful that the component datasets for training bigger models, because the number of objects contained is much larger than in those. In the integrated dataset, the spatial position of objects within an image is more uniform, and the absolute number of objects in all horizontal and vertical positions are much larger, than in the component datasets. To build a specialized integrated dataset with a specific range of density, size, shape, intensity, and position values for other objects of interest, including important objects, the ratio of items from each dataset can be adjusted appropriately. For example, if the goal is to segment human regions reliably regardless of size and color, the ratio of the Mapillary dataset in the integrated dataset can be increased.

6 Conclusion

Published datasets for use in semantic scene segmentation have different characteristics, such as the number of classes that have been defined and labeled, the image size, the range of regions in which the images were obtained, the realism of the graphic, and the diversity of the landscapes. Therefore, to learn a deeper neural network, a many images that include various characteristics should be acquired. In this paper, we compare the basic information of five representative datasets, then analyzed the distribution characteristics by defining three frame attributes and five object attributes. We also performed class matching to construct new datasets that incorporate these five datasets. Statistical results show that the image complexity of the virtual image dataset (SYN-

THIA, GTA-V) is relatively higher than that of the real image dataset, and that the Cityscapes dataset includes a variety of road scenes. In addition, for certain important objects, the datasets with flat distribution ranges are different for each attribute, so the proportional contribution of each dataset in the integrated dataset should be optimized to best match the situation of the research field to which it is to be applied. In the future, we will analyze how the method of constructing integrated datasets affects segmentation accuracy, and will study how to learn the deep neural network by using the integrated datasets to improve accuracy.

References

[Bileschi,] S. Bileschi. Cbcl streetscenes: towards scene understanding in still images. Technical report, MIT.

[Brostow *et al.*, 2008] G J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. 2008.

[Brostow *et al.*, 2009] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object class in video: a high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[Buda *et al.*, 2017] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. 2017.

[Chawla *et al.*,] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16.

[Collins *et al.*, 2001] R. T. Collins, A. J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, 2001.

[Cordts *et al.*, 2015] M. Cordts, M. Omran, S. Ramos, T. Scharwachter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset. 2015.

[Cordts *et al.*, 2016] M. Cordts, M. Omran, S. Ramos, T. Scharwachter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. 2016.

[Drummond and Holte, 2003] C. Drummond and R. C. Holte. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. 2003.

[Fei-Fei *et al.*, 2006] I. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[Griffin *et al.*,] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, Caltech.

[Haixiang *et al.*, 2016] G. Haixiang, L. Yijing, J. Shang, G. Mingyun an dH. Yuanyue, and G. Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73(1):220–239, 2016.

[Han *et al.*,] H. Han, W. Y. Wang, and B. H. Mao. Borderline-smote: a new over sampling method in imbalanced data sets learing. *Advances in Intelligent Computing*.

[Jaccard *et al.*, 2017] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin. Detection of concealed cars in complex cargo x-ray imagery using deep learning. *Journal of X-Ray Science and Technology*, 25(3):323–339, 2017.

[Janowezyk and Madabhush,] A. Janowezyk and A. Madabhush. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7.

[Martin *et al.*, 2001] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. 2001.

[Miller, 1995] G. A. Miller. Wordnet: a lexical database for english. volume 38, pages 39–41, 1995.

[Neuhold *et al.*, 2017] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. 2017.

[Perazzi *et al.*, 2016] F. Perazzi, J. P. Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. S. Hornung. A benchmark dataset and evaluation methodology for video object segmentation. 2016.

[Richter *et al.*, 2016] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: ground truth from computer games. 2016.

[Ros *et al.*, 2016] G. Ros, L. Sellart, J. Materzynska, D. Vazques, and A. M. Lopez. The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. 2016.

[Russell *et al.*,] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1).

[Shen *et al.*, 2016] L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. 2016.

[Shotton *et al.*, 2006] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance shape and context modeling for multi-class object recognition and segmentation. 2006.

[Smeaton *et al.*, 2006] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. 2006.

[Technologies,] U. Technologies. Unity development platform. Technical report.

[Wang *et al.*,] K. J. Wang, B. Makond, K. H. Chen, and K. M. Wang. A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing*, 20.

[Yao *et al.*, 2007] B. Yao, X. Yang, and S. C. Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. 2007.