Large scale labelled video data augmentation for semantic segmentation in driving scenarios

Ignas Budvytis¹, Patrick Sauer², Thomas Roddick¹, Kesar Breen¹, Roberto Cipolla¹
¹Department of Engineering, University of Cambridge, UK
²Toyota Motor Europe, Belgium
{ib255@eng.cam.ac.uk,patrick.sauer@toyota-europe.com,tr346@eng.cam.ac.uk,kesar.breen@gmail.com,rc10001@eng.cam.ac.uk}

Abstract

In this paper we present an analysis of the effect of large scale video data augmentation for semantic segmentation in driving scenarios. Our work is motivated by a strong correlation between the high performance of most recent deep learning based methods and the availability of large volumes of ground truth labels. To generate additional labelled data, we make use of an occlusion-aware and uncertainty-enabled label propagation algorithm [8]. As a result we increase the availability of high-resolution labelled frames by a factor of 20, yielding in a 6.8% to 10.8% rise in average classification accuracy and/or IoU scores for several semantic segmentation networks.

Our key contributions include: (a) augmented CityScapes and CamVid datasets providing 56.2K and 6.5K additional labelled frames of object classes respectively, (b) detailed empirical analysis of the effect of the use of augmented data as well as (c) extension of proposed framework to instance segmentation.

1. Introduction

Semantic segmentation is one of the most important sub-problems of autonomous driving. Its progress has been accelerated by the developments in the state-of-the-art in image classification [15] [52] and advances in training and inference procedures (e.g. dropout or batch normalisation) as well as architectural innovation in deep learning in general. However, in contrast to image classification and some deep learning lead problems of computer vision, semantic segmentation (especially for autonomous driving) still operates on limited size datasets which do not exceed 5000 labelled frames [10]. As labelling by hand takes approximately 1 hour per single frame, alternative methods for obtaining dense labelled data for semantic segmentation must be employed.

In this work we propose using a simplified version of the label propagation algorithm of [8] in order to increase the quantity of available ground truth labels by an order of magnitude. The chosen label propagation algorithm handles occlusions and label uncertainty elegantly, which is essential in order to avoid generating erroneous labelled data.
We evaluate the benefits of our proposed data augmentation procedure on two standard datasets for semantic segmentation: CityScapes [10] and CamVid [6]. We observe that for many models the use of augmented data leads to a significant increase in performance, and even in the worse case does not lead to a reduction larger than 2%. For some networks, such as E-Net [28], Dilation Network [36], SegNet [2] and PSPNet [37] an increase of 10.8%, 8.2%, 6.6% and 4.1% in average classification accuracy and/or IoU score is observed - significantly larger than previously reported [26]. Qualitatively the most notable improvement is in increase of class accuracy of smaller classes as illustrated in Figure 1 and reported in Section 5.

We also extend our data augmentation framework and corresponding analysis to to instance segmentation (see Figure 1 and Section 5.5). As a part of our evaluation effort we include the CamVid-Instance dataset consisting of instance labels for people and cars from the original CamVid dataset.

We proceed with a brief overview of related work in Section 2. We explain our data augmentation algorithm based on a simplified tree structured graphical model [8, 1] in Section 3. We then provide details of our experimental setup, results and analysis in Sections 4 and 5. Final remarks are provided in Section 6.

2. Related work

In this section we examine the pros and cons of various approaches aiming to directly or indirectly increase the quantity of label data.

Larger datasets. CamVid [6] and Leuven [18] datasets are among the earliest mid-sized datasets for semantic segmentation in autonomous driving, providing up to 701 hand labelled frames. Over the last four years we have seen a rapid increase of datasets covering driving scenarios such as the Ford Campus Vision and Lidar Dataset [27], KITTI Dataset [13], Daimler Urban Segmentation [31] with the largest so far being the CityScapes Dataset [10] providing 5000 hand labelled frames. The recently released Oxford Robotcar dataset [23] contains an unparalleled number of image frames on the order of 20 million, but no ground truth semantic segmentation labels are currently available.

Label Propagation. Video-based label propagation algorithms [8, 1, 26, 20] use fully or partially (e.g. paint strokes) labelled frames and propagate labels across the video via established frame to frame correspondences. Label propagation algorithms vary in (a) their methods of establishing correspondences between neighbouring frames, (b) the label inference method used and (c) the choice of unaries. Frame to frame correspondences are often established with optical flow [26] or patch matching [8]. However, as suggested in [1], unless the optical flow has high accuracy occlusion awareness, it is likely to propagate erroneous labels near occlusion boundaries. The most frequently used inference schemes include marginal posterior inference [8], sliding window inference [33], continuous MRF [14], shortest path calculation [4] and max-marginals [17]. Among these the marginal posterior inference produces the most intuitive segmentations by increasing predicted label uncertainty further away from labelled data in a chosen model (as explained in [7]). Finally, unaries are often provided via CRF [20], Random Forests [8] or other classifiers. However these require human interaction to provide highly accurate results. It is also important to note that while there are a plethora of label propagation algorithms, very few have evaluated the application of propagated labels for classifier training at scale. Exceptions include [26] and [1], however large gains in supervised training scenarios have not yet been demonstrated.

Label Transfer. Label transfer methods are very similar to label propagation methods in video, but they often use a different type of data as intermediate representation. For example, a 3D reconstruction of a scene [35, 34] can be used to achieve similar results to labels propagated in videos. Such an approach has benefits at occlusion boundaries (esp. if non-vision based 3D data is used), but may suffer lower accuracies at labelling small classes. The results of [34] are promising, however as no images or labelled frames have been disclosed, no comparison is possible. Another example of a label transfer approach is the leveraging of aerial images [25] where labelling some classes (e.g. side-walk or road) at scale turns out to be more efficient.

Artificial Data. Using artificial data [24, 9] provides another alternative for obtaining large amounts of high quality labelled data. Artificial datasets such as Synthia [30] and Virtual KITTI [12] seem to be on the rise. Despite the low cost, the direct value of such data for popular semantic segmentation benchmarks is yet to be proven. Key challenges remain obtaining photo-realistic images and modelling real world scenes, yet rapid progress has been demonstrated. Other. Other approaches include the use of weak labels [22] as well as additional sensors such as 3D point clouds [11] or point supervision [5] by humans.

3. Label Propagation

In this section, we describe the model, inference and implementation details of the label propagation method used for data augmentation.

3.1. Model

We use a simplified version of the model of [8] in which parameter learning (including class unaries) and variational inference are avoided. We define a joint probability of pixel class labels as follows:

$$P(Z) \propto \prod_{\forall k,p,j} \Psi(Z_{k+d,T_k+d,j}, Z_{k,p,j})$$  
(1)
where $Z$ is a set of discrete random variables $Z_{k,p(j)}$ taking values in the range $1..L$ corresponding to the class label of a pixel $j$ in a patch $p$ of frame $k$. Here $\Psi$ is a potential favouring same class prediction

$$\Psi(a, b) = \begin{cases} 1 - \delta, & \text{if } a = b \\ \delta, & \text{otherwise.} \end{cases}$$

(2)

Furthermore $Z_{k+d,T_{k+d,p(j)}}$ corresponds to a class label of a pixel $j$ in a patch $T_{k+d,p}$ in frame $k + d$. Here $T_{k+d,p}$ corresponds to the best matching patch of frame $k + d$ to patch $p$ in frame $k$. Finally, $d$ is a constant which builds correspondences from the current frame to the previous frame or to the next frame when set to $-1$ and $1$ respectively.

### 3.2. Inference

As each pixel is restricted to have exactly one best match for each pixel, the aforementioned joint distribution can be represented as a tree-structured factor graph. As a result, the exact inference of the marginal posterior for each variable $P(Z_{k,p(j)} = l)$ can be performed using message passing. The final per pixel class distributions are obtained by summing over distributions of overlapping pixels as follows

$$R(k, i, l) = \frac{1}{K} \sum_{s.t. p(j) = i} P(Z_{k,p(j)} = l)$$

(3)

where $K$ is a normalisation constant.

### 3.3. Implementation

The data augmentation algorithm is has two phases. During the first phase neighbouring frame correspondences are calculated by finding the highest cross-correlation score of a patch $p$ in window $W \times H$ around this patch in frame $k + d$ as detailed in [8]. During the second phase obtained correspondences are used to calculate per-pixel marginal posterior distributions as described in the previous section. Processing steps performed for semantic class labels and for instance label propagation are described below.

#### Class label augmentation.

To obtain class labels for training, we perform three steps. First, for each pixel $i$ in frame $k$, we assign the most likely class label $\arg \max_{l'} R(k, i, l')$. For pixels where the most likely label has a probability lower than a threshold $\frac{1}{L} + 0.0001$ we assign the “void” label to avoid mislabelling caused by numerical instability. Examples of labels for $d = -1$ and $d = 1$ for one sequence from the Bochum dataset (CityScapes) are presented in rows (b) and (c) in Figure 2. Unlike in [8,11], we produce the final result by taking a label image difference (i.e. assigning a class label if both frames agree and a “void” label if they disagree) as opposed to averaging the backward ($d = -1$) and forward ($d = 1$) built structures. Example comparisons between using image differencing and averaging can be found in rows (e) and (d) respectively of the left section of Figure 2. Although more pixel labels are obtained when using averaging, taking an image differ-
due to low representation of rare classes, most studies have

driving (road, side-walk, car, etc.) are provided, however

tion set. A total of 32 label types related to autonomous

bridge, UK. Labels are provided every 30 frames for the

ering a total of 10 minutes of 30Hz video of driving in Cam-

4.1. CamVid dataset

The CamVid Dataset [6] consists of 701 labelled images:

2048 × 1024 resolution. Each densely labelled frame is surrounded by
30 unlabelled adjacent video frames. We performed label
propagation for 20 surrounding frames resulting in a total
of 62475 annotated frames using the protocol described in
Section 4.1. 5% of the annotated frames were filtered manu-
ally. Key modes of failure included label leakage caused by
large object displacement and ego-motion as well as sudden
change in lighting.

To evaluate the quality of the labelling, we first sampled
9 frames from the CityScapes training dataset at random
at four different locations (−8,−4,+4 and +8) from the seed
label (see Figure 3). We then hand labelled the selected
images following the original protocol of CityScapes [10].
As shown in Table 2 all the classes had accuracy higher
than 70%, except for the bus class.

4.3. CamVid-Instance Dataset

For our third set of experiments we augmented the original
CamVid [6] dataset with instance level annotation of
people and cars. Several examples of instance labels ob-
tained are shown in Figure 2.

In order to obtain instances, we used original boundaries
of cars and people, only introducing new boundary anno-
tations where two instances of the same class were over-
lapping. The number of instances in a single frame ranged
from 0 to 27. More examples of CamVid-Instance dataset
images can be found in the supplementary material.

4.4. Model Training

In this section we provide brief training details of vari-
ous models for class and instance segmentation.

Class segmentation. For our experiments on the CamVid
dataset we trained six commonly used segmentation archi-
tectures: FCN [21], SegNet [2], Bayesian SegNet [2], E-
Net [28], Dilation Network [36] and PSPNet [37] on both

4. Experiment Setup

In this section we describe experiments on three datasets:
CamVid [6], CityScapes [10] and the novel CamVid-
Instance dataset. For all datasets we use the experiment-
protocol described in Section 4.1 unless stated otherwise.
Below we provide more details of each dataset and corre-
sponding steps taken to produce augmented labels.

4.1. CamVid dataset

The CamVid Dataset [6] consists of 701 labelled images:
367 for training, 233 for testing and 101 for validation, cov-
ering a total of 10 minutes of 30Hz video of driving in Cam-
bridge, UK. Labels are provided every 30 frames for the
training and testing set and every 2 frames for the valida-
tion set. A total of 32 label types related to autonomous
driving (road, side-walk, car, etc.) are provided, however
due to low representation of rare classes, most studies have

focused on evaluating classifiers on a subset containing 11
classes.

To obtain label propagation results we calculated the
 correspondences for every pair of neighbouring labelled
frames. We then performed label propagation for the 11
classes, representing the void class via a uniform distribu-
tion. Inference took on average of 5 seconds per frame at
full resolution of 960 × 720. The window size (W × H)
for establishing correspondences was set to 140 × 100, δ to
0.001.

We obtained propagated labels for all the images in the
train, test and validation datasets, however we use only the
labels in the training dataset in the experiments described
here.

4.2. CityScapes dataset

The CityScapes dataset consists of 5000 densely labelled
frames and 20000 coarsely labelled frames of 2048 × 1024
resolution. We obtained propagated labels for all the images in
the train, test and validation datasets, however we use only the
labels in the training dataset in the experiments described
here.
Figure 4. Three graphs in this figure compare segmentation network performance using global accuracy, class average accuracy and average IoU measures for results obtained when training on the original hand labelled CamVid data ("Hand") as well as on augmented labels obtained from uncertainty averaging ("Avg.") and label differencing ("Aug."), as explained in Section 3.3. The FCN network (left) does not benefit from additional hand labels in our experiments. Broadly this corresponds to the findings reported in [26]. On the other hand, Segnet [2] (middle) trained on augmented data produced using label differencing shows a significant increase in either class average or IoU when compared to Avg. or Hand labels. The graph on the right shows class accuracy and IoU scores on train and test datasets for E-Net [28] for the first 100 epochs. The high (and increasing) average class accuracy on training data for E-Net trained on hand labels indicates overfitting and explains the large class average and IoU score differences on the test data.

We repeated the same exercise for four models on the CityScapes dataset. We excluded Bayesian SegNet and FCN due to time constraints. Note that the aim of this work is to demonstrate the value of using augmented data and not to outperform state-of-the-art benchmarks.

5. Results

In this section we present the quantitative and qualitative results of our experiments on CamVid, CityScapes and CamVid-Instance datasets.

5.1. CamVid Dataset

Figure 4 illustrates the evolution of test accuracies during training of FCN [21], SegNet [2] and E-Net [28] on the CamVid dataset with and without data augmentation. It highlights the advantage of using augmented labels obtained by frame differencing as opposed to uncertainty averaging. It also indicates the potential of using augmented data to prevent overfitting. See the caption of Figure 4 for more details.

Table 1 compares six different state-of-the-art-models trained on hand labelled and on augmented data. Per class accuracies as well as average classification accuracy, global accuracy and mean intersection over union (mean IoU) scores are reported. Note that matching previously reported accuracies is not straightforward hence we include an evaluation of our own. Each network was trained either until convergence or for an equal number of iterations with its counterpart.

Overall we have found that using augmented labels often led to a significant increase (4% or more) with only a slight decrease (less than 1.5%) in the worst case. A particularly large increase in both average class accuracy (10.8%) and IoU score (9.8%) was observed for E-Net and in average class accuracy (6.6%) for SegNet. The slight
Table 1. This table provides a quantitative evaluation of six different methods trained on hand labelled (H) and augmented (A) data from CamVid [6] dataset. The number of iterations used to train each network is reported along with per-class segmentation accuracies, average classification accuracy, global accuracy and average per-class intersection over union (IoU) metric. All methods were trained on images of 480 × 360 resolution. Corresponding results of each network reported in the literature are also listed in rows marked with type R (reported). However, due to the different strategies followed in training and optimising deep networks, we encourage the reader to focus on the relative performance difference between the results we report. Large increases in class average and/or average IoU metrics are observed for E-Net and Segnet when trained on augmented data. The performance of Bayesian SegNet, FCN, Dilation Network and PSPNet is largely unaffected. * - note that a significant difference in previously reported accuracies [36] and the ones obtained by us of Dilation Network and Segnet when trained on augmented data. The performance of Bayesian SegNet, FCN, Dilation Network and PSPNet is largely unaffected. + - note that a significant difference in previously reported accuracies [36] and the ones obtained by us of Dilation Network may be explained by the use of lower resolution images instead of 640 × 480.

Table 2. The first row of this table contains the evaluation of augmented labels against hand labelled frames at various offsets from the seed label frame as illustrated in Figure 3. The rest of this table contains quantitative evaluation of four different methods trained on the hand labelled (H) and augmented (A) data from the CityScapes dataset. Similar metrics as in Table 1 are reported. The validation portion of the dataset is used for testing as the submission frequency of the CityScapes benchmark is highly limited. SegNet, Dilation Network and PSPNet show significant increase in class average accuracy and IoU metric scores when trained on augmented data.

decrease (less than 1.2%) in both metrics observed for the FCN network can be explained by its limited ability in handling small object classes at low resolution. A slight decrease in IoU matched by a higher class average observed for Bayesian SegNet is likely to have been caused by an unsaturated training requiring a significantly larger number of iterations for convergence due to the use of dropout in training.

5.2. CityScapes Dataset

Table 2 provides a quantitative analysis of SegNet [2], E-Net [28], Dilation Network [36] and PSPNet [37] tested on the validation portion of CityScapes dataset. SegNet and Dilation Network show a large increase in class average accuracy (6.6% and 8.1% respectively) and in IoU (3.3% and 8.2%). The increased performance of Dilation Network can be explained by existence of more pronounced small classes in CityScapes dataset. PSPNet shows a moderate increase.
in both class average (2.1%) and IoU (4.1%). Note while PSPNet does not reach reported \cite{37} accuracy, maximising overall performance is out of the scope of this work. Multiple experimental changes could have contributed to lower than reported performance: no pre-training was used, network input and image resolution was halved, only 1 week of a single GPU training time had been dedicated. Finally, note that a slightly decrease in the IoU and average class accuracy of E-Net may indicate the limitations of this very compact model when trained on a significantly more complex dataset than CamVid.

### 5.3. Instance Segmentation

Table 3 and Figure 5 show quantitative and qualitative comparison of Deep Watershed \cite{3} and Deep Mask \cite{29} instance segmentation networks when trained on hand labelled and on augmented data. Standard average precision metrics\cite{19} under different IoU thresholds (as shown in Table 3) are used to perform the quantitative evaluation.

As with semantic segmentation experiments both metrics (AP-0.5-ALL-100 and AP-0.5:0.95-ALL-100) show relatively larger increase (from 0.207 to 0.252) than decrease (from 0.095 to 0.094) when augmented data is used for training. Moreover a larger increase seen in both average precision metrics for Deep Watershed Transform Network translates to a qualitatively better separation of car instances shown in Figure 5.

### 6. Conclusion

In this paper we presented the analysis of the effect of large scale labelled video data augmentation for semantic segmentation in driving scenarios. For some networks, such as E-Net \cite{28}, SegNet \cite{2} or Dilation Network \cite{36} we observed increases of either average class accuracy or intersection over union (IoU) score in range from 6.6% to 10.8%. We also demonstrated the potential for using augmented data to improve the accuracy of instance level segmentation. The augmented data for CityScapes \cite{10}, CamVid \cite{6} and CamVid-Instance datasets will be made available to the research community for further evaluation.

### References


