Semantics-aware Visual Localization under Challenging Perceptual Conditions

Tayyab Naseer Gabriel L. Oliveira Thomas Brox Wolfram Burgard

***Abstract*— Visual place recognition under difficult perceptual conditions remains a challenging problem due to changing weather conditions, illumination and seasons. Long-term visual navigation approaches for robot localization should be robust to these dynamics of the environment. Existing methods typically leverage feature descriptions of whole images or image regions from Deep Convolutional Neural Networks. Some approaches also exploit sequential information to alleviate the problem of spatially inconsistent and non-perfect image matches. In this paper, we propose a novel approach for learning a discriminative holistic image representation which exploits the image content to create a dense and salient scene description. These salient descriptions are learnt over a variety of datasets under large perceptual changes. Such an approach enables us to precisely segment the regions of an image which are geometrically stable over large time lags. We combine features from these salient regions and an off-the-shelf holistic representation to form a more robust scene descriptor. We also introduce a semantically labeled dataset which captures extreme perceptual and structural scene dynamics over the course of 3 years. We evaluated our approach with extensive experiments on data collected over several kilometers in Freiburg and show that our learnt image representation outperforms off-the-shelf features from the deep networks and hand-crafted features.**

I. INTRODUCTION

Robust camera-only navigation in difficult outdoor environments is a challenging problem. The environments undergo changes due to different weather conditions, seasons, construction and dynamic objects. Lifelong operation of autonomous robots require localization approaches to be robust to all these appearance changes. Traditionally, approaches relied on point descriptors such as SIFT [15] for place recognition. Recently, generic image descriptors extracted from Deep Convolutional Neural Networks (DCNNs) have shown to outperform the traditional point-descriptors. Recent research has shown that these descriptors can be combined with various region detectors to achieve impressive robust visual localization [22], [26]. However, these approaches aggregate features from image regions without considering the semantic information encapsulated in them. Linegar *et al.* [12] proposed to learn significant visual elements by discovering discriminative visual patches of a specific map for robust visual tracking.

In this paper, we propose a novel method for learning visually discriminant image regions by training a DCNN in a one-vs-all scheme to achieve robust global localization. Deep learning approaches have achieved astounding results for semantic segmentation. One specific type of convolutional architecture, called up-convolutional Networks surpassed other approaches for such applications [13], [14], [23]. These architectures weigh each class prediction with probabilities in a classification network. Such maps have the desirable characteristic of providing specific regions of interest. In this work, we use these stable regions in order to confine the feature extraction for robust scene description. We build on top of a recently proposed up-convolutional architecture, Fast-Net to learn the regions of interest [23]. We use training examples from the public datasets Cityscapes and Virtual KITTI [4], [7]. These datasets do not cover large seasonal or weather variations. Therefore, we also present a new labeled dataset collected by driving around in

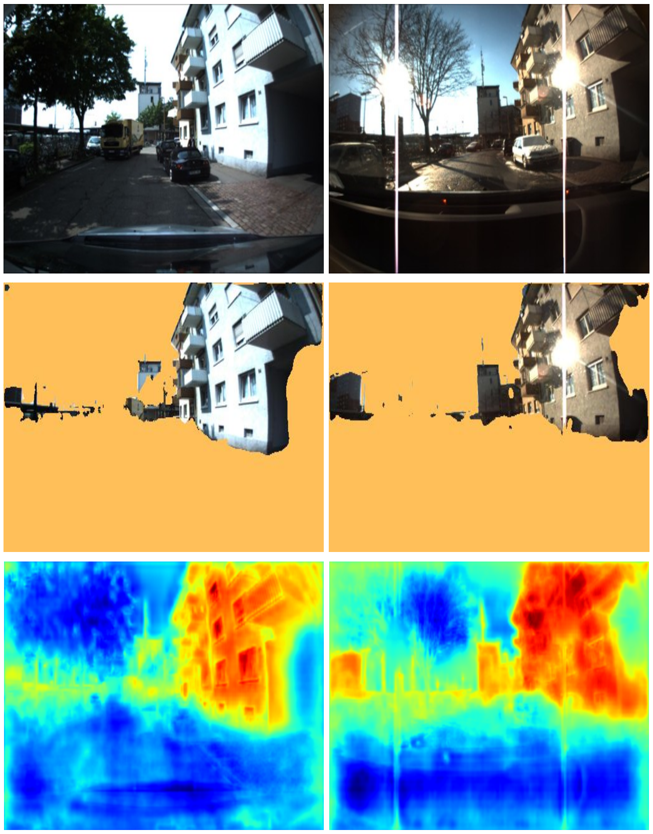


Fig. 1: Our proposed method provides a robust scene description under harsh visual conditions by segmenting geometrically stable image regions (*middle row*). The (*bottom row*) shows the contributing weights of each pixel for the salient region. The red areas correspond to stable parts of the image and the blue areas define the parts of the image which potentially could change over time. This figure is best viewed in color.

All the authors are with the Department of Computer Science, University of Freiburg, 79110, Germany.

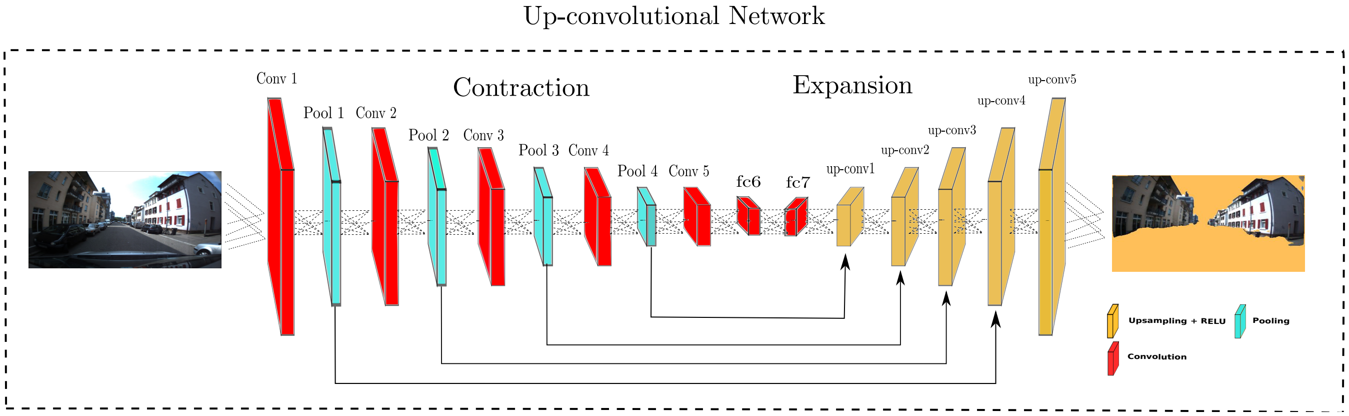


Fig. 2: Fast-Net architecture used in our approach for image segmentation. Only convolutional, pooling, and up-convolutional layers are visualized.

Freiburg City, Germany. The Freiburg dataset contains a wide range of appearance and structural changes spanning over the course of 3 years. We labeled all the potentially non-stable regions of the training images as non-discriminative. These regions correspond to objects which change over different perceptual conditions, e.g., roads, sky, trees and the mercurial objects, like people, vehicles etc. We also labeled all the geometrically stable structures in the scene as discriminative regions. Our trained network provides dense pixel-wise segmentation of the discriminative regions as shown in Fig. 1. Subsequently, we extract deep features from our segmented and original images and aggregate them to form a robust scene descriptor. Our approach is in spirit similar to the point feature-based instance retrieval method [10], which detects common distracting patterns using GPS information and suppresses them for image matching. The advantage of our method is that existing region-based place recognition approaches can be combined with our approach for more accurate and informative image matchings.

Overall our contributions can be summarized as follows:

– We present a learning approach for robust binary segmentation and feature aggregation of deep networks.
– We show that our method outperforms off-the-shelf features from deep networks for robust place recognition over a variety of datasets. Our approach runs online at 14 Hz on a single GPU.
– We present a coarsely labeled dataset for semantic saliency in dynamic and perceptually changing urban environments which captures long-term weather, seasonal, and structural changes.

II. Related Work

Visual place recognition has been studied extensively in both the computer vision and robotics communities. In computer vision, the research is typically focused on instance retrievals [9], [24], where sophisticated point feature-based approaches are used with Bag-of-Words (BoW) model to achieve robust place recognition. Arandjelović *et al.* [2] proposed an approach to take the semantic nature of the visual words into account to prune false matches. Inspired by the confusing feature suppression approach in [10], Mousavian *et al.* [19] proposed a method to leverage semantic segmentation to select feature points only from buildings to improve the precision of the BoW model. These approaches are based on keypoint-based handcrafted feature descriptors and work reasonably well under mild perceptual changes, but with extreme environmental changes as shown in Fig. 1, they do not show much robustness [18], [21]. Badino *et al.* [3] proposed an approach which fuses LiDAR and image data with a particle sequential filter framework to perform loop closure detection across large time lags. McManus *et al.* [17] proposed to learn discriminative patches per place by using a significant amount of aligned training data for each location to perform robust place recognition. Lowry *et al.* [16] proposed an approach to remove the perceptual changes caused by weather using Principal Component Analysis to make the scene descriptor more robust. These methods can also leverage semantic information as proposed in our method to further boost localization performance.

Recently, DCNNs have shown to outperform handcrafted features in various image recognition and classification tasks [11]. Robust descriptors from these deep networks have been successfully applied to place recognition and loop closure detection tasks [27]. A very recent approach [29] combines various feature descriptors in a sequence-based loop closure scheme to achieve precise localization. Their main idea leverages the sparsity of the loop closure problem and they formulate it as an optimization problem to find the optimal solution. In our work, we mainly aim to increase the performance of purely image-based feature matching, which makes it easy to integrate with other methods. Sünderhauf *et al.* proposed to increase the robustness of deep features to viewpoint variance by extracting object features followed by their concatenation instead of holistic descriptors [26]. We believe such approaches can be integrated with our semantic content-aware method to store and match objects which are more relevant for place recognition. We evaluate our approach on datasets collected from real world driving scenarios, hence it also contains all natural occurring driving maneuvers, viewpoint changes and occlusions. Dymczyk *et al.* [6] propose an approach to summarize and update maps for long-term navigation by keeping a minimal number of landmarks in the memory for localization. Such landmarks can also be pruned based on their content aided by their semantic saliency. Therefore, for a robust scene description, we aim to highly weight the features at spatial locations which are geometrically stable over longer periods of time.

III. Technical Approach

In this section, we describe the integral components of our proposed approach: (i) Up-Convolutional network architecture, (ii) training data, (iii) robust feature description, (iv) feature embedding and matching. An illustration of our method is shown in Fig. 3.

A. *Up-Convolutional Network Architecture*

Up-convolutional approaches have recently raised the bar on multiple semantic segmentation tasks [13], [14], [23]. This class of techniques are efficient and capable of end-to-end training. Up-convolutional networks have two parts, contractive and expansive. Given a high resolution input the contractive part produces a low resolution segmentation mask. Usually a network designed for image recognition is fine tuned for such a task, for instance the VGG architecture [25]. The output of the contractive side of the network is still too coarse for many applications, thus returning to the input resolution is needed. In this work, we build upon a recently proposed up-convolutional network called Fast-Net. The architecture is shown in Fig. 2.

Fast-Net is mainly designed to provide near real-time semantic segmentation with no loss in discriminative efficiency. The main characteristics of the network are the refinement layers and the identification and suppression of the computational bottlenecks of up-convolutional architectures. The refinement stage consists of upsampling and convolving low resolution segmentation masks and fusing them with previous pooling layers to obtain higher resolution outputs. The main computational burden of previous architectures are the last two convolutional layers of the contractive side. Previous architectures have four times more filters in these layers and make use of larger filter sizes, i.e., 7×7 compared to 3×3 of Fast-Net.

Network Training: In this work, we do not aim to provide highly precise segmentation for all the objects in an urban environment, rather we aim to provide saliency maps for a place, based mainly on man-made structures. These saliency maps contain information about geometrically stable structures of the place. Therefore, we carry out binary segmentation of the input images.

We define a pixel-wise labeling for an image $\mathcal{I}$ as $(X_n, Y_n), n = 1, \ldots, N$, where $X_n = \{x_i, i = 1, \ldots, |X|\}$ defines the set of image pixels, $Y_n = \{y_i, i = 1, \ldots, |Y|\}$, $y_i \in \{0, 1\}$ denotes the corresponding ground truth with 2 classes in our case. We denote the input image with $\mathcal{I}_\text{o}$ and the segmented image with $\mathcal{I}_\text{s}$. The activation function of the

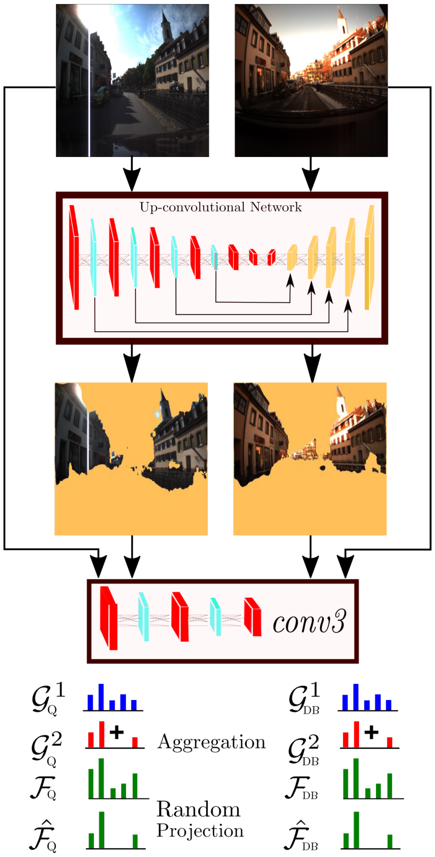


Fig. 3: Our approach segments input images for salient region detection, aggregates deep features from both the images and is followed by sparse random projection for dimensionality reduction.

network is represented as $f(x_i; \theta)$, where θ are the parameters of the network. We perform multi-stage training with one refinement at a time. This training strategy is computationally efficient and yields a less complex learning process. We use off-the-shelf VGG weights to initialize the contractive part of the network. The network is trained by backpropagation using Stochastic Gradient Descent with momentum and it learns the features which minimize the cross entropy loss between the estimated class probabilities and the ground truth. The loss is computed according to Eq. (1) and then we compute θ^* as in Eq. (2) which minimizes this loss function over all the pixels.

$$\mathcal{L}(p, q) = -\sum_{k \in 0,1} p_k \log q_k \tag{1}$$

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N \times |X|} \mathcal{L}((f(x^i; \theta)), y^i) \tag{2}$$

We use a minibatch of size one, hence we use momentum 0.99 to keep as much of the previous gradients as possible. The learning rate is fixed to $1e^{-10}$ due to the non-normalized loss. Our trained network yields a very fine segmentation of the image, so we perform morphological dilation over the segmented binary mask of $\mathcal{I}_\text{s}$ with a structural element $\mathbf{B}$ as in Eq. (3) to relax the tightly bounded segmentation for really thin structures.

$\mathcal{I}_{\mathrm{s}b}$ is the translation of $\mathcal{I}_{\mathrm{s}}$ by $\mathbf{B}$. In our case $\mathbf{B}$ is a 4×4 grid-based structuring element.

$$\mathcal{I}_{\mathrm{s}} \oplus \mathbf{B} = \bigcup_{b \in \mathbf{B}} \mathcal{I}_{\mathrm{s}b} \tag{3}$$

The dilation increases the recall rate of distant, occluded and fine structures.

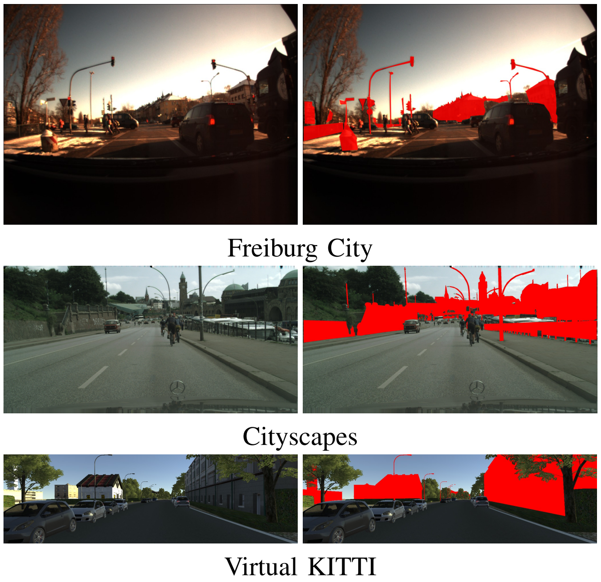
B. Training Data

We use training examples from the two public datasets, Cityscapes and Virtual KITTI and from our new Freiburg dataset[1]. The Cityscapes dataset provides dense pixel-wise semantic labels from real world driving scenarios. The Virtual KITTI dataset is rendered in a game engine and the synthetic images are rendered in close proximity to the real world KITTI dataset [8]. Although, these synthetic environments provide flexibility to render various environmental conditions, these datasets do not provide all the naturally occurring driving conditions and perceptual changes. Hence, we introduce a new semantically labelled dataset (Freiburg City), which models environmental conditions encountered over years, e.g, snow, foliage changes, structural changes, various driving maneuvers, and variable traffic conditions. We use $3900, 700,$ and 430 training examples from the Cityscapes, Virtual KITTI and Freiburg datasets respectively to make our segmentation robust towards various environment types and conditions. The examples of images and the corresponding labels are shown in Fig. 4.

Data Augmentation: Deep networks require a large amount of training data. Due to the low number of training examples, we augmented our data with a series of transformations. We applied scaling, rotation, color distortion, and skew for the augmentation. We also considered other image transformations, e.g., cropping, flipping and vignetting, however, the improvement was marginal.

C. Place Description

We aggregate features from the original and the segmented images to form the final place description. The idea is to boost the activations of the neurons at the salient spatial locations and leverage original activations for images which do not contain enough geometrically stable structure in it. The feature tensor, $\mathcal{G} \in \mathbb{R}^{C \times W \times H}$ is extracted for both the images from the *conv3* layer of a DCNN, as this layer is shown to be most robust for appearance changes [28]. $W = 13$, $H = 13$, and $C = 384$ are the width, height and number of channels of the tensor $\mathcal{G}$ and it has $\approx 65\mathrm{K}$ dimensions. We aggregate both the feature tensors ($\mathcal{G}^1 + \mathcal{G}^2$) to form a robust place descriptor ($\mathcal{F}$). We then L2-normalize $\mathcal{F}$. The feature vector size of the resulting descriptor is large for real-time applications so we project this vector to lower dimensions as explained in the next subsection.



Freiburg City

Cityscapes

Virtual KITTI

Fig. 4: Training images with their corresponding labels. We label all the man-made structures which are geometrically more stable over longer periods of time.

D. Feature Embedding and Matching

The aggregated feature vectors have a large dimensionality, which affects the scalability of the approach. We use Sparse Random Projection [1] for embedding our high-dimensional feature vectors into lower dimensions. The *Johnson-Lindenstrauss Lemma* states that for $m > 4$ samples in $\mathbb{R}^N$, distortion of $\epsilon \in (0, \frac{1}{2}]$, and $k = \frac{20 \log m}{\epsilon^2}$, we can linearly map set $\mathbf{V}$ of m points in $\mathbb{R}^N$ $\exists\ f\ : \mathbb{R}^N \rightarrow \mathbb{R}^k$ s.t $\forall \boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{V}$

$$(1 - \epsilon)\|\boldsymbol{a} - \boldsymbol{b}\|^2 \leq \|f(\boldsymbol{a}) - f(\boldsymbol{b})\|^2 \leq (1 + \epsilon)\|\boldsymbol{a} - \boldsymbol{b}\|^2 \tag{4}$$

The linear mapping function f in this case is $\frac{1}{\sqrt{k}}\boldsymbol{R}$, where $\boldsymbol{R} \in \mathbb{R}^{k \times N}$ is the random projection matrix with i.i.d entries.

$$\boldsymbol{r}_{ij} = \sqrt{3}\begin{cases} 1 & \text{with prob.} \quad \frac{1}{6} \\ 0 & \text{with prob.} \quad \frac{2}{3} \\ -1 & \text{with prob.} \quad \frac{1}{6} \end{cases}$$

This projection is fast and sparse as only one third of the data is projected. We project the high-dimensional feature descriptors onto lower dimensions using Eq. (5).

$$\hat{\mathcal{F}} = \frac{1}{\sqrt{k}}\boldsymbol{R}\mathcal{F} \in \mathbb{R}^k \tag{5}$$

The projected query feature vector $\mathcal{F}_{\mathrm{Q}}$ and database feature vector $\mathcal{F}_{\mathrm{DB}}$ are matched using cosine distance as in Eq. (6)

$$\mathrm{sim}(\hat{\mathcal{F}_{\mathrm{Q}}}, \hat{\mathcal{F}_{\mathrm{DB}}}) = \frac{\hat{\mathcal{F}_{\mathrm{Q}}} \cdot \hat{\mathcal{F}_{\mathrm{DB}}}}{\|\hat{\mathcal{F}_{\mathrm{Q}}}\|\|\hat{\mathcal{F}_{\mathrm{DB}}}\|} \tag{6}$$

For our experiments we chose $k = 4096$, which resulted in significant data reduction and speed-up at the cost of nominal accuracy loss.

[1] We intend to make the dataset publicly available in the near future.

IV. Experiments

We evaluated our approach on an extensive set of perceptually challenging data collected by driving over 100 km in Freiburg, Germany over the span of 3 years. The dataset encompasses a wide range of occlusions, dense traffic conditions, and seasonal variations as well as noise like strong sun glare. The first session consists of a 10 km trajectory recorded in Summer 2012 (**Localization-1**), and another 50 km trajectory recorded in Winter 2012 (**Database**). The second session consists of 40 km trajectory recorded in Summer 2015 (**Localization-2**). Altogether these trajectories contain a set of $\sim$ 45000 images. The trajectories are overlayed on Google Maps as shown in Fig. 6.

A. *Robust Descriptor Evaluation*

In this section, we detail the quantitative evaluation of our robust descriptor and compare it with off-the-shelf feature descriptors from a DCNN [11]. We report best $\mathbf{F_1}$ scores according to

$$\mathbf{F_1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

to quantify the localization performance. These experiments bespeak the effectiveness of content-aware feature boosting for the problem of single-shot place recognition. In all the results, 'Ours' corresponds to our proposed method where we aggregate features from the segmented images and the original images, 'Ours-4096' corresponds to our method with 4096-D descriptors, 'DCNN' corresponds to off-the-shelf feature descriptors, and 'HOG' is the handcrafted-feature proposed by [5]. Please note that no post-processing is applied to the similarity matrix in this work as in [20], so that the approach can be run online instead of batch-processing. The matches are considered to be a true-positive if they lie within ±3 frames of the query location. The quantitative results are summarized in Table I and the precision-recall curves are shown in Fig. 5.

The first dataset (Parallel Blocks) consists of image sequences recorded by driving around in parallel blocks of the city where the spatial global context of different places appear to be similar and would cause perceptual aliasing. It consists of 821 query images and 570 database images. Note that, although objects like trees can be regarded as significant scene signatures, they can cause perceptual aliasing as shown in Fig. 7. Therefore, boosting feature activations from man-made structures would help mitigate such problems. Our proposed method achieves the best $\mathbf{F_1}$ score of 0.59 and the projection of the feature vectors yields the best F1 score of 0.58, hence the loss in accuracy is nominal. We achieve 11% increase in performance over off-the-shelf feature descriptions in this case.

The second dataset exhibits severe occlusions and dense traffic scenarios (Dense Traffic and Occlusions). This dataset consists of 1213 query images and 596 database images. Suppressing any activations from non-relevant objects like cars in this case and highlighting the salient objects for place recognition also increases the localization performance in this case. Our method achieves best $\mathbf{F_1}$ score of 0.70, as compared to the DCNN score of 0.63. The second row of Fig. 7 shows an example from this dataset where the scene is occluded by a truck, affected by strong sun glare and the salient objects in this image are the distant buildings and a pole. Our method boosts the features from only these objects which help discriminate the place better and features from dynamic distracting objects are down weighted, which might contribute as an equal weight in the DCNN holistic representation.

The third dataset (Cluttered Environments) demonstrates

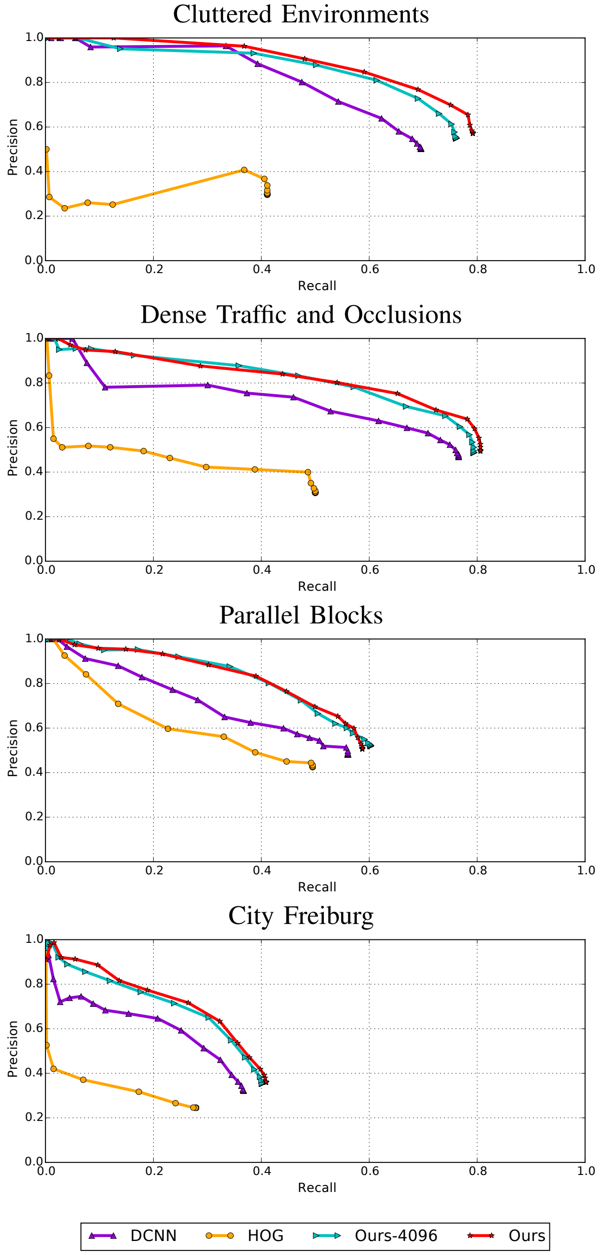


Fig. 5: *Top:* The precision recall curves exhibits the advantage of leveraging salient semantic information over all the datasets. Our proposed feature aggregation method outperforms off-the-shelf features.

Method	Parallel Blocks	Dense Traffic	Cluttered Environment	City Freiburg
HOG	0.47	0.44	0.39	0.26
DCNN	0.53	0.63	0.63	0.38
Ours	**0.59**	**0.70**	**0.73**	**0.43**
Ours-4096	0.58	0.69	0.71	0.42

TABLE I: Our method outperforms DCNN-features over all the datasets. The performance loss due to the feature projection (blue) is insignificant.

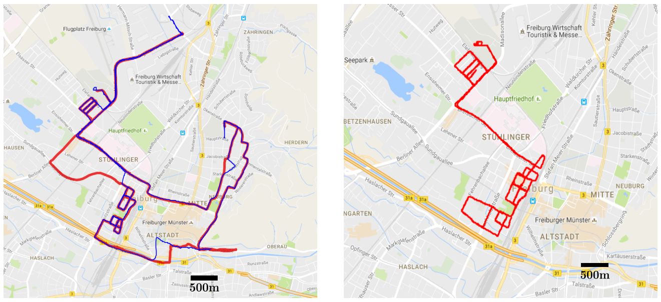


Fig. 6: *Left:* The trajectories traversed in Summer 2015, and Winter 2012 in Freiburg, Germany. *Right:* The trajectory recorded in Summer 2012.

Operation	GPU (ms)	CPU (ms)
Image Segmentation	**70.0**	6000.0
Feature Extraction	**8.0**	31.0
Feature Matching	**0.004**	1.0

TABLE II: Segmentation consumes most of the computational resources and GPU performs 120× faster than CPU on average.

how feature aggregation provides complimentary description in cluttered environments as shown in Fig. 7. The tree on the right corner is really dominant and is an important landmark in this case, however DCNN retrieves a false database image which has a tree at a similar spatial location. The feature aggregation includes features from the tree and amplifies the activations from the tower and traffic sign. In this way, we are able to cope with areas which are occluded by trees and do not contain enough man-made structure. Our method achieves 15.9% higher best $\mathbf{F_1}$ score than DCNN-features. This test set contains 781 query images and 1328 database images.

The fourth experiment shows the scalability of our approach. We evaluated it over the complete **Localization-2** dataset. The **database** and the localization datasets have a time lag of 3 years in this case, encapsulating extreme perceptual changes, which makes the dataset very challenging. The database contains 30790 images and the localization run contains 5392 images. We achieve the best $\mathbf{F_1}$ score of 0.43, which is an improvement of 13.1% over DCNN. Hence, our approach outperforms off-the-shelf feature description over large scale environments and generalizes well over different environment types and conditions. Our approach can be further combined with sequential information and region-based descriptions to achieve higher localization performance.

B. Timing Comparisons

We evaluated three main components of our approach on CPU and GPU for the timing analysis. Our approach runs at 14 Hz on the TITAN-X GPU. Image segmentation is the most computationally intense task, taking 70 ms to segment an image of 256×512 pixels. The timing for feature matching is reported for single image-pair and the timing comparison between CPU and GPU is shown in Table II.

V. Conclusion

We have proposed an approach that learns stable image regions over longer periods of time and harsh perceptual conditions. We showed that boosting features from these salient image regions significantly improves off-the-shelf feature descriptors extracted from DCNNs. We also introduced a new dataset that captures a large range of perceptual conditions in real world driving scenarios. We showed with extensive set of evaluations that our approach improves generic image descriptors and also has the potential to be combined with other approaches to achieve higher retrieval accuracy and robustness.

References

[1] D. Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[2] R. Arandjelović and A. Zisserman, "Visual vocabulary with a semantic twist," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2014.

[3] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2012.

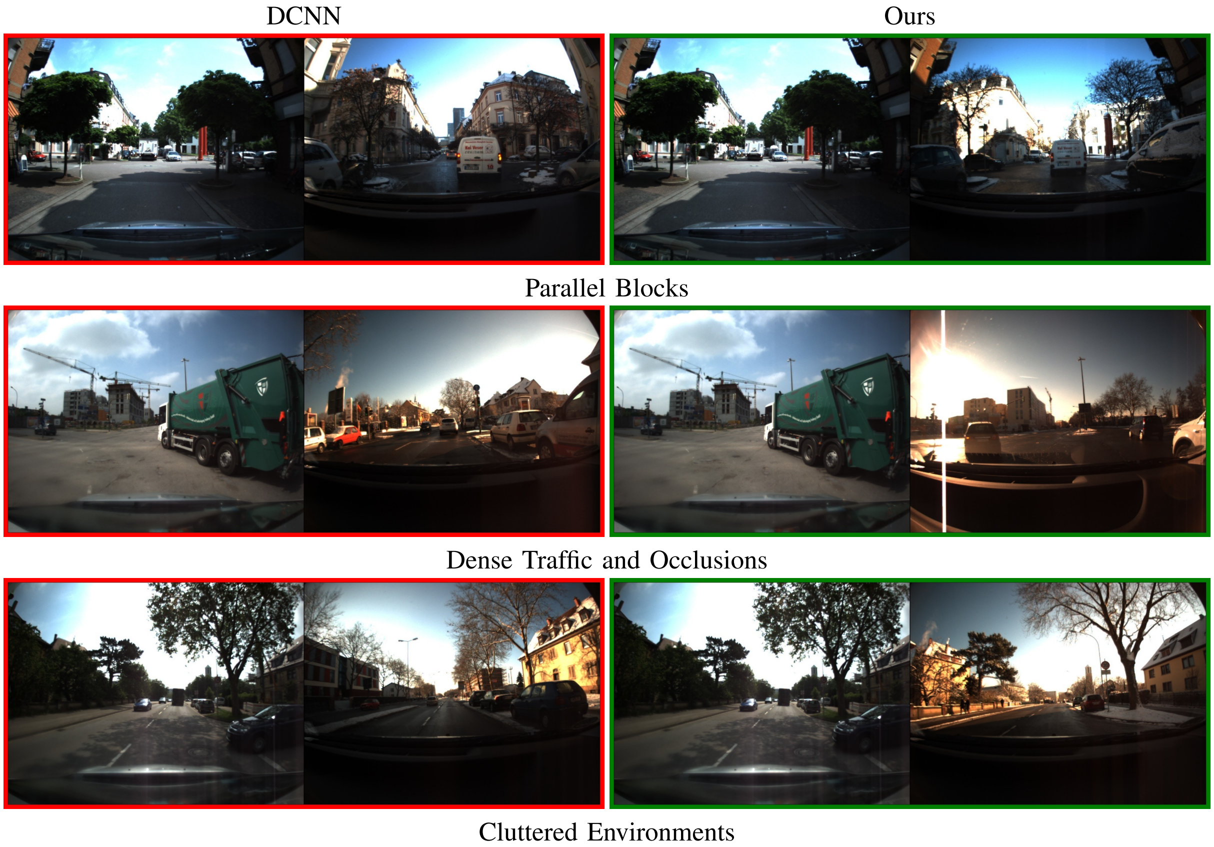
[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[6] M. Dymczyk, S. Lynen, T. Cieslewski, M. Bosse, R. Siegwart, and P. Furgale, "The gist of maps-summarizing experience for lifelong localization," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2015.

[7] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," *arXiv preprint arXiv:1605.06457*, 2016.

[8] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.



Parallel Blocks

Dense Traffic and Occlusions

Cluttered Environments

Fig. 7: The correct matches are shown in *Green* borders and the false matches are shown in *Red*. Our approach helps to diminish false matches by boosting features from stable image regions, suppressing spatial ambiguities and leveraging original DCNN activations.

[9] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 907–914.

[10] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *European Conference on Computer Vision*. Springer, 2010.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[12] C. Linegar, W. Churchill, and P. Newman, "Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2016.

[13] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv: 1506.04579*, 2015. [Online]. Available: http://arxiv.org/abs/1506.04579

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[16] S. M. Lowry, G. F. Wyeth, and M. J. Milford, "Towards training-free appearance-based localization: probabilistic models for whole-image descriptors," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2014.

[17] C. McManus, B. Upcroft, and P. Newman, "Learning place-dependant features for long-term vision-based localisation," *Robotics & Autonomous Systems*, vol. 39, no. 3, pp. 363–387, 2015.

[18] M. Milford and G. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights." in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2012.

[19] A. Mousavian, J. Košecká, and J.-M. Lien, "Semantically guided location recognition for outdoors scenes," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2015.

[20] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard, "Robust visual slam across seasons," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[21] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows." in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2014.

[22] P. Neubert and P. Protzel, "Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 484–491, 2016.

[23] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep methods for monocular road segmentation," *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2016.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Int. Conf. on Learning Representations (ICLR)*, 2015.

[26] N. Suenderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. of Robotics: Science and Systems (RSS)*, Rome, Italy, July 2015.

[27] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," *arXiv preprint arXiv:1501.04158*, 2015.

[28] N. Sünderhauf, S. Shirazi, A. Jacobson, E. Pepperell, F. Dayoub, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. of Robotics: Science and Systems (RSS)*, 2015.

[29] H. Zhang, F. Han, and H. Wang, "Robust multimodal sequence-based loop closure detection via structured sparsity," in *Proc. of Robotics: Science and Systems (RSS)*, 2016.