05-07

16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

The CASE Dataset of Candidate Spaces for Advert Implantation

Soumyabrata Dev[1], Murhaf Hossari[1], Matthew Nicholson[1], Killian McCabe[1], Atul Nautiyal[1]
Clare Conran[1], Jian Tang[3], Wei Xu[3], and François Pitié[1,2] * †

[1]The ADAPT SFI Research Centre, Trinity College Dublin
[2]Department of Electronic & Electrical Engineering, Trinity College Dublin
[3]Huawei Ireland Research Center, Dublin

Abstract

With the advent of faster internet services and growth of multimedia content, we observe a massive growth in the number of online videos. The users generate these video contents at an unprecedented rate, owing to the use of smart-phones and other hand-held video capturing devices. This creates immense potential for the advertising and marketing agencies to create personalized content for the users. In this paper, we attempt to assist the video editors to generate augmented video content, by proposing candidate spaces in video frames. We propose and release a large-scale dataset of outdoor scenes, along with manually annotated maps for candidate spaces. We also benchmark several deep-learning based semantic segmentation algorithms on this proposed dataset.

1 Introduction

Impressive technological advances have provided a platform to create and share on-demand video contents over the internet. The consumers have an insatiable appetite for media, and artificial intelligence (AI)-based techniques assist us in understanding such media [3]. This creates opportunity for the marketing and the advertisement agencies to generate personalized marketing products, according to the consumers' likes and dislikes. However, the current *skip-ad* generation have a limited attention span [1], and therefore, new strategies of advertisement needs to be employed. One such technique include product placement or embedded marketing, that involves the video editors to artificially integrate new adverts into original scenes.

Online videos provide several opportunities for integrating new advertisements in the original videos. We developed a deep-learning based advert creation system [6] that can automatically identify the frames in a video containing an advertisement [4]. Additionally, we also identify the specific frames in a video where new advertisements can be artificially augmented, conforming to the subjective human judgement. It is important to select the candidate space in a frame in a systematic manner, such that the new advert can be augmented in a seamless and sensical manner. For example, adding advertisement boards by the side of the road, and onto empty wall spaces are good candidates for advert integration. Figure 1 describes such a scenario, wherein candidate spaces are marked based on the scene understanding. These candidate spaces of new advertisements in a video frame are generally manually selected by the video-editors during the post-processing stage.
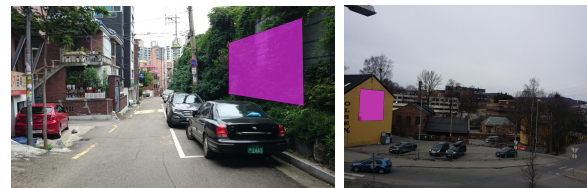
(a) Floating candidate   (b) Anchored candidate

Figure 1: Potential spaces for advertisement integration in outdoor scenes. The candidate spaces can either be floating across the scene, or anchored against walls in the scene.

In most cases, the video editors uses his/her understanding of the scene, and proposes a new space for advert integration. In the literature, there are no related works that can automatically parse the scene in the image, and propose a candidate space for integrating advertisements. We attempt to bridge this gap in the literature, by releasing the first large-scale dataset of outdoor scenes, with manually annotated candidate spaces.

The main contributions of this paper are as follows: (a) we propose and release the first large scale dataset of candidate placements in outdoor scene images, (b) we also provide a systematic and detailed benchmarking results of several popular deep learning based segmentation frameworks in our proposed dataset. We believe that this dataset will greatly serve the multimedia and advertisement communities, and assist in providing interesting research in this field. The rest of the paper is arranged as follows: Section 2 describes the dataset, and its characteristics. We also provide a few benchmarking results on the proposed dataset in Section 3. Finally, Section 4 concludes the paper, and describes our future work.

2 Dataset

We refer our dataset as **CASE** dataset, that stands for **CA**ndidate **S**paces for adv**E**rt implantation dataset [1]. The sources of the images in this dataset is Cityscapes dataset [2] comprising street view images. We randomly select 10000 images from the Cityscapes dataset that for annotating with potential placements for advertisements.

---

[1]The download link of the dataset is available here: https://www.adaptcentre.ie/case-studies/ai-based-advert-creation-system-for-next-generation-publicity
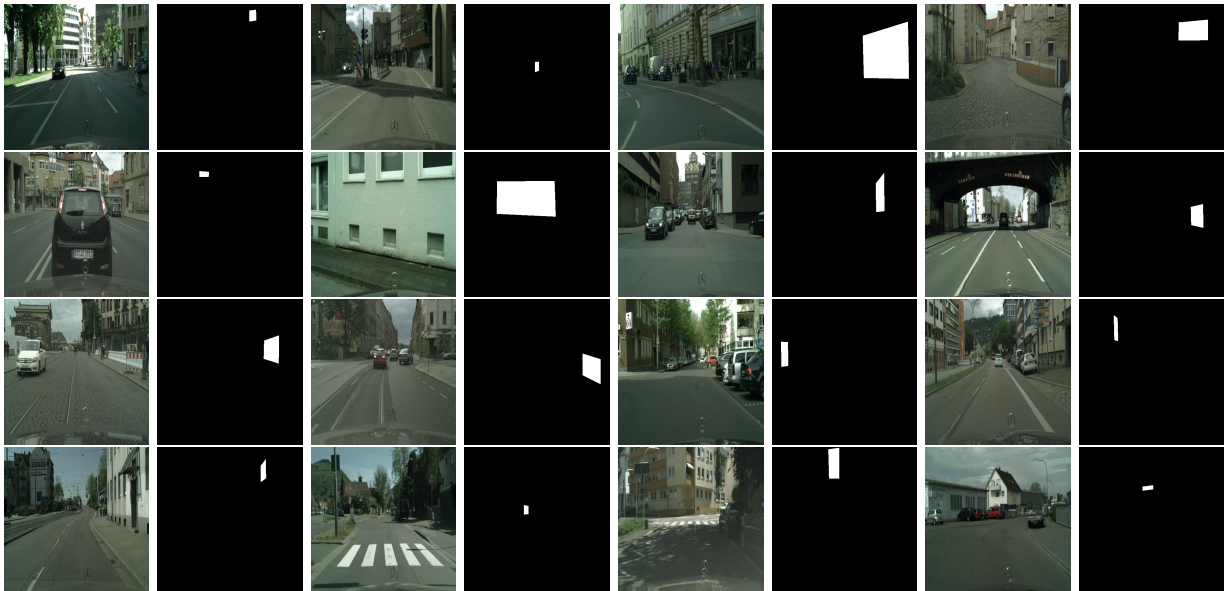
Figure 2: Representative images of the CASE dataset, along with manually annotated binary candidate maps for integrating new adverts.

2.1 Creation of candidate spaces

We are interested in learning candidate placement of objects within a scenes. We have chosen to investigate placing regularly shaped billboard in street view images. In our application case study, billboards are considered as rectangular areas that can be added to the sides of buildings, or fixed by the pathways and road sides for placement of posters or advertisements. We employ paid volunteers to manually go through each of our selected 10000 images, and mark the location where they can assume that a billboard could be placed. This process is done carefully, such that the embedded advertisement matches the perspective of the scene. We use the VGG Image Annotator tool [2] for annotating the candidate spaces. The user annotate each image in the dataset by selecting 4 points, that either form an anchored placement (against wall or surface) or floating placement in the scene.

Although there can be several potential spaces to embed the new advertisement, we restrict the annotators to include only a single candidate placement in an image. This is intentionally done, such that the annotator chooses the *best* candidate space for integration, according to his/her subjective judgment. Examples of good placement include, but not limited to, placing advert on a blank wall, beside existing billboards, hanging beside a lamppost, or over a window/group of windows. We do not annotate existing billboards or posters. This process is completed for all the selected images of our dataset. Post the annotation process, we manually go through all the generated candidate spaces, to ensure that the annotations are performed properly.

Figure 2 shows a few representative images of the CASE dataset, along with its manually annotated binary masks. The binary masks are generated from the four corners of the annotated billboard.

2.2 Dataset characteristics

The candidate spaces in the images are annotated with varying sizes. We define the amount of image area covered with a candidate space as the advert coverage value. We ensure that the CASE dataset has a varying values of advert coverage values, across all the images. Figure 3 describes the distribution of the coverage of the candidate spaces.
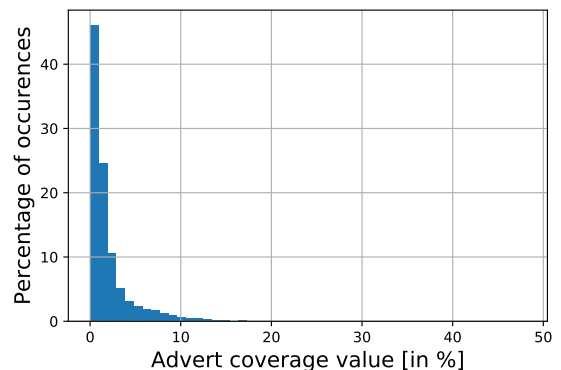

Figure 3: Statistical distribution of the percentage of occurrences of the advertisement spaces w.r.t. the advert coverage values.

3 Benchmarking Experiments

In addition to the proposed dataset of candidate spaces, we also provide a few benchmarking experiments of several semantic segmentation algorithms in the CASE dataset. With the success of deep neural networks, several convolutional neural networks have been proposed that provides impressive results on various semantic segmentation tasks. We benchmark the performance of fully convolutional neural network (FCN) [5], pyramid scene parsing network (PSP-Net) [8], and U-Net [7].

The FCN network by Long et al. employs only locally connected layers, without the use of any dense layers in its architecture. The layers in FCN include

convolution, pooling and upsampling layers. It provides dense segmentation masks of input images of any dimension. Zhao et al. in [8] introduced a pyramid pooling module, that can better grasp the context of the input image. The PSPNet produced impressive results on the ImageNet scene parsing challenge [9]. The advantage of the pyramid pooling module is that it captures information at varying scales, across different sub-regions of the input image. Recently, Ronneberger et al. introduced a convolutional neural network called U-Net that provided impressive results in biomedical image segmentation. The U-Net network has a symmetric structure, and consists of three primary parts – contracting, bottleneck, and expanding path [7]. It provides skip connection between the contracting and expanding paths, that provides the local information to the global information, during the upsampling process. These networks are popularly used in the area of semantic segmentation, for an efficient generation of segmentation masks. We use these networks as benchmarking models for our proposed dataset. Currently, there are no networks that are specifically designed for identifying candidate spaces in input images.

3.1 Subjective Evaluation

Figure 4 shows a few sample results of the benchmarking algorithms on the CASE dataset. We visualize the results obtained from these algorithms, prior to the final *softmax* layer. This is useful, as it provides us a probabilistic viewpoint on the possibility of a candidate space in an image. Most of these approaches generate a lot of false positives in the generated binary maps. This is understandable because there is no *unique* position of advertisement in an input image. The position of the embedded advertisement is completely dependent on the subjective judgement of the annotator. It is interesting to observe from Fig. 4(d) and (e), that PSPNet and U-Net identifies the buildings and the sideways of the roads as possible placements for new adverts. The FCN network, on the other hand is conservative in nature – it produces significantly fewer false positives, and do not propose a particular area of the image as a candidate space for advertisement integration. However, we observe from the visual results, that most of these networks viz. PSPNet and U-Net can successfully learn the semantics of the input scenes, and provide reasonable maps of candidate spaces.

3.2 Objective Evaluation

In addition to the subjective evaluation of the benchmarking algorithms, we also provide an objective evaluation of the different approaches. We report the average values of the following metrics – pixel accuracy, mean accuracy, mean intersection over union, and frequency weighted intersection over union. These metrics are variations of classification accuracy and intersection over union (IOU). Let us assume that the total number of pixels belonging to class i, and predicted to belong to class j, is n_{ij}. The total number of classes in this particular task of semantic segmentation is assumed to be n_{cl}.

$$\text{Pixel Accuracy} = \frac{\sum_i n_{ii}}{\sum_i t_i} \quad (1)$$

$$\text{Mean Accuracy} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i} \quad (2)$$

$$\text{Mean Intersection Over Union} = \frac{1}{n_{cl}} \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \quad (3)$$

$$\text{Frequency Weighted Intersection Over Union}$$
$$= \frac{1}{\sum_k t_k} \frac{\sum_i t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}, \quad (4)$$

where $t_i = \sum_{j=1}^{n_{cl}} n_{ij}$ is the total number of pixels in class i.

Table 1 summarizes the results of the various benchmarking algorithms in CASE dataset.

	Pixel Accuracy	Mean Accuracy	Mean IOU	Frequency Weighted IOU
FCN	**0.978**	0.509	**0.498**	**0.959**
PSPNet	0.545	0.625	0.284	0.529
U-Net	0.619	**0.727**	0.327	0.601

Table 1: Benchmarking of the CASE dataset with various deep-learning based segmentation algorithms. The best performance for each metric is indicated in bold.

Amongst the benchmarking networks, we observe that the FCN network has the best evaluation scores. The FCN network can learn the semantics of the annotated scene, and provides a good understanding of a candidate space in an image. The results obtained from PSPNet and U-Net can be further improved by training these networks on larger dataset of diverse image types, and employing batch normalization across multiple graphics processing units (GPUs). These results provide a detailed benchmark of the popular segmentation algorithms on the candidate space dataset. The results can be further improved, by designing a bespoke neural network that is specifically designed for the task of identifying candidate spaces in outdoor scenes.

4 Conclusion and Future Work

In this paper, we propose CASE – a large-scale dataset of outdoor scenes with manually annotated binary maps of candidate spaces. This dataset is the first of its kind that assists the video editors an efficient and systematic manner of creating augmented video contents with new advertisements. In the future, we intend to relax this criterion of outdoor scenes, and also further extend this dataset, by including images from indoor scenes and entertainment television shows.

Acknowledgement

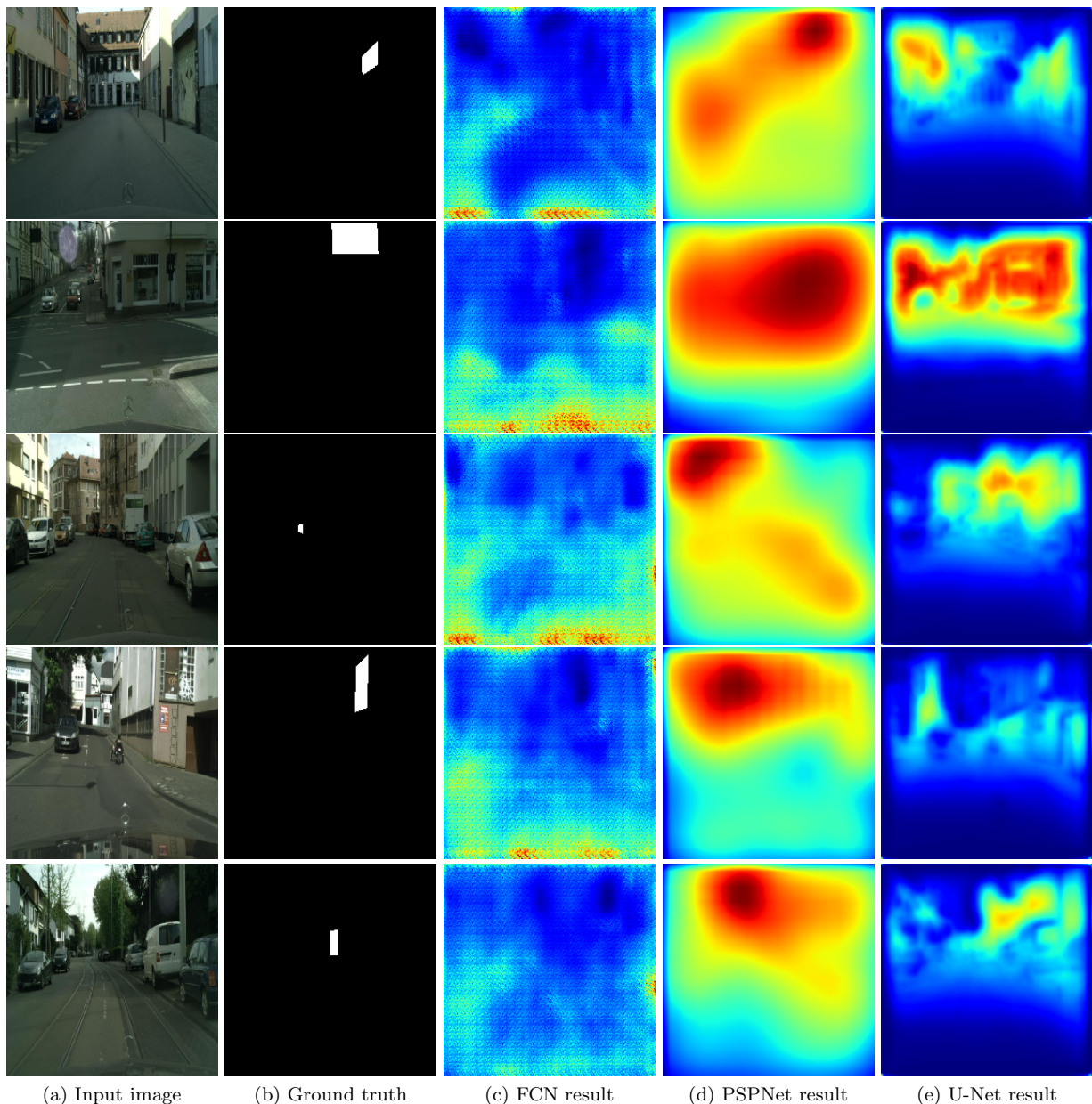| (a) Input image | (b) Ground truth | (c) FCN result | (d) PSPNet result | (e) U-Net result |

Figure 4: Subjective evaluation of various segmentation algorithms on CASE dataset. The results in (c–e) are visualized, prior to the final *softmax* layer of the individual networks.

References

[1] Chalawsky, M.: Ad skip feature for characterizing advertisement effectiveness (June 18 2013), US Patent 8,468,056

[2] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)

[3] Hossari, M., Dev, S., Kelleher, J.D.: TEST: A terminology extraction system for technology related terms. arXiv preprint arXiv:1812.09541 (2018)

[4] Hossari, M., Dev, S., Nicholson, M., McCabe, K., Nautiyal, A., Conran, C., Tang, J., Xu, W., Pitié, F.: ADNet: A deep network for detecting adverts. arXiv preprint arXiv:1811.04115 (2018)

[5] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

[6] Nautiyal, A., McCabe, K., Hossari, M., Dev, S., Nicholson, M., Conran, C., McKibben, D., Tang, J., Wei, X., Pitié, F.: An advert creation system for next-gen publicity. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) (2018)

[7] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

[8] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)

[9] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. arXiv preprint arXiv:1608.05442 (2016)