

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, *Senior Member, IEEE*,

Abstract—We present a novel and practical deep fully convolutional neural network architecture for semantic pixel-wise segmentation termed SegNet. This core trainable segmentation engine consists of an encoder network, a corresponding decoder network followed by a pixel-wise classification layer. The architecture of the encoder network is topologically identical to the 13 convolutional layers in the VGG16 network [1]. The role of the decoder network is to map the low resolution encoder feature maps to full input resolution feature maps for pixel-wise classification. The novelty of SegNet lies in the manner in which the decoder upsamples its lower resolution input feature map(s). Specifically, the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This eliminates the need for learning to upsample. The upsampled maps are sparse and are then convolved with trainable filters to produce dense feature maps. We compare our proposed architecture with the fully convolutional network [2] architecture and its variants. This comparison reveals the memory versus accuracy trade-off involved in achieving good segmentation performance.

The design of SegNet was primarily motivated by road scene understanding applications. Hence, it is efficient both in terms of memory and computational time during inference. It is also significantly smaller in the number of trainable parameters than competing architectures and can be trained end-to-end using stochastic gradient descent without complex training protocols. We also benchmark the performance of SegNet on Pascal VOC12 salient object segmentation and the recent SUN RGB-D indoor scene understanding challenge. These quantitative assessments show that SegNet provides competitive performance although it is significantly faster than other architectures. We also provide a Caffe implementation of SegNet and a web demo at <http://mi.eng.cam.ac.uk/projects/segnet/>.

Index Terms—Deep Convolutional Neural Networks, Semantic Pixel-Wise Segmentation, Encoder, Decoder, Pooling, Upsampling.



1 INTRODUCTION

Semantic segmentation has a wide array of applications ranging from scene understanding, inferring support-relationships among objects to autonomous driving. Early methods that relied on low-level vision cues have fast been superseded by popular machine learning algorithms. In particular, deep learning has seen huge success lately in handwritten digit recognition, speech, categorising whole images and detecting objects in images [4], [5]. Now there is an active interest for semantic pixel-wise labelling [6] [7], [8], [2], [9], [10], [11], [12], [13], [14]. However, some of these recent approaches have tried to directly adopt deep architectures designed for category prediction to pixel-wise labelling [6]. The results, although very encouraging, appear coarse [14]. This is primarily because max pooling and sub-sampling reduce feature map resolution. Our motivation to design SegNet arises from this need to map low resolution features to input resolution for pixel-wise classification. This mapping must produce features which are useful for accurate boundary localization.

Our architecture, SegNet, is designed to be a *core segmentation engine* for pixel-wise semantic segmentation. It is primarily motivated by road scene understanding applications which require the ability to model appearance (road, building), shape (cars, pedestrians) and understand the spatial-relationship (context) be-

tween different classes such as road and side-walk. In typical road scenes, the majority of the pixels belong to large classes such as road, building and hence the network must produce smooth segmentations. The engine must also have the ability to delineate moving and other objects based on their shape despite their small size. Hence it is important to retain boundary information in the extracted image representation. From a computational perspective, it is necessary for the network to be efficient in terms of both memory and computation time during inference. It must also be able to train end-to-end in order to jointly optimise all the weights in the network using an efficient weight update technique such as stochastic gradient descent (SGD) [15]. Networks that are trained end-to-end or equivalently those that do not use multi-stage training [2] or other supporting aids such as region proposals [9] help establish benchmarks that are more easily repeatable. The design of SegNet arose from a need to match these criteria.

The encoder network in SegNet is topologically identical to the convolutional layers in VGG16 [1]. We remove the fully connected layers of VGG16 which makes the SegNet encoder network significantly smaller and easier to train than many other recent architectures [2], [9], [11], [16]. The key component of SegNet is the decoder network which consists of a hierarchy of decoders one corresponding to each encoder. Of these, the appropriate decoders use the max-pooling indices received from the corresponding encoder to perform non-linear upsampling of their input feature maps. This idea was inspired from an architecture designed for unsupervised feature learning [17]. Reusing max-pooling indices in the decoding process has several practical

- V. Badrinarayanan, A. Kendall, R. Cipolla are with the Machine Intelligence Lab, Department of Engineering, University of Cambridge, UK. E-mail: vb292,agk34,cipolla@eng.cam.ac.uk

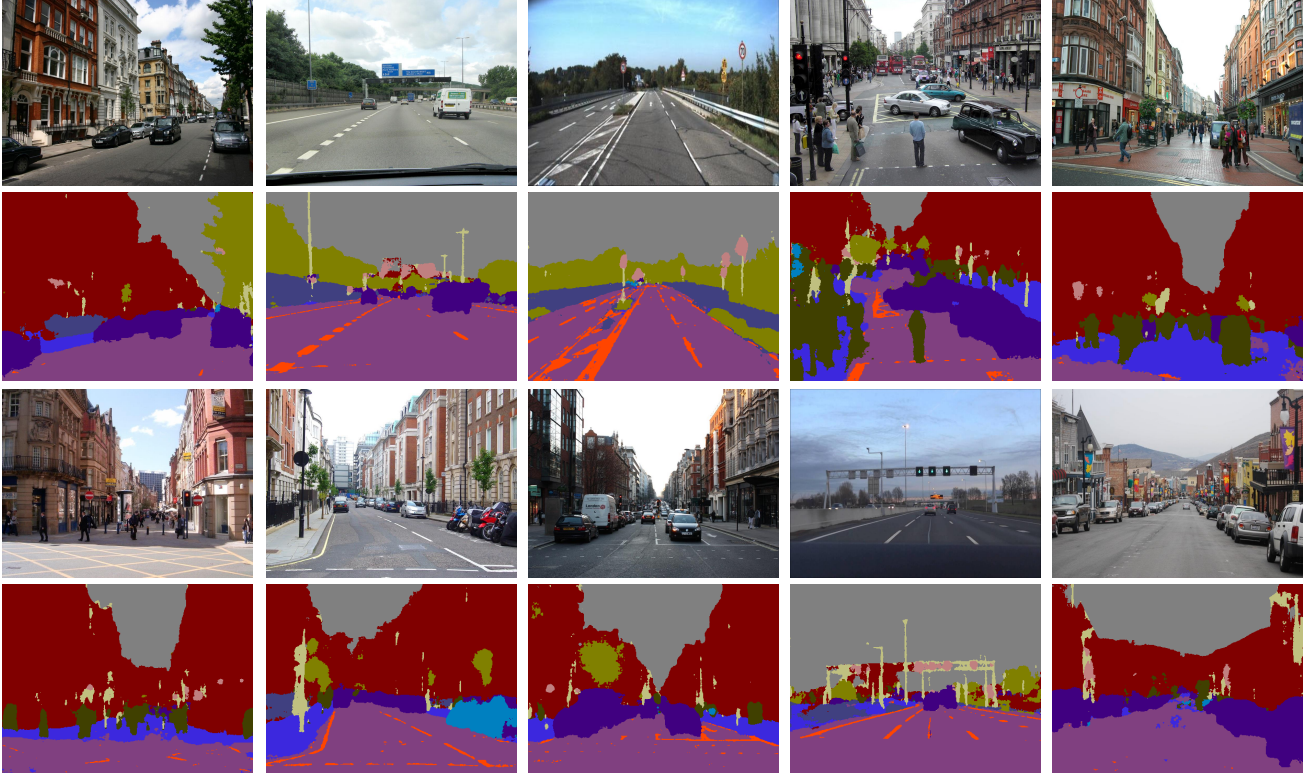


Fig. 1. SegNet predictions on urban and highway scene test samples from the wild. The class colour codes can be obtained from Brostow *et al.* [3]. To try our system yourself, please see our online web demo at <http://mi.eng.cam.ac.uk/projects/segnet/>.

advantages; (i) it improves boundary delineation, (ii) it reduces the number of parameters enabling end-to-end training, and (iii) this form of upsampling can be incorporated into any encoder-decoder architecture such as [2], [10] with only a little modification.

One of the main contributions of this paper is our analysis of the SegNet decoding technique and the widely used Fully Convolutional Network (FCN) [2]. This is in order to convey the practical trade-offs involved in designing segmentation architectures. Most recent deep architectures for segmentation have identical encoder networks, i.e. VGG16, but differ in the form of the decoder network, training and inference. Another common feature is they have trainable parameters in the order of hundreds of millions and thus encounter difficulties in performing end-to-end training [9]. The difficulty of training these networks has led to multi-stage training [2], appending networks to a pre-trained core segmentation engine such as FCN [10], use of supporting aids such as region proposals for inference [9], disjoint training of classification and segmentation networks [16] and use of additional training data for pre-training [11] [18] or for full training [10]. In addition, performance boosting post-processing techniques [14] have also been popular. Although all these factors improve performance on challenging benchmarks [19], it is unfortunately difficult from their quantitative results to disentangle the key design factors necessary to achieve good performance. We therefore analysed the decoding process used in some of these approaches [2], [9] and reveal their pros and cons.

We evaluate the performance of SegNet on PascalVOC12 salient object(s) segmentation [19], [20] and scene understanding challenges such as CamVid road scene segmentation [3] and SUN RGB-D indoor scene segmentation [21]. Pascal VOC12 [19] is the benchmark for segmentation due to its size and challenges, but the

majority of this task has one or two foreground classes surrounded by a highly varied background. This implicitly favours techniques used for detection as shown by the recent work on a decoupled classification-segmentation network [16] where the classification network can be trained with a large set of weakly labelled data and the independent segmentation network performance is improved. The method of [14] also use the feature maps of the classification network with an independent CRF post-processing technique to perform segmentation. The performance can also be boosted by the use additional inference aids such as region proposals [9], [22]. Therefore, it is different from scene understanding where the idea is to exploit co-occurrences of objects and other spatial-context to perform robust segmentation. To demonstrate the efficacy of SegNet, we present a real-time online demo of road scene segmentation into 11 classes of interest for autonomous driving (see Fig. 1). Some example test results produced on randomly sampled road scene images from Google are shown in Fig. 1.

The remainder of the paper is organized as follows. In Sec. 2 we review related recent literature. We describe the SegNet architecture and its analysis in Sec. 3. In Sec. 4 we evaluate the performance of SegNet on several benchmark datasets. This is followed by a general discussion regarding our approach with pointers to future work in Sec. 5. We conclude in Sec. 6.

2 LITERATURE REVIEW

Semantic pixel-wise segmentation is an active topic of research, fuelled by challenging datasets [3], [19], [21], [23], [24]. Before the arrival of deep networks, the best performing methods mostly relied on hand engineered features classifying pixels independently. Typically, a patch is fed into a classifier *e.g.* Random

Forest [25], [26] or Boosting [27], [28] to predict the class probabilities of the center pixel. Features based on appearance [25] or SfM and appearance [26], [27], [28] have been explored for the CamVid road scene understanding test [3]. These per-pixel noisy predictions (often called *unary* terms) from the classifiers are then smoothed by using a pair-wise or higher order CRF [27], [28] to improve the accuracy. More recent approaches have aimed to produce high quality unaries by trying to predict the labels for all the pixels in a patch as opposed to only the center pixel. This improves the results of Random Forest based unaries [29] but thin structured classes are classified poorly. Dense depth maps computed from the CamVid video have also been used as input for classification using Random Forests [30]. Another approach argues for the use of a combination of popular hand designed features and spatio temporal super-pixelization to obtain higher accuracy [31]. The best performing technique on the CamVid test [28] addresses the imbalance among label frequencies by combining object detection outputs with classifier predictions in a CRF framework. The result of all these techniques indicate the need for improved features for classification.

Indoor RGBD pixel-wise semantic segmentation has also gained popularity since the release of the NYU dataset [23]. This dataset showed the usefulness of the depth channel to improve segmentation. Their approach used features such as RGB-SIFT, depth-SIFT and pixel location as input to a neural network classifier to predict pixel unaries. The noisy unaries are then smoothed using a CRF. Improvements were made using a richer feature set including LBP and region segmentation to obtain higher accuracy [32] followed by a CRF. In more recent work [23], both class segmentation and support relationships are inferred together using a combination of RGB and depth based cues. Another approach focuses on real-time joint reconstruction and semantic segmentation, where Random Forests are used as the classifier [33]. Gupta *et al.* [34] use boundary detection and hierarchical grouping before performing category segmentation. The common attribute in all these approaches is the use of hand engineered features for classification of either RGB or RGBD images.

The success of deep convolutional neural networks for object classification has more recently led to researchers to exploit their feature learning capabilities for structured prediction problems such as segmentation. There have also been attempts to apply networks designed for object categorization to segmentation, particularly by replicating the deepest layer features in blocks to match image dimensions [6], [35], [36], [37]. However, the resulting classification is blocky [36]. Another approach using recurrent neural networks [38] merges several low resolution predictions to create input image resolution predictions. These techniques are already an improvement over hand engineered features [6] but their ability to delineate boundaries is poor.

Newer deep architectures [2], [9], [10], [13], [16] particularly designed for segmentation have advanced the state-of-the-art by learning to decode or map low resolution image representations to pixel-wise predictions. The encoder network which produces these low resolution representations in all of these architectures is the VGG16 classification network [1] which has 13 convolutional layers and 3 fully connected layers. This encoder network weights are typically pre-trained on the large ImageNet object classification dataset [39]. The decoder network varies between these architectures and is the part which is responsible for producing multi-dimensional features for each pixel for classification.

Each decoder in the Fully Convolutional Network (FCN)

architecture [2] learns to upsample its input feature map(s) and combines them with the corresponding encoder feature map to produce the input to the next decoder. It is a core segmentation engine which has a large number of trainable parameters in the encoder network (134M) but a very small decoder network (0.5M). The overall large size of this network makes it hard to train end-to-end on a relevant task. Therefore, the authors use a stage-wise training process. Here each decoder in the decoder network is progressively added to an existing trained network. The network is grown until no further increase in performance is observed. This growth is stopped after three decoders thus ignoring high resolution feature maps can certainly lead to loss of edge information [9]. Apart from training related issues, the need to reuse the encoder feature maps in the decoder makes it memory intensive in test time. We study this network in more detail as it the core of other recent architectures [10], [11].

The predictive performance of FCN has been improved further by appending the FCN with a recurrent neural network (RNN) [10] and fine-tuning them on large datasets [19], [40]. The RNN layers mimic the sharp boundary delineation capabilities of CRFs while exploiting the feature representation power of FCN's. They show a significant improvement over FCN-8 but also show that this difference is reduced when more training data is used to train FCN-8. The main advantage of the CRF-RNN is when it is jointly trained with a core segmentation engine such as the FCN-8. The fact that joint training helps is also shown in other recent results [41], [42]. Interestingly, the deconvolutional network [9] performs significantly better than FCN although at the cost of a more complex training and inference. This however raises the question as to whether the perceived advantage of the CRF-RNN would be reduced as the core feed-forward segmentation engine is made better. In any case, the CRF-RNN network can be appended to any core segmentation engine including SegNet.

Multi-scale deep architectures are also being pursued [13], [42]. They come in two flavours, (i) those which use input images at a few scales and corresponding deep feature extraction networks, and (ii) those which combine feature maps from different layers of a single deep architecture [43] [11]. The common idea is to use features extracted at multiple scales to provide both local and global context [44] and the using feature maps of the early encoding layers retain more high frequency detail leading to sharper class boundaries. Some of these architectures are difficult to train due to their parameter size [13]. Thus a multi-stage training process is employed along with data augmentation. The inference is also expensive with multiple convolutional pathways for feature extraction. Others [42] append a CRF to their multi-scale network and jointly train them. However, these are not feed-forward at test time and require optimization to determine the MAP labels.

Several of the recently proposed deep architectures for segmentation are not feed-forward in inference time [9], [14], [16]. They require either MAP inference over a CRF [42], [41] or aids such as region proposals [9] for inference. We believe the perceived performance increase obtained by using a CRF is due to the lack of good decoding techniques in their core feed-forward segmentation engine. SegNet on the other hand uses decoders to obtain features for accurate pixel-wise classification.

The recently proposed Deconvolutional Network [9] and its semi-supervised variant the Decoupled network [16] use the max locations of the encoder feature maps (pooling indices) to perform non-linear upsampling in the decoder network. The authors of these architectures, independently of SegNet (first submitted to

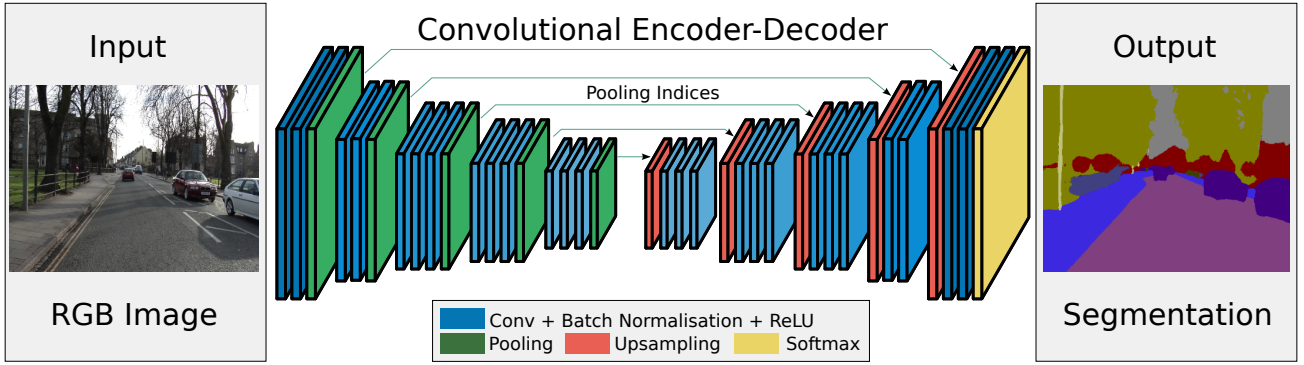


Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

CVPR 2015 [12]), proposed this idea of decoding in the decoder network. However, their encoder network consists of the fully connected layers from the VGG-16 network which consists of about 90% of the parameters of their entire network. This makes training of their network very difficult and thus require additional aids such as the use of region proposals to enable training. Moreover, during inference these proposals are used and this increases inference time significantly. From a benchmarking point of view, this also makes it difficult to evaluate the performance of their core segmentation engine (encoder-decoder network) without other aids. In this work we discard the fully connected layers of the VGG16 encoder network which enables us to train the network using the relevant training set using SGD optimization. Another recent method [14] shows the benefit of reducing the number of parameters significantly without sacrificing performance, reducing memory consumption and improving inference time.

Our work was inspired by the unsupervised feature learning architecture proposed by Ranzato *et al.* [17]. The key learning module is an encoder-decoder network. An encoder consists of convolution with a filter bank, element-wise tanh non-linearity, max-pooling and sub-sampling to obtain the feature maps. For each sample, the indices of the max locations computed during pooling are stored and passed to the decoder. The decoder upsamples the feature maps by using the stored pooled indices. It convolves this upsampled map using a trainable decoder filter bank to reconstruct the input image. This architecture was used for unsupervised pre-training for classification. A somewhat similar decoding technique is used for visualizing trained convolutional networks [45] for classification. The architecture of Ranzato *et al.* mainly focused on layer-wise feature learning using small input patches. This was extended by Kavukcuoglu *et al.* [46] to accept full image sizes as input to learn hierarchical encoders. Both these approaches however did not attempt to use *deep encoder-decoder* networks for unsupervised feature training as they discarded the decoders after each encoder training. Here, SegNet differs from these architectures as the deep encoder-decoder network is trained jointly for a supervised learning task and hence the decoders are an integral part of the network in test time.

Other applications where pixel wise predictions are made using deep networks are image super-resolution [47] and depth map prediction from a single image [48]. The authors in [48] discuss the need for learning to upsample from low resolution feature maps which is the central topic of this paper.

3 ARCHITECTURE

SegNet has an encoder network and a corresponding decoder network, followed by a final pixelwise classification layer. This architecture is illustrated in Fig. 3. The encoder network consists of 13 convolutional layers which correspond to the first 13 convolutional layers in the VGG16 network [1] designed for object classification. We can therefore initialize the training process from weights trained for classification on large datasets [39]. We can also discard the fully connected layers in favour of retaining higher resolution feature maps at the deepest encoder output. This also reduces the number of parameters in the SegNet encoder network significantly (from 134M to 14.7M) as compared to other recent architectures [2], [9] (see. Table 5). Each encoder layer has a corresponding decoder layer and hence the decoder network has 13 layers. The final decoder output is fed to a multi-class soft-max classifier to produce class probabilities for each pixel independently.

Each *encoder* in the encoder network performs convolution with a filter bank to produce a set of feature maps. These are then batch normalized [49]). Then an element-wise rectified-linear non-linearity (ReLU) $\max(0, x)$ is applied. Following that, max-pooling with a 2×2 window and stride 2 (non-overlapping window) is performed and the resulting output is sub-sampled by a factor of 2. Max-pooling is used to achieve translation invariance over small spatial shifts in the input image. Sub-sampling results in a large input image context (spatial window) for each pixel in the feature map. While several layers of max-pooling and sub-sampling can achieve more translation invariance for robust classification correspondingly there is a loss of spatial resolution of the feature maps. The increasingly lossy (boundary detail) image representation is not beneficial for segmentation where boundary delineation is vital. Therefore, it is necessary to *capture and store* boundary information in the encoder feature maps before sub-sampling is performed. If memory during inference is not constrained, then all the encoder feature maps (after sub-sampling) can be stored. This is usually not the case in practical applications and hence we propose a more efficient way to store this information. It involves storing only the max-pooling *indices*, i.e, the locations of the maximum feature value in each pooling window is memorized for each encoder feature map. In principle, this can be done using 2 bits for each 2×2 pooling window and is thus much more efficient to store as compared to memorizing feature map(s) in float precision. As we show later in this work,

this lower memory storage results in a slight loss of accuracy but is still suitable for practical applications.

The appropriate *decoder* in the decoder network upsamples its input feature map(s) using the memorized max-pooling indices from the corresponding encoder feature map(s). This step produces sparse feature map(s). This SegNet decoding technique is illustrated in Fig. 3. These feature maps are then convolved with a trainable decoder filter bank to produce dense feature maps. A batch normalization step is then applied to each of these maps. Note that the decoder corresponding to the first encoder (closest to the input image) produces a multi-channel feature map, although its encoder input has 3 channels (RGB). This is unlike the other decoders in the network which produce feature maps with the same number of size and channels as their encoder inputs. The high dimensional feature representation at the output of the final decoder is fed to a trainable soft-max classifier. This soft-max classifies each pixel independently. The output of the soft-max classifier is a K channel image of probabilities where K is the number of classes. The predicted segmentation corresponds to the class with maximum probability at each pixel.

3.1 Decoder Variants

Many segmentation architectures [2], [9], [14] share the same encoder network and they only vary in the form of their decoder network. Of these we choose to compare the SegNet decoding technique with the widely used Fully Convolutional Network (FCN) decoding technique [2], [10].

In order to analyse SegNet and compare its performance with FCN (decoder variants) we use a smaller version of SegNet, termed **SegNet-Basic**¹, which has 4 encoders and 4 decoders. All the encoders in SegNet-Basic perform max-pooling and sub-sampling and the corresponding decoders upsample its input using the received max-pooling indices. Batch normalization is used after each convolutional layer in both the encoder and decoder network. No biases are used after convolutions and no ReLU non-linearity is present in the decoder network. Further, a constant kernel size of 7×7 over all the encoder and decoder layers is chosen to provide a wide context for smooth labelling *i.e.* a pixel in the deepest layer feature map (layer 4) can be traced back to a context window in the input image of 106×106 pixels. This small size of SegNet-Basic allows us to explore many different variants (decoders) and train them in reasonable time. Similarly we create **FCN-Basic**, a comparable version of FCN for our analysis which shares the same encoder network as SegNet-Basic but with the FCN decoding technique in the corresponding decoders.

On the left in Fig. 3 is the decoding technique used by SegNet (also SegNet-Basic), where there is no learning involved in the upsampling step. However, the upsampled maps are convolved with trainable multi-channel decoder filters to densify the sparse inputs. Each decoder filter has the same number of channels as the number of upsampled feature maps. A smaller variant is one where the decoder filters are single channel, *i.e.* they only convolve their corresponding upsampled feature map. This variant (**SegNet-Basic-SingleChannelDecoder**) reduces the number of trainable parameters and inference time significantly.

On the right in Fig. 3 is the FCN (also FCN-Basic) decoding technique. The important design element of the FCN model is dimensionality reduction step of the encoder feature maps. This

compresses the encoder feature maps which are then used in the corresponding decoders. Dimensionality reduction of the encoder feature maps, say of 64 channels, is performed by convolving them with $1 \times 1 \times 64 \times K$ trainable filters, where K is the number of classes. The compressed K channel final encoder layer feature maps are the input to the decoder network. In a decoder of this network, upsampling is performed by convolution using a trainable *multi-channel upsampling kernel*. We set the kernel size to 8×8 . This manner of upsampling is also termed as *deconvolution*. Note that in SegNet the multi-channel convolution using trainable decoder filters is performed after upsampling to densifying feature maps. The upsampled feature map in FCN is then added to the corresponding resolution encoder feature map to produce the output decoder feature map. The upsampling kernels are initialized using bilinear interpolation weights [2].

The FCN decoder model requires storing encoder feature maps during inference. This can be memory intensive, for *e.g.* storing 64 feature maps of the first layer of FCN-Basic at 180×240 resolution in 32 bit floating point precision takes 11MB. This can be made smaller using dimensionality reduction to the 11 feature maps which requires ≈ 1.9 MB storage. SegNet on the other hand requires almost negligible storage cost for the pooling indices (.17MB if stored using 2 bits per 2×2 pooling window). We can also create a variant of the FCN-Basic model which discards the encoder feature map addition step and only learns the upsampling kernels (**FCN-Basic-NoAddition**).

In addition to the above variants, we study upsampling using fixed bilinear interpolation weights which therefore requires no learning for upsampling (**Bilinear-Interpolation**). At the other extreme, we can add 64 encoder feature maps at each layer to the corresponding output feature maps from the SegNet decoder to create a more memory intensive variant of SegNet (**SegNet-Basic-EncoderAddition**). Another and more memory intensive FCN-Basic variant (**FCN-Basic-NoDimReduction**) is where there is no dimensionality reduction performed for the encoder feature maps. Finally, please note that to encourage reproduction of our results we release the Caffe implementation of all the variants².

We also tried other generic variants where feature maps are simply upsampled by *replication* [6], or by using a fixed (and sparse) array of indices for upsampling. These performed quite poorly in comparison to the above variants. A variant without max-pooling and sub-sampling in the encoder network (decoders are redundant) consumes more memory, takes longer to converge and performs poorly.

3.2 Training

We use the CamVid road scenes dataset to benchmark the performance of the decoder variants. This dataset is small, consisting of 367 training and 233 testing RGB images (day and dusk scenes) at 360×480 resolution. The challenge is to segment 11 classes such as road, building, cars, pedestrians, signs, poles, side-walk etc. We perform local contrast normalization [50] to the RGB input.

The encoder and decoder weights were all initialized using the technique described in He *et al.* [51]. To train all the variants we use stochastic gradient descent (SGD) with a fixed learning rate of 0.1 and momentum of 0.9 [15] using our Caffe implementation of SegNet-Basic [52]. We train the variants until the training loss converges. Before each epoch, the training set is shuffled and each

1. SegNet-Basic was earlier termed SegNet in a archival version of this paper [12]

2. See <http://mi.eng.cam.ac.uk/projects/segnet/> for our SegNet code and web demo.

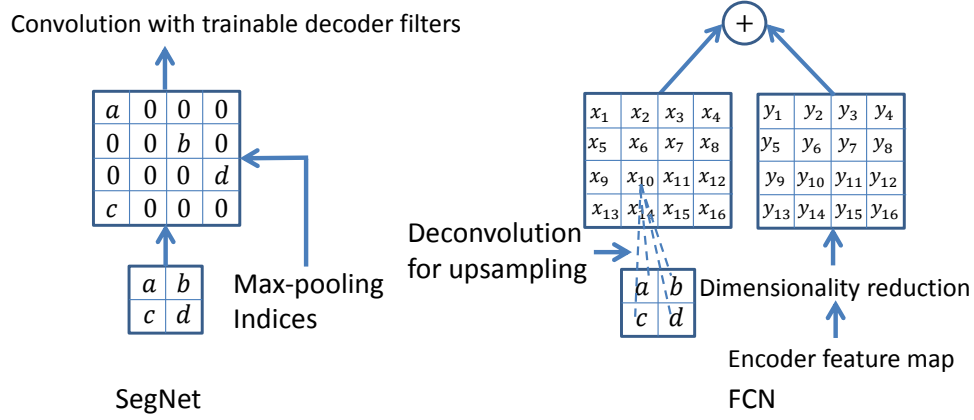


Fig. 3. An illustration of SegNet and FCN [2] decoders. a, b, c, d correspond to values in a feature map. SegNet uses the max pooling indices to upsample (without learning) the feature map(s) and convolves with a trainable decoder filter bank. FCN upsamples by learning to deconvolve the input feature map and adds the corresponding encoder feature map to produce the decoder output. This feature map is the output of the max-pooling layer (includes sub-sampling) in the corresponding encoder. Note that there are no trainable decoder filters in FCN.

Variant	Params (M)	Encoder storage (MB)	Infer time (ms)	Median frequency balancing						Natural frequency balancing					
				Test			Train			Test			Train		
				G	C	I/U	G	C	I/U	G	C	I/U	G	C	I/U
Fixed upsampling															
Bilinear-Interpolation	0.625	0	24.2	77.9	61.1	43.3	89.1	90.2	82.7	82.7	52.5	43.8	93.5	74.1	59.9
Upsampling using max-pooling indices															
SegNet-Basic	1.425	1x	52.6	82.7	62.0	47.7	94.7	96.2	92.7	84.0	54.6	46.3	96.1	83.9	73.3
SegNet-Basic-EncoderAddition	1.425	64x	53.0	83.4	63.6	48.5	94.3	95.8	92.0	84.2	56.5	47.7	95.3	80.9	68.9
SegNet-Basic-SingleChannelDecoder	0.625	1x	33.1	81.2	60.7	46.1	93.2	94.8	90.3	83.5	53.9	45.2	92.6	68.4	52.8
Learning to upsample (bilinear initialisation)															
FCN-Basic	0.65	11x	24.2	81.7	62.4	47.3	92.8	93.6	88.1	83.9	55.6	45.0	92.0	66.8	50.7
FCN-Basic-NoAddition	0.65	n/a	23.8	80.5	58.6	44.1	92.5	93.0	87.2	82.3	53.9	44.2	93.1	72.8	57.6
FCN-Basic-NoDimReduction	1.625	64x	44.8	84.1	63.4	50.1	95.1	96.5	93.2	83.5	57.3	47.0	97.2	91.7	84.8
FCN-Basic-NoAddition-NoDimReduction	1.625	0	43.9	80.5	61.6	45.9	92.5	94.6	89.9	83.7	54.8	45.5	95.0	80.2	67.8

TABLE 1

Comparison of decoder variants. We quantify the performance using global (G), class average (C) and mean of intersection over union (I/U) metrics. The testing and training accuracies are shown as percentages for both natural frequency and median frequency balanced training loss function. SegNet-Basic performs at the same level as FCN-Basic but requires only storing max-pooling indices and is therefore more memory efficient during inference. Note that the theoretical memory requirement reported is based only on the size of the first layer encoder feature map. Networks with larger decoders and those using the encoder feature maps in full perform best, although they are least efficient.

mini-batch (12 images) is then picked in order thus ensuring that each image is used only once in an epoch. We select the model which performs highest on a validation dataset.

We use the cross-entropy loss [2] as the objective function for training the network. The loss is summed up over all the pixels in a mini-batch. When there is large variation in the number of pixels in each class in the training set (e.g road, sky and building pixels dominate the CamVid dataset) then there is a need to weight the loss differently based on the true class. This is termed *class balancing*. We use *median frequency balancing* [13] where the weight assigned to a class in the loss function is the ratio of the median of class frequencies computed on the entire training set divided by the class frequency. This implies that larger classes in the training set have a weight smaller than 1 and the weights of the smallest classes are the highest. We also experimented with training the different variants without class balancing or equivalently using *natural frequency balancing*.

3.3 Analysis

To compare the quantitative performance of the different decoder variants, we use three commonly used performance measures: global accuracy (G) which measures the percentage of pixels

correctly classified in the dataset, class average accuracy (C) is the mean of the predictive accuracy over all classes and mean intersection over union (I/U) over all classes as used in the Pascal VOC12 challenge [19]. The mean I/U metric is the hardest metric since it penalizes false positive predictions unlike class average accuracy. However, I/U metric is not optimized for directly through the class balanced cross-entropy loss.

We test each variant after each 1000 iterations of optimization on the CamVid validation set until the training loss converges. With a training mini-batch size of 12 this corresponds to testing approximately every 33 epochs (passes) through the training set. We select the iteration wherein the global accuracy is highest amongst the evaluations on the validation set. We report all the three measures of performance at this point on the held-out CamVid test set. Although we use class balancing while training the variants, it is still important to achieve high global accuracy to result in an overall smooth segmentation. Another reason is that the contribution of segmentation towards autonomous driving is mainly for delineating classes such as roads, buildings, side-walk, sky. These classes dominate the majority of the pixels in an image and a high global accuracy corresponds to good segmentation of these important classes. We also observed that reporting the

numerical performance when class average is highest can often correspond to low global accuracy indicating a perceptually noisy segmentation output.

In Table 1 we report the numerical results of our analysis. We also show the size of the trainable parameters and the highest resolution feature map or pooling indices storage memory, i.e., of the first layer feature maps after max-pooling and sub-sampling. We show the average time for one forward pass with our Caffe implementation, averaged over 50 measurements using a 360×480 input on an NVIDIA Titan GPU with cuDNN v3 acceleration. We note that the upsampling layers in the SegNet variants are not optimised using cuDNN acceleration. We show the results for both testing and training for all the variants at the selected iteration. The results are also tabulated without class balancing (natural frequency) for training and testing accuracies. Below we analyse the results with class balancing.

From the Table 1, we see that bilinear interpolation based upsampling without any learning performs the worst based on all the three measures of accuracy. All the other methods which either use learning for upsampling (FCN-Basic and variants) or learning decoder filters after upsampling (SegNet-Basic and its variants) perform significantly better. This emphasizes the need to learn decoders for segmentation. This is also supported by experimental evidence gathered by other authors when comparing FCN with SegNet-type decoding techniques [9].

When we compare SegNet-Basic and FCN-Basic we see that both perform equally well on this test over all the three measures of accuracy. The difference is that SegNet uses less **memory** during inference since it only stores max-pooling indices. On the other hand FCN-Basic stores **encoder feature maps** in full which consumes much more memory (11 times more). SegNet-Basic has a decoder with 64 feature maps in each decoder layer. In comparison FCN-Basic, which uses dimensionality reduction, has fewer (11) feature maps in each decoder layer. This reduces the number of convolutions in the decoder network and hence FCN-Basic is faster during inference (forward pass). From another perspective, the decoder network in SegNet-Basic makes it overall a larger network than FCN-Basic. This endows it with more flexibility and hence achieves higher training accuracy than FCN-Basic for the same number of iterations. Overall we see that SegNet-Basic has an advantage over FCN-Basic when memory during inference is constrained but where inference time can be compromised to an extent.

SegNet-Basic is most similar to FCN-Basic-NoAddition in terms of their decoders, although the decoder of SegNet is larger. Both learn to produce dense feature maps, either directly by learning to perform deconvolution as in FCN-Basic-NoAddition or by first upsampling and then convolving with trained decoder filters. The performance of SegNet-Basic is superior, in part due to its larger decoder size. Now, the accuracy of FCN-Basic-NoAddition is also lower as compared to FCN-Basic. This shows that it is important to capture the information present in the encoder feature maps for better performance. This can also explain the part of the reason why SegNet-Basic outperforms FCN-Basic-NoAddition.

The size of the FCN-Basic-NoAddition-NoDimReduction model is slightly larger than SegNet-Basic and this makes it a fair comparison. The performance of this FCN variant is poorer than SegNet-Basic in test but also its training accuracy is lower for the same number of training epochs. This shows that using a larger decoder is not enough but it is also important to capture encoder feature map information to learn better. Here it is also interesting

to see that SegNet-Basic has a competitive training accuracy when compared to larger models such FCN-Basic-NoDimReduction.

Another interesting comparison between FCN-Basic-NoAddition and SegNet-Basic-SingleChannelDecoder shows that using max-pooling indices for upsampling and an overall larger decoder leads to better performance. This also lends evidence to SegNet being a good architecture for segmentation, particularly when there is a need to find a compromise between storage cos, accuracy versus inference time. In the best case, when both memory and inference time is not constrained, larger models such as FCN-Basic-NoDimReduction and SegNet-EncoderAddition are both more accurate than the other variants. Particularly, discarding dimensionality reduction in the FCN-Basic model leads to the best performance amongst the FCN-Basic variants. This once again emphasizes the trade-off involved between memory and accuracy in segmentation architectures.

The last two columns of Table 1 show the result when no class balancing is used (natural frequency). Here, we can observe that without weighting the results are poorer for all the variants, particularly for class average accuracy and mean I/U metric. The global accuracy is the highest without weighting since the majority of the scene is dominated by sky, road and building pixels. Apart from this all the inference from the comparative analysis of variants holds true for natural frequency balancing too. SegNet-Basic performs as well as FCN-Basic and is better than the larger FCN-Basic-NoAddition-NoDimReduction. The bigger but less efficient models FCN-Basic-NoDimReduction and SegNet-EncoderAddition perform better than the other variants.

We can now summarize the above analysis with the following general points.

- 1) The best performance is achieved when encoder feature maps are stored in full.
- 2) When memory during inference is constrained, then compressed forms of encoder feature maps (dimensionality reduction, max-pooling indices) can be stored and used with an appropriate decoder (e.g. SegNet type) to improve performance.
- 3) Larger decoders increase performance for a given encoder network.

4 BENCHMARKING

We quantify the performance of SegNet on three different benchmarks using our Caffe implementation³. Through this process we demonstrate the efficacy of SegNet for various scene segmentation tasks which have practical applications. In the first experiment, we test the performance of SegNet on the CamVid road scene dataset (see Sec. 3.2 for more information about this data). We use this result to compare SegNet with several methods including Random Forests [25], Boosting [25], [27] in combination with CRF based methods [28]. We also trained SegNet on a larger dataset of road scenes collected from various publicly available datasets [3], [53], [54] and show that this leads to a large improvement in accuracy.

SUN RGB-D [21] is a very challenging and large dataset of indoor scenes with 5285 training and 5050 testing images. The images are captured by different sensors and hence come in various resolutions. The task is to segment 37 indoor scene classes including wall, floor, ceiling, table, chair, sofa etc. This task is

3. Our web demo and Caffe implementation is available for evaluation at <http://mi.eng.cam.ac.uk/projects/segnet/>

made hard by the fact that object classes come in various shapes, sizes and in different poses. There are frequent partial occlusions since there are typically many different classes present in each of the test images. These factors make this one of the hardest segmentation challenges. We only use the RGB modality for our training and testing. Using the depth modality would necessitate architectural modifications/redesign [2]. Also the quality of depth images from current cameras require careful post-processing to fill-in missing measurements. They may also require using fusion of many frames to robustly extract features for segmentation. Therefore we believe using depth for segmentation merits a separate body of work which is not in the scope of this paper. We also note that an earlier benchmark dataset NYUv2 [23] is included as part of this dataset.

Pascal VOC12 [19] is a RGB dataset for segmentation with 12031 combined training and validation images of indoor and outdoor scenes. The task is to segment 21 classes such as bus, horse, cat, dog, boat from a varied and large background class. The foreground classes often occupy a small part of an image. The evaluation is performed remotely on 1456 images.

In all three benchmark experiments, we scale the images to 224×224 resolution for training and testing. We used SGD with momentum to train SegNet. The learning rate was fixed to 0.001 and momentum to 0.9. The mini-batch size was 4. The optimization was performed for 100 epochs and then tested.

4.1 CamVid Road Scenes

A number of outdoor scene datasets are available for semantic parsing [3], [24], [55], [56]. Of these we choose to benchmark SegNet using the CamVid dataset [3] as it contains video sequences. This enables us to compare our proposed architecture with those which use motion and structure [26], [27], [28] and video segments [31]. We also combine [3], [24], [55], [56] to form an ensemble of 3433 images to train SegNet for an additional benchmark. For a web demo (see footnote 3) of road scene segmentation, we include the CamVid test set to this larger dataset.

The qualitative comparisons of SegNet-Basic and SegNet predictions with several well known algorithms (unaries, unaries+CRF) are shown in Fig. 4 along with the input modalities used to train the methods. The qualitative results show the ability of the proposed architectures to segment small (cars, pedestrians, bicyclist) classes while producing a smooth segmentation of the overall scene. The unary only methods like Random Forests, Boosting which are trained to predict the label of the center-pixel of a small patch produce low quality segmentations. Smoothing unaries with CRF's improve the segmentation quality considerably with higher order CRF's performing best. Although the CRF based results appear smooth, upon close scrutiny, we see that shape segmentation of smaller but important classes such as bicyclists, pedestrians are poor. In addition, natural shapes of classes like trees are not preserved and the details like the wheels of cars are lost. More dense CRF models [57] can be better but with additional cost of inference. SegNet-Basic and SegNet (without large dataset training) clearly indicate their ability to retain the natural shapes of classes such as bicyclists, cars, trees, poles etc better than CRF based approaches. The overall segmentation quality is also smooth except for the side-walk class. This is because the side-walk class is highly varied in terms of size and texture and this cannot be captured with a small training set. Another explanation could be that the size of the receptive fields of the deepest layer

feature units are smaller than their theoretical estimates [11], [58] and hence unable to group all the side-walk pixels into one class. Illumination variations also affect performance on cars in the dusk examples. However, several of these issues can be ameliorated by using larger amounts of training data. In particular, we see that smaller classes such as pedestrians, cars, bicyclists, column-pole are segmented better than other methods in terms of shape retention. The side-walk class is also segmented smoothly.

The quantitative results in Table 2 show SegNet-Basic, SegNet obtain competitive results even without CRF based processing. This shows the ability of the deep architecture to extract meaningful features from the input image and map it to accurate and smooth class segment labels. SegNet is better in performance than SegNet-Basic although trained with the same (small) training set. This indicates the importance of using pre-trained encoder weights and choice of the optimization technique. Interestingly, the use of the bigger and deeper SegNet architecture improves the accuracy of the larger classes as compared to SegNet-Basic and not the smaller classes as one might expect. We also find that SegNet-Basic [12] trained in a layer-wise manner using L-BFGS [59] also performs competitively and is better than SegNet-Basic trained with SGD (see Sec. 3.2). This is an interesting training approach but needs further research in order for it scale to larger datasets.

The most interesting result is the approximately 15% performance improvement in class average accuracy that is obtained when a large training dataset, obtained by combining [3], [24], [55], [56], is used to train SegNet. The mean of intersection over union metric is also very high. Correspondingly, the qualitative results of SegNet (see Fig. 4) are clearly superior to the rest of the methods. It is able to segment both small and large classes well. In addition, there is an overall smooth quality of segmentation much like what is typically obtained with CRF post-processing. Although the fact that results improve with larger training sets is not surprising, the percentage improvement obtained using pre-trained encoder network and this training set indicates that this architecture can potentially be deployed for practical applications. Our random testing on urban and highway images from the internet (see Fig. 1) demonstrates that SegNet can *absorb* a large training set and generalize well to unseen images. It also indicates the contribution of the prior (CRF) can be lessened when sufficient amount of training data is made available.

4.2 SUN RGB-D Indoor Scenes

Road scene images have limited variation, both in terms of the classes of interest and their spatial arrangements, especially when captured from a moving vehicle. In comparison, images of indoor scenes are more complex since the view points can vary a lot and there is less regularity in both the number of classes present in a scene and their spatial arrangement. Another difficulty is caused by the widely varying sizes of the object classes in the scene. Some test samples from the recent SUN RGB-D dataset [21] are shown in Fig. 5. We observe some scenes with few large classes and some others with dense clutter (bottom row and right). The appearance (texture and shape) can also widely vary in indoor scenes. Therefore, we believe this is the hardest challenge for segmentation architectures and methods in computer vision. Other challenges such as Pascal VOC12 [19] salient object segmentation have occupied researchers more, but indoor scene segmentation is more challenging and has more practical applications such as in robotics. To encourage more research in this direction we set a new benchmark on the large SUN RGB-D dataset.

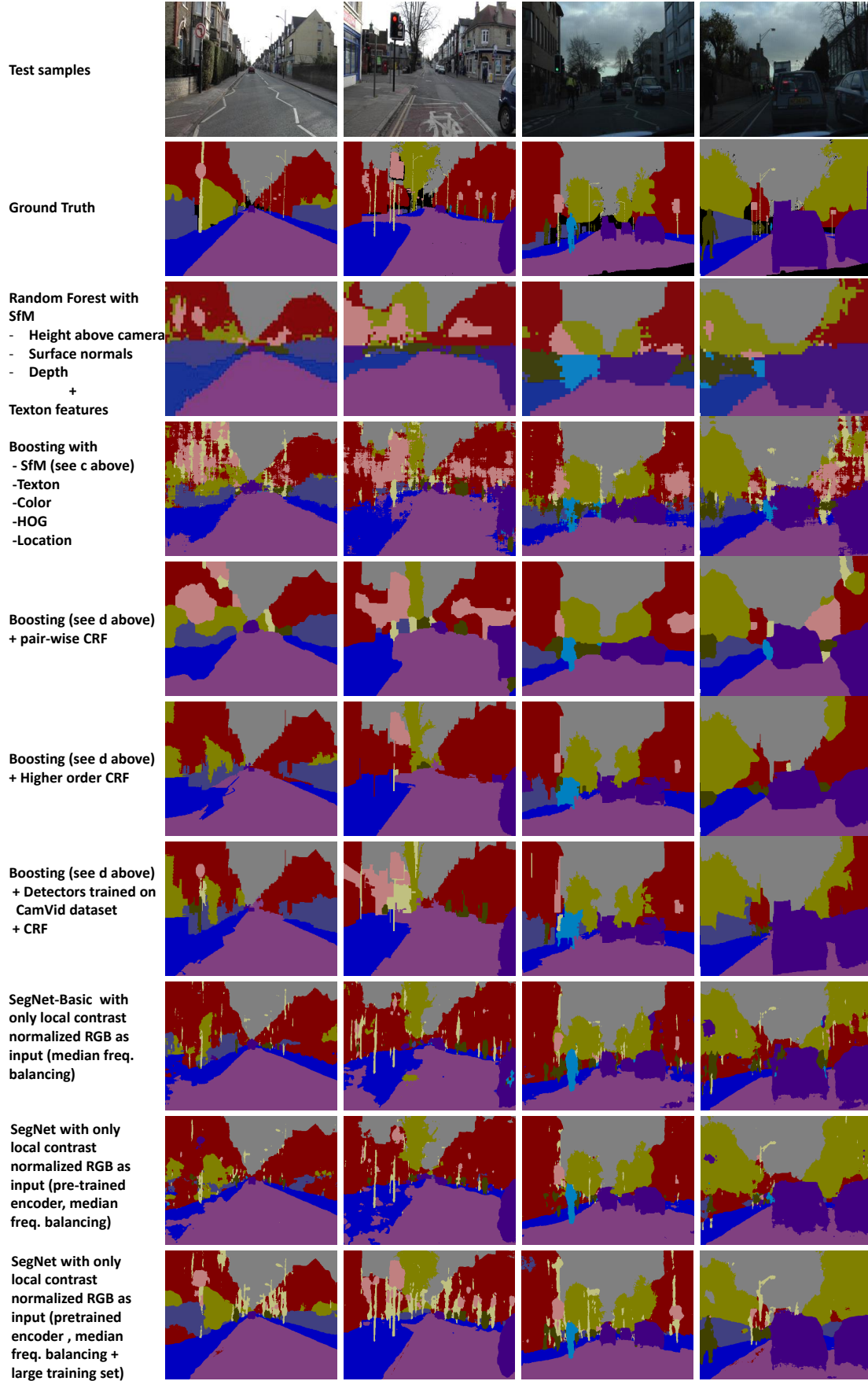


Fig. 4. Results on CamVid day and dusk test samples. The evolution of results from various patch based predictions [25], [26], then combined with CRF smoothing models [27], [28]. Unlike CRF based methods, SegNet-Basic and SegNet predictions retain the shape of small categories such as poles (column 2,4), bicyclist (column 3), far side side-walk (column 2) better. Large sized side-walk is not smoothly segmented possibly because the empirical size of the receptive fields of deep feature units is not large enough [11], [58]. When SegNet is trained on a large dataset it produces the most accurate predictions with an overall smooth appearance.

Method	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Side-walk	Bicyclist	Class avg.	Global avg.	Mean I/U
SfM+Appearance [26]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1	n/a
Boosting [27]	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4	n/a
Dense Depth Maps [30]	85.3	57.3	95.4	69.2	46.5	98.5	23.8	44.3	22.0	38.1	28.7	55.4	82.1	n/a
Structured Random Forests [29]	n/a											51.4	72.5	n/a
Neural Decision Forests [60]	n/a											56.1	82.1	n/a
Local Label Descriptors [61]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	36.3	73.6	n/a
Super Parsing [31]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3	n/a
SegNet-Basic	81.3	72.0	93.0	81.3	14.8	93.3	62.4	31.5	36.3	73.7	42.6	62.0	82.7	47.7
SegNet-Basic (layer-wise training [12])	75.0	84.6	91.2	82.7	36.9	93.3	55.0	37.5	44.8	74.1	16.0	62.9	84.3	n/a
SegNet	88.8	87.3	92.4	82.1	20.5	97.2	57.1	49.3	27.5	84.4	30.7	65.2	88.5	55.6
SegNet (3.5K dataset training)	73.9	90.6	90.1	86.4	69.8	94.5	86.8	67.9	74.0	94.7	52.9	80.1	86.7	60.4
CRF based approaches														
Boosting + pairwise CRF [27]	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8	n/a
Boosting+Higher order [27]	84.5	72.6	97.5	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8	n/a
Boosting+Detectors+CRF [28]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	62.5	83.8	n/a

TABLE 2

Quantitative results on CamVid [3] consisting of 11 road scene categories. SegNet outperforms all the other methods, including those using depth, video and/or CRF's. In comparison with the CRF based methods SegNet predictions are more accurate in 8 out of the 11 classes. It also shows a good $\approx 15\%$ improvement in class average accuracy when trained on a large dataset of 3.5K images and this sets a new benchmark for the majority of the individual classes. Particularly noteworthy are the significant improvements in accuracy for the smaller/thinner classes.

The qualitative results of SegNet on some images of indoor scenes of different types such as bedroom, kitchen, bathroom, classroom etc. are shown in Fig. 5. We see that SegNet obtains sharp boundaries between classes when the scene consists of reasonable sized classes but even when view point changes are present (see bed segmentation from different view points). This is particularly interesting since the input modality is only RGB. It indicates the ability of SegNet to extract features from RGB images which are useful for view-point invariant segmentation provided there is sufficient training data (here 5285 images). RGB images are also useful to segment thinner structures such as the legs of chairs and tables, lamps which is difficult to achieve using depth images from currently available sensors. It is also useful to segment decorative objects such as paintings on the wall.

In Table 4 we report the quantitative results on the 37 class segmentation task. We first note here that the other methods that have been benchmarked are not based on deep architectures and they only report class average accuracy. The existing top performing method [32] relies on hand engineered features using colour, gradients and surface normals for describing super-pixels and then smooths super-pixel labels with a CRF. For SegNet we achieve a high global accuracy which correlates with an overall smooth segmentation. This is also an indicator that largest classes such as wall, floor, bed, table, sofa are segmented well in spite of view-point changes and appearance variations. However, the class average accuracy and mean I/U metric are poor, but at the same level as the hand engineered method which also includes the depth channel as input. This shows that smaller and thinner classes which have lesser training data are not segmented well. The individual class accuracies are reported in Table 3. From these we see that there is a clear correlation between the size and natural frequency of occurrence of classes and their individual accuracies. It is also informative to note RGB input is useful to segment widely varying (shape, texture) categories such as wall, floor, ceiling, table, chair, sofa with reasonable accuracy.

Method	Global avg.	Class avg.	Mean I/U
RGB			
Liu <i>et al.</i> [62]	n/a	9.3	n/a
SegNet	70.3	35.6	26.3
RGB-D			
Liu <i>et al.</i> [62]	n/a	10.0	n/a
Ren <i>et. al</i> [32]	n/a	36.3	n/a

TABLE 4

Quantitative comparison on the SUN RGB-D dataset which consists of 5050 test images of indoor scenes with 37 classes. SegNet RGB based predictions have a high global accuracy and also matches the RGB-D based predictions [32] in terms of class average accuracy.

4.3 Pascal VOC12 Segmentation Challenge

The Pascal VOC12 segmentation challenge [19] consists of segmenting a few salient object classes from a widely varying background class. It is unlike the segmentation for scene understanding benchmarks described earlier which require learning both classes and their spatial context. A number of techniques have been proposed based on this challenge which are increasingly more accurate and complex ⁴. Our efforts in this benchmarking experiment have not been diverted towards attaining the top rank by either using multi-stage training [2], other datasets for pre-training such as MS-COCO [10], [40], training and inference aids such as object proposals [9], [22] or post-processing using CRF based methods [9], [14]. Although these supporting techniques clearly have value towards increasing the performance it unfortunately does not reveal the true performance of the deep architecture which is the *core segmentation engine*. It however does indicate that some of the large deep networks are difficult to train end-to-end on this task even with pre-trained encoder weights. Therefore, to encourage more controlled benchmarking, we trained SegNet end-to-end without other aids and report this performance.

In Table 5 we show the class average accuracy for some recent methods based on deep architectures. To the best of our ability, we

4. See the leader board at <http://host.robots.ox.ac.uk:8080/leaderboard>

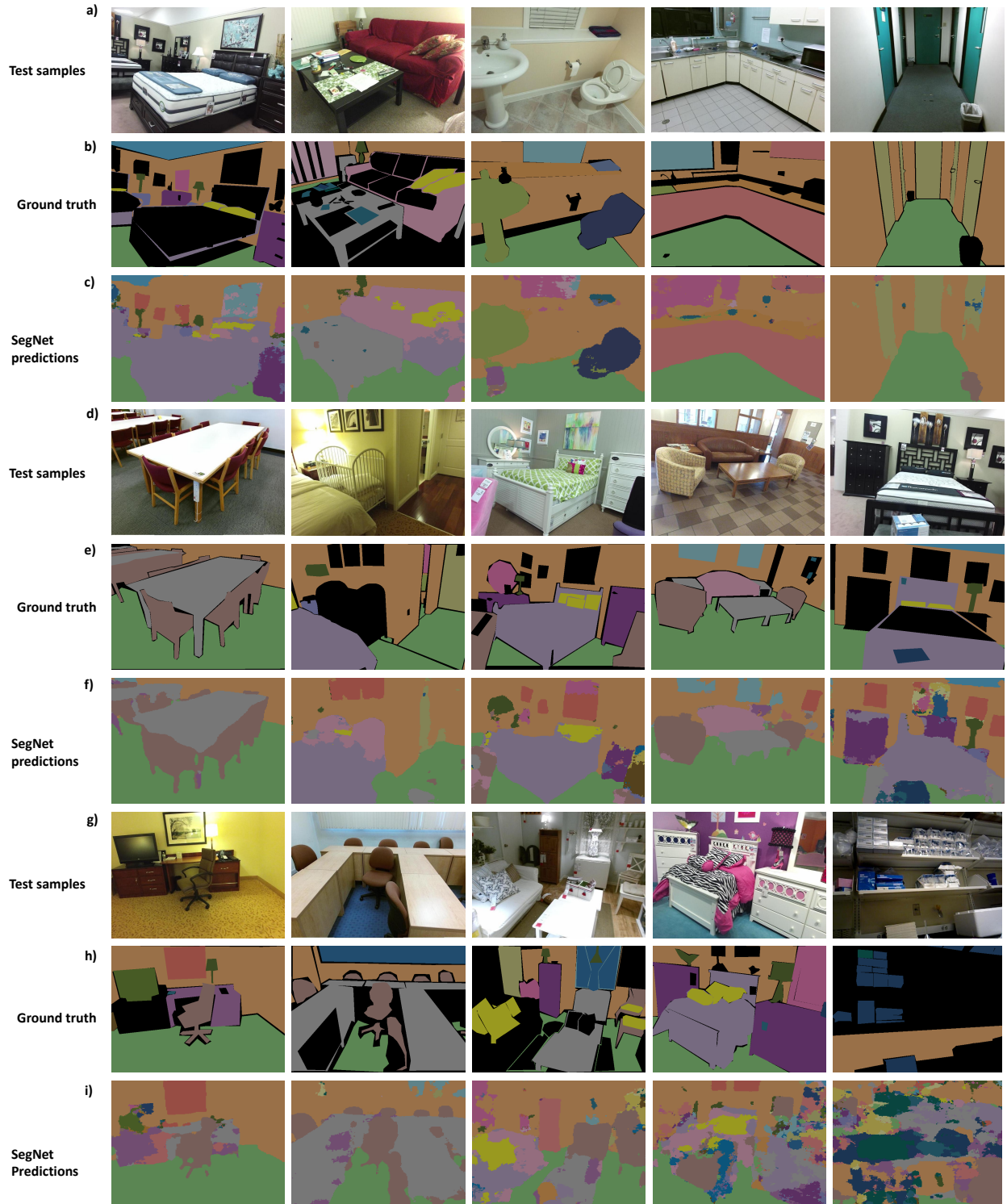


Fig. 5. Qualitative assessment of SegNet predictions on RGB indoor test scenes from the recently released SUN RGB-D dataset [21]. In this hard challenge, SegNet predictions delineate inter class boundaries well for object classes in a variety of scenes and their view-points. The segmentation quality is good when object classes are reasonably sized (rows (c,f)) but suffers when the scene is more cluttered (last two samples in row (i)). Note that often parts of an image of a scene do not have ground truth labels and these are shown in black colour. These parts are not masked in the corresponding SegNet predictions that are shown.

Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door	Window	Bookshelf	Picture	Counter	Blinds
86.6	92.0	52.4	68.4	76.0	54.3	59.3	37.4	53.8	29.2	49.7	32.5	31.2
Desk	Shelves	Curtain	Dresser	Pillow	Mirror	Floor mat	Clothes	Ceiling	Books	Fridge	TV	Paper
17.8	5.3	53.2	28.8	36.5	29.6	0.0	14.4	67.7	32.4	10.2	18.3	19.2
Towel	Shower curtain	Box	Whiteboard	Person	Night stand	Toilet	Sink	Lamp	Bathtub	Bag		
11.5	0.0	8.9	38.7	4.9	22.6	55.6	52.7	27.9	29.9	8.1		

TABLE 3

Class average accuracy of SegNet predictions for the 37 indoor scene classes in the SUN RGB-D benchmark dataset.

have tried to gather the performance measures of the competing methods for their runs using minimum supporting techniques. We also specify when a method reports the performance of its core engine on the smaller validation set of 346 images [14]. We find the performance on the full test set to be approximately 1% less as compared to the smaller validation set.

From the results in Table 5 we can see that the best performing networks are either very large (and slow) [9] and/or they use a CRF [10]. The CRF encourages large segments with a single label and this suits the Pascal challenge wherein there are one or two salient objects in the center of the image. This prior also has a larger impact when training data is limited. This is shown in the experiments using CRF-RNN [10] wherein the core FCN-8 model predictions are less accurate without extra training data.

It is interesting that the DeepLab [14] architecture which is simply upsampling the FCN encoder features using bilinear interpolation performs reasonably well (on the validation set). The fact that a coarse segmentation is enough to produce this performance shows that this challenge is unlike scene understanding wherein many classes of varying shape and size need to be segmented.

Methods using object proposals during training and/or inference [9], [43] are very slow in inference time and it is hard to measure their true performance. These aids are necessitated by the very large size of their deep network [9] and also because the Pascal data can also be processed by a detect and segment approach. In comparison, SegNet is smaller by virtue of discarding the fully connected layers in the VGG16 [1]. The authors of DeepLab [14] have also reported little loss in performance by reducing the size of the fully connected layers. The smaller size of SegNet makes end-to-end training possible for benchmarking. Although it may be argued that larger networks perform better, it is at the cost of a complex training mechanism, increased memory and inference time. This makes them unsuitable for real-time applications such as road scene understanding.

The individual class accuracies for SegNet predictions on Pascal VOC12 are shown in Table 6. From this we once again see larger and more frequently occurring classes such as aeroplane, bus, cat etc. have higher accuracy and smaller/thinner classes such as potted plant, bicycle are poorly segmented. We believe more training data [40] can help improve the performance of SegNet.

5 DISCUSSION AND FUTURE WORK

Deep learning models have often achieved increasing success due to the availability of massive datasets and expanding model depth and parameterisation. The use of millions of training parameters in these networks has resulted in performance gains and this effect can be observed in our analysis of various models (see Sec. 3.3). However, less attention has been paid to smaller and more efficient models for real-time applications (of segmentation) such as road scene understanding. This was the primary motivation behind the proposal of SegNet, which is significantly smaller and faster than

other competing architectures, but efficient for tasks such as road scene understanding.

Although in some tasks it may be sufficient to simply replicate the encoder features and post-process with a CRF, it is less elegant and does not exploit the potential of deep learning for feed-forward segmentation. Using SegNet as a candidate architecture, we have shown the need for designing trainable decoders which can learn to map the output of the encoder network to input resolution feature maps for classification. Our controlled analysis reveals the trade-offs involved in designing architectures for segmentation, mostly involving inference memory, time and accuracy. In particular, transferring more (feature) information to the decoders can lead to simpler decoding process but with additional storage costs. When compressed forms of encoder feature maps are stored and transferred to the decoders then, to achieve similar performance, the decoders need to be more complex (more training parameters).

Our experience with benchmarking experiments have shown the efficacy of our proposed architecture for segmentation on both indoor and outdoor scenes. We outperform all the state-of-the-art methods for road scene understanding. Indoor scene understanding is very challenging and SegNet has set a new benchmark on a large dataset for RGB based scene understanding. An important issue we faced while conducting these experiments is the lack of controlled baselines for many of the recent architectures. Many of these architectures have used a host of supporting techniques to arrive at high accuracies on datasets but this makes it difficult to gather conclusive evidence about their true performance. This is further worsened by the fact that different variants of SGD is used to train these architectures and the learning rate is tuned manually by observing the progress of the loss. To avoid this and in the interest of future research, we trained SegNet end-to-end using SGD with a fixed learning rate and momentum throughout the training process for a predefined number of epochs. We also provide our Caffe implementation of SegNet-Basic (and its variants), SegNet and a web demo for evaluation of the real-time road scene segmentation. For the future, we would like to exploit our understanding of segmentation architectures gathered from our analysis to design more efficient architectures for real-time applications. We are also interested in experimenting with Dropout [63] during training and testing. This can help estimate the uncertainty of the predictions for scene understanding.

6 CONCLUSION

We presented SegNet, a deep convolutional network architecture for semantic segmentation. The main motivation behind SegNet was the need to design an efficient architecture for road scene understanding which is efficient both in terms of memory and computational time. We analysed SegNet and compared it with other important variants to reveal the trade-offs involved in designing architectures for segmentation. Those which store the encoder network feature maps in full perform best but consume more

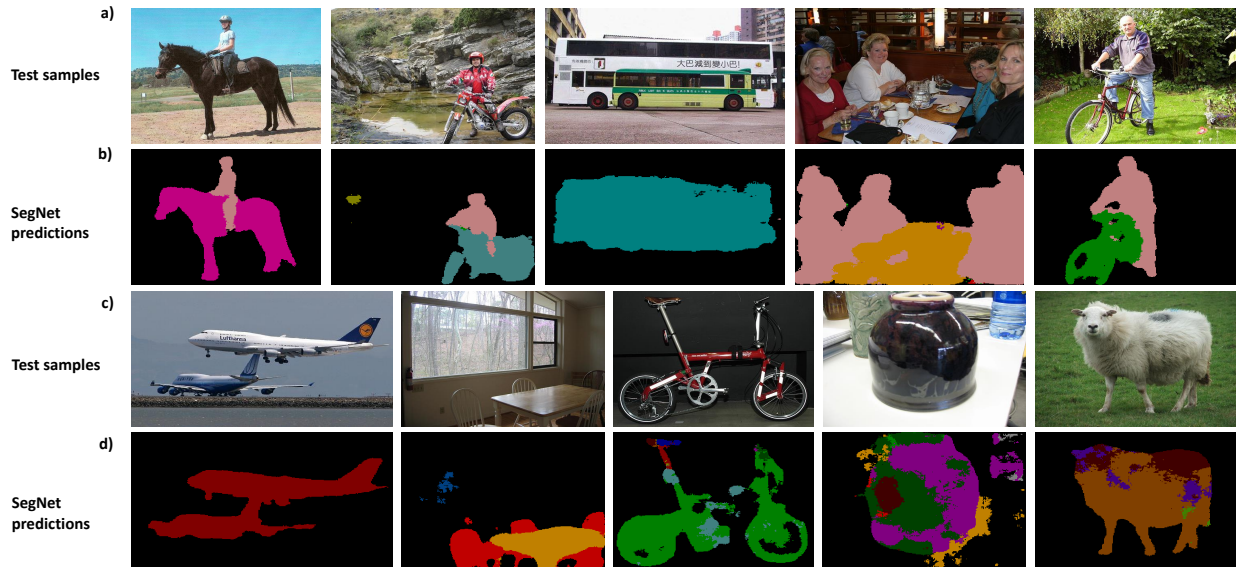


Fig. 6. Qualitative assessment of SegNet predictions on test samples from Pascal VOC12 [19] dataset. SegNet performs competitively (row (b)) on several object classes of varying shape and appearance. However, it lacks smoothness particularly on large objects (see row(d)). This can be perhaps be attributed to the smaller empirical size of the receptive field of the feature units in the deepest encoder layer size [58].

Method	Encoder size (M)	Decoder size (M)	Total size (M)	Class avg. acc.	Inference 500×500 pixels	Inference 224×224 pixels
DeepLab [14] (validation set)	n/a	n/a	< 134.5	58	n/a	n/a
FCN-8 [2] (multi-stage training)	134	0.5	134.5	62.2	210ms	n/a
Hypercolumns [43] (object proposals)	n/a	n/a	> 134.5	62.6	n/a	n/a
DeconvNet [9] (object proposals)	138.35	138.35	276.7	69.6	n/a	92ms ($\times 50$)
CRF-RNN [10] (multi-stage training)	n/a	n/a	> 134.5	69.6	n/a	n/a
SegNet	14.725	14.725	29.45	59.1	94ms	28ms

TABLE 5

Quantitative comparison on Pascal VOC12 dataset. The accuracies for the competing architectures are gathered for their inference run using the least number of supporting training and inference techniques. However, since they are not trained end-to-end like SegNet and use aids such as object proposals, we have added corresponding qualifying comments. The first three columns show the number of trainable parameters in the encoder, decoder and full network. Many of the models are approximately the same size as FCN. In comparison, SegNet is considerably smaller but achieves a competitive accuracy without resorting to supporting training or inference aids. This results in SegNet being significantly faster than other models in terms of inference time.

Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining table
74.5	30.6	61.4	50.8	49.8	76.2	64.3	69.7	23.8	60.8	54.7
Dog	Horse	Motor bike	Person	Potted plant	Sheep	Sofa	Train	TV	Background	
62.0	66.4	70.2	74.1	37.5	63.7	40.6	67.8	53.0	88.6	

TABLE 6

Individual class accuracies of SegNet predictions on the Pascal VOC12 segmentation benchmark consisting of 21 object classes.

memory during inference time. SegNet on the other hand is more efficient since it only stores the max-pooling indices of the feature maps and uses them in its decoder network to achieve good performance. However, the inference time is increased since the decoder network is required to be larger. On large and well known datasets SegNet performs competitively on challenges such as indoor scene understanding and Pascal VOC12. It sets a new benchmark for road scene understanding when trained on a large dataset. With its efficient architecture and competitive performance SegNet is well suited for scene understanding applications. In future, we would also like to estimate the uncertainty of the labelling using Bayesian neural network techniques [64].

REFERENCES

- [1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [2] 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*, pp. 3431–3440, 2015.
- [3] G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *PRL*, vol. 30(2), pp. 88–97, 2009.
- [4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.
- [5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
- [6] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE PAMI*, vol. 35, no. 8, pp. 1915–1929, 2013.
- [7] N. Hft, H. Schulz, and S. Behnke, "Fast semantic segmentation of rgb-d scenes with gpu-accelerated deep neural networks," in *KI 2014: Advances in Artificial Intelligence* (C. Lutz and M. Thielscher, eds.), vol. 8736 of *Lecture Notes in Computer Science*, pp. 80–85, Springer International Publishing, 2014.
- [8] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *ICML*, pp. 129–

- 136, 2011.
- [9] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," *CoRR*, vol. abs/1505.04366, 2015.
 - [10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," *arXiv preprint arXiv:1502.03240*, 2015.
 - [11] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," *CoRR*, vol. abs/1506.04579, 2015.
 - [12] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *CoRR*, vol. abs/1505.07293, 2015.
 - [13] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *arXiv preprint arXiv:1411.4734*, 2014.
 - [14] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014.
 - [15] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186, Springer, 2010.
 - [16] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," *CoRR*, vol. abs/1506.04924, 2015.
 - [17] M. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *CVPR*, 2007.
 - [18] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, et al., "The role of context for object detection and semantic segmentation in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 891–898, IEEE, 2014.
 - [19] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136.
 - [20] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 991–998, IEEE, 2011.
 - [21] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015.
 - [22] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision–ECCV 2014*, pp. 391–405, Springer, 2014.
 - [23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *ECCV*, pp. 746–760, Springer, 2012.
 - [24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, pp. 3354–3361, 2012.
 - [25] J. Shotton, M. Johnson, and R. Cipolla, "Semantic textron forests for image categorization and segmentation," in *CVPR*, 2008.
 - [26] G. Brostow, J. Shotton, J., and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV, Marseille*, 2008.
 - [27] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr, "Combining appearance and structure from motion features for road scene understanding," in *BMVC*, 2009.
 - [28] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr, "What, where and how many? combining object detectors and crfs," in *ECCV*, pp. 424–437, 2010.
 - [29] P. Kotschieder, S. R. Bulo, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *ICCV*, pp. 2190–2197, IEEE, 2011.
 - [30] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *ECCV*, pp. 708–721, Springer, 2010.
 - [31] J. Tighe and S. Lazebnik, "Superparsing," *IJCV*, vol. 101, no. 2, pp. 329–349, 2013.
 - [32] X. Ren, L. Bo, and D. Fox, "Rgb-d scene labeling: Features and algorithms," in *CVPR*, pp. 2759–2766, IEEE, 2012.
 - [33] A. Hermans, G. Floros, and B. Leibe, "Dense 3D Semantic Mapping of Indoor Scenes from RGB-D Images," in *ICRA*, 2014.
 - [34] S. Gupta, P. Arbeláez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *CVPR*, pp. 564–571, IEEE, 2013.
 - [35] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," in *ICML*, 2012.
 - [36] D. Grangier, L. Bottou, and R. Collobert, "Deep convolutional networks for scene parsing," in *ICML Workshop on Deep Learning*, 2009.
 - [37] C. Gatta, A. Romero, and J. van de Weijer, "Unrolling loopy top-down semantic feedback in convolutional deep networks," in *CVPR Workshop on Deep Vision*, 2014.
 - [38] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *ICML*, pp. 82–90, 2014.
 - [39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, pp. 1–42, April 2015.
 - [40] 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 *Computer Vision–ECCV 2014*, pp. 740–755, Springer, 2014.
 - [41] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," *arXiv preprint arXiv:1503.02351*, 2015.
 - [42] G. Lin, C. Shen, I. Reid, et al., "Efficient piecewise training of deep structured models for semantic segmentation," *arXiv preprint arXiv:1504.01013*, 2015.
 - [43] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," *CoRR*, vol. abs/1411.5752, 2014.
 - [44] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," *arXiv preprint arXiv:1412.0774*, 2014.
 - [45] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *CVPR*, pp. 2528–2535, IEEE, 2010.
 - [46] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. LeCun, "Learning convolutional feature hierarchies for visual recognition," in *NIPS*, pp. 1090–1098, 2010.
 - [47] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, pp. 184–199, Springer, 2014.
 - [48] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *arXiv preprint arXiv:1406.2283*, 2014.
 - [49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.
 - [50] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *ICCV*, pp. 2146–2153, 2009.
 - [51] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv preprint arXiv:1502.01852*, 2015.
 - [52] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
 - [53] A. Torralba, B. C. Russell, and J. Yuen, "Labelme: online image annotation and applications," tech. rep., MIT CSAIL Technical Report, 2009.
 - [54] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. Lopez, "Vision-based offline-online perception paradigm for autonomous driving," in *WACV*, 2015.
 - [55] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *ICCV*, pp. 1–8, IEEE, 2009.
 - [56] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *IJCV*, vol. 77, no. 1-3, pp. 157–173, 2008.
 - [57] V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *In: NIPS (2011)*, 2011.
 - [58] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.
 - [59] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer, 2nd ed., 2006.
 - [60] S. Bulo, S. Rota, and P. Kotschieder, "Neural decision forests for semantic image labelling," in *CVPR*, 2014.
 - [61] Y. Yang, Z. Li, L. Zhang, C. Murphy, J. Ver Hoeve, and H. Jiang, "Local label descriptor for example based semantic image labeling," in *ECCV*, pp. 361–375, Springer, 2012.
 - [62] C. Liu, J. Yuen, A. Torralba, and J. Sivic, "Sift flow: Dense correspondence across different scenes," in *ECCV, Marseille*, 2008.
 - [63] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Insights and applications," in *Deep Learning Workshop, ICML*, 2015.
 - [64] A. Kendall, V. Badrinarayanan and R. Cipolla, "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding," *arXiv preprint arXiv:1511.02680*, 2015.