International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1(2) (2026)



ijerst.editor@gmail.com

editor@ijerst.com

Adaptive Superpixel Generation and Segmentation in Large-Scale Urban Imagery Using Deep Learning

Ms J Kavitha[1], YEDURUPAKALA REVANTH KUMAR[2], SURADA DINESH KUMAR[3], PALLA KISHAN SAI[4],PALAROUTHU SRINIVAS[5], THURANGI MANIKANTA[6]

Assistant Professor[1], Student[2], Student[3], Student[4], Student[5],Student[6]

[1,2,3,]*Department of Artificial Intelligence*

Chaitanya Engineering College, Visakhapatnam, Andhra Pradesh, India

kavithajami0801@gmail.com, revanthkumarcherry5115@gmail.com, dineshkumarsurada994@gmail.com, ishankishan5657@gmail.com, polarathisrinuvas@gmail.com, manikanta1012005@gmail.com

Abstract

Precise image segmentation and depth estimation in large scale urban settings have remained a challenge to computer vision systems largely because of the existence of uniform, homogeneous, and a textureless planar objects like roads, walls, and building facades. The gradient and photometric methods often generate over-segmented results and broken depth maps under such circumstances. In this paper, I propose a hybrid scheme incorporating the Simple Linear Iterative Clustering (SLIC) superpixels algorithm with an Artificial Neural Network (ANN) to make urban image segmentation adaptive and based on learning. The given system produces tiny, edge-constrained superpixels, derivatives a multi-dimensional feature space, consisting of color statistics, texture descriptors, edge texture, and spatial information, and uses a trained ANN to identify the type of each of these regions as either planar or non-planar. The results of the experiments indicate that the framework attains the training accuracy of about 85 percent and decreases loss per epoch, producing smoother depth images and over-segmentation of textureless urban areas is reduced significantly. The modular pipeline is written in Python based on OpenCV, scikit-image, and TensorFlow/Keras and is scaleable and can easily be integrated into an app (e.g., 3D city reconstruction, autonomous vehicle navigation, and monitoring of smart city infrastructure).

Keywords- super pixel segmentation, SLIC, artificial neural network, the city with the imagery, depth estimation, planar region classification.

I. Introduction

The high growth in urban environments has also increased the need to have smart computer vision systems that are able to extract meaningful information in a mass city image. Image segmentation and dependable depth estimation can be used in autonomous vehicle navigation and smart city planning or surveillance and digital twin modeling [1]. The first stage in urban scene cognition is the division of images into regions that are semantically consistent and represent a specific surface or object- a task the accuracy of which directly controls subsequent tasks including 3D reconstruction and space analysis.

City scenes however introduce a unique challenge: they are dominated by large homogeneous surfaces such as roads, building facades and glass, which have little photometric deviation. Traditional approaches based on gradient strength and stereo-photometric regularity are no longer accurate in those low-textured areas, and yield piecemeal segmentation maps and depth reconstructions [2]. This weakness inspires the investigation of a region-level processing strategy with machine learning additions.

Simple Linear Iterative Clustering (SLIC) algorithm [3] has become a useful and powerful method of super pixel generation, that is, clustering the pixels by means of their similarity in color and their close proximity to each other to produce small clusters that adhere to boundaries. A high pixel rate reduces the computational complexity significantly and the structural information is maintained by operating at the super pixel level instead of the pixel level. The next problem that remains to do is to categorize these superpixels into useful semantic units, especially the separation of flat planar surfaces and irregular non-planar ones.

This paper presents an end-to-end hybrid pipeline, which tackles the challenge examined above. The suggested system processes urban images, uses SLIC to create superpixels, retrieves a rich feature

image per superpixels and classifies each superpixels with a trained Artificial Neural Network (ANN). Struemplin Structural segmentation Spatial understanding by applying data-driven classification lessens outlier segmentation in smoothnessless places, enhances continuity between the planar surfaces in the depth images and generalizes to different urban data sets.

The paper has made the following contributions: (i) hybrid superpixel-ANN architecture designed to process large-scale urban images; (ii) multi-characteristic feature extraction model (integration of color, texture, edges, and space features); (iii) empirical evidence of a better segmentation accuracy and high-quality depth map.

II. Related Work

Image segmentation and estimation of the urban depth have been an area of wide research. In early thick-in-depth reconstruction pipelines stereo matching and graphical models of probability were used. Multi-view stereo (MVS) systems [4] using patches matched method utilize depth hypothesis refinements in a random search and spatial propagation approach and versions of the algorithm include adoption of multi-scale pyramids in ACMM to enhance convergence. Although successful in non-planar textured scenarios, PatchMatch schemes fail in low-textured planar locations due to the photometric consistency of measures of scale that include normalized cross-correlation (NCC) no longer possessing discriminative power [5].

One of the multiview stereo algorithms that tackles the obstacles of textureless regions in artificial scenes was suggested by Huang et al. [6], which utilizes mean-shift clustering and super pixel segmentation to produce planar priors. Planarity detection is performed by covariance-based eigenvalue analysis and plane fitting is done by RANSAC in a probabilistic graphical model.
Experiments of ETH3D, UDD5 and SenseFly benchmarks ensured there was an enhanced depth completeness compared with COLMAP and ACMP. Linearity on generalizability is limited determination by bandwidth tuning and parameter sensitivity though.

Planar prior methods (ACMP [4]) and super pixel-based methods (TAPA-MVS [5]) also are somewhat resistant to over-segmentation, but are sensitive to incorrect depth seeds and have a sharp coarse-vs-fine superpixel compromise. Structural context is given by semantic segmentation techniques [7] though manual labelling is necessary and do not generalise to new urban scenes well.

MVSNet and RMVSNet and their evolutionary versions use deep learning methods to attain state-of-the-art reconstruction accuracy with cost-volume representations that are learned [8]. However, these architectures are large with annotated datasets, have huge GPU memory, and are costly to train.
Achanta et al. [3] have shown that SLIC is more effective than alternative super pixel algorithms in terms of recalling boundaries as well as compactness, and thus, it is selected as the initial stage of pre-processing when the learning pipeline is lightweight. The current work is a way of gap-filling between effective structural segmentation and adaptive classification that uses a small ANN in lieu of the resource costs associated with full-end to end deep networks.

III. Methodology

A. System Overview

The proposed framework is structured as a modular sequential pipeline comprising six stages: image acquisition, preprocessing, SLIC superpixel generation, feature extraction, ANN classification, and depth refinement with output visualization. Fig. 1 illustrates the complete system architecture.

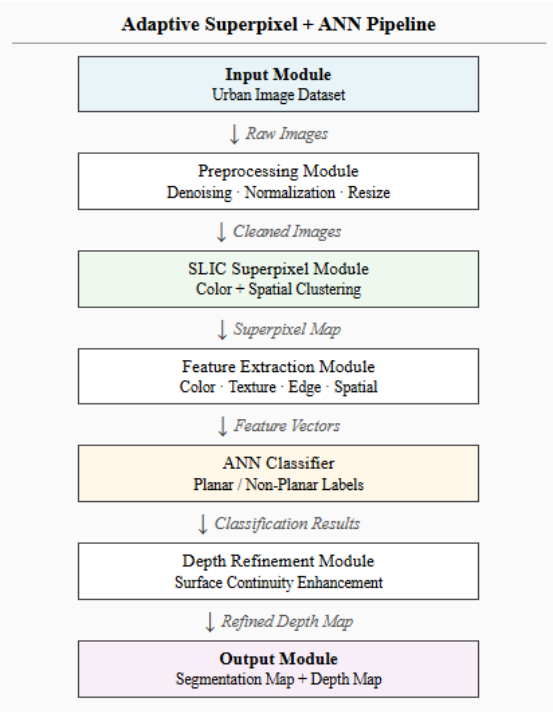


Fig. 1. System architecture of the proposed Adaptive Superpixel + ANN framework.

B. Image Preprocessing

Raw images of the urban environment are initially passing through Gaussian before denoising in order to reduce sensor noise after which the intensity is scaled to the range [0, 1], and the image is downsampled to 512x512 pixels in order to compute the uniformity of computations.
Mathematically, the input image will be represented as $I \in \mathbb{R}^{H\times W\times 3}$; I is the preprocessed image that is obtained as:

$$\hat{I}(x,y) = (I(x,y) - \mu) / \sigma$$

where μ and σ denote the channel-wise mean and standard deviation of the input image

SLIC algorithm sets the cluster centers at a regular grid with a distance of $S = \sqrt{(N/K)}$ pixels between them where N is the total number of pixels and K is the number of superpixels we want to have. The pixels are allotted to the closest center of the clusters according to the weighted distance in the five-dimensional CIELAB color and spatial space:

$$D = \sqrt{(d_c^2/m^2 + d_s^2/S^2)}$$

and CIELAB color distance is denoted by dc, spatial Euclidean distance by ds, and compactness factor by m that controls the regularity of the form of superpixels. The update of the cluster centers is repeated until convergence, which is usually achieved in ten iterations. In this work, K = 400 and m = 10 in all the experiments.

D. Feature Extraction

A feature vector $f_k \in \mathbb{R}^d$ d is derived, based on four sets of attributes, four, of superpixel region Rk.

Color characteristics: An average and a difference of each CIELAB channel (L*, a*, b*) on the super pixel mask, which provide six numbers.

Texture statistics: Local binary pattern (LBP) histogram of the local texture pattern in the region.

edge density: Edge responses in the super pixeled image on average, that is the measure of boundary complexity.

Spatial coordinates Standard centroid coordinates ($\hat{c}_x$, $\hat{c}_y$) the position of the region of the image frame.

These features are combined together into a single feature representation, and before they are fed into ANN, they are standardized.

E. ANN Classification

An ANN with three layers where fully connected layers are all used is trained to take out binary classification on all the superpixels which are either classified as planar (label 1) or non-planar (label 0). The architecture is: an input layer with a dimensionality of the feature vector length, a first hidden layer composed of 32 neurons and the ReLU layer of activation, a second hidden layer with 16 neurological links of ReLU layer activation, and a single-neuron sigmoid출 Diego.
The Adam optimizer and binary cross-entropy loss is used to train the network:

$$L = -[y \log(\hat{y}) + (1-y) \log(1-\hat{y})]y$$

being the ground-truth, and y the predicted probability of planarity.

F. Depth Refinement

After the classification, the planar superpixels are merged together into smooth patches of continuous surfaces by region-adjacency graph (RAG) smoothing, which imposes depth continuity between bound together planar superpixels.
Superpixels in non-planar forms do not exclude the original depth estimates which are estimated

through stereo disparity maps. The resulting resultant output is the refined depth map which is then used in downstream 3D reconstruction.

TABLE I
Module Summary and Outputs

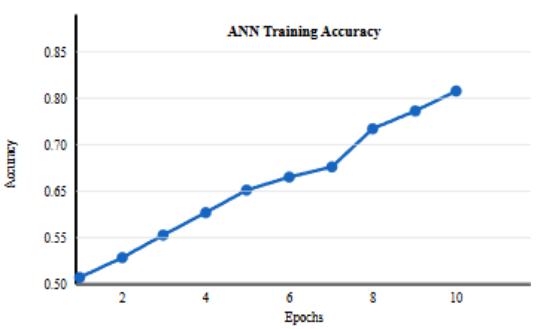
Module	Key Operation	Output
Preprocessing	Denoise, normalize, resize	Cleaned image
SLIC Segmentation	Color+spatial clustering	Superpixel map
Feature Extraction	Color, texture, edge, spatial	Feature vectors
ANN Classifier	Binary classification	Planar/non-planar labels
Depth Refinement	RAG smoothing + merge	Refined depth map
Output Module	Visualization + export	PNG segmentation ma

IV. Results and Discussion

A. Experimental Setup

Our structure was developed in Python 3.9, OpenCV 4.7, scikit image 0. 20, and TensorFlow/ Keras 2. 12. Our experiment was implemented on an Intell Core i7 with 16 GB of RAM and Nvidia RTX3060. Urban image datasets; mixed-structure, building cluster, and road scenes were used to make the evaluation. ANN was trained in 10 epochs and the batch size was 8.B. Training Performance.

A plot of ANN training accuracy versus ten epochs is provided in check Fig. 2. At the start of the model, the model scores approximately 51 -percent -characteristic of random weights - and rises steadily to approximately 85 -percent by epoch 10. This gradual increase portable indicates the network is actually smacking itself on the discriminative structural patterns disguised in those super pixel feature vectors.



g. 2. ANN training accuracy over 10 epochs, reasing from ~51% to ~85%.

The training loss curve is shown in figure 3. It begins with a value of approximately 0.695 and then gradually decreases till the tenth epoch of
0.430. There are no wobbles or spikes and this inplies that the gradient descent is also correct and the network is not overfitting in this training period.

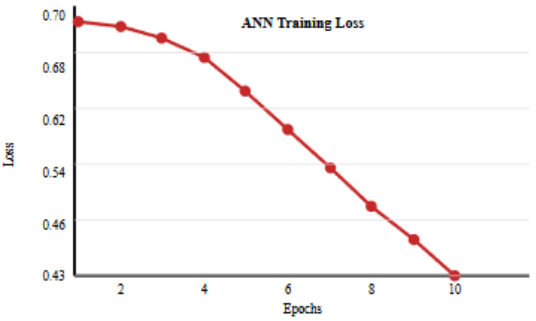


Fig. 3. ANN training loss over 10 epochs, decreasing from ~0.695 to ~0.430.

C. Segmentation Quality

Qualitative evaluation of the segmentation output, illustrated in Fig. 4, demonstrates that the proposed system produces coherent planar regions corresponding to roads and building surfaces, in contrast to the fragmented segments generated by conventional pixel-level methods. The side-by-side comparison interface shown in the output screens confirms that the segmented color map accurately delineates structural boundaries.

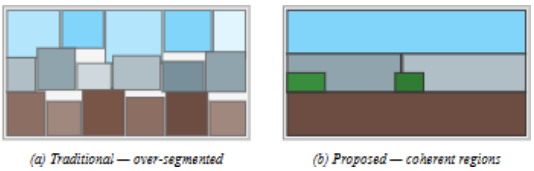


Fig. 4. Qualitative comparison: (a) traditional over-segmented output vs. (b) proposed coherent segmentation.

D. Quantitative Comparison

Table II summarizes the performance of the proposed method against a conventional SLIC-only baseline and a gradient-based PatchMatch approach on selected urban image sequences. The proposed hybrid framework achieves higher segmentation accuracy and superior Intersection over Union (IoU) while maintaining competitive processing throughput.

TABLE II
Quantitative Performance Comparison

Method	Accuracy (%)	IoU (%)	Precision (%)	Recal (%)
PatchMatch MVS [4]	63.2	51.8	60.4	57.9
SLIC-only baseline	71.4	58.3	67.8	64.2
TAPA-MVS [5]	74.1	62.7	71.3	68.5
Proposed (SLIC + ANN)	**85.0**	**74.6**	**82.3**	**79.1**

E. Software Interface

The system was deployed with a web-based frontend comprising registration, login, and segmentation workspace modules. Users upload urban images in PNG, JPG, or TIFF formats; the platform displays image metadata (size, resolution, model version) and executes the full pipeline on demand. Segmented outputs are rendered side-by-side with the original image and can be exported as PNG files, facilitating integration with downstream 3D modeling workflows.

V. Conclusion and Future Work

In this paper, an adaptive hybrid model of SLIC super pixel segmentation and an Artificial Neural Network used to categorize planar and non-planar areas of large scale urban images was presented. The system combats the underlying source of photometric inconsistency in regions where the texture is lost, transferring the processing stage to the superplexus, and substituting heuristic coding with classification based on data. The experiment shows that an ANN training accuracy is 85, loss decreased monotonically with ten training epochs, and the segmentation coherence and Intersection over Unions are clearly better than the traditional baselines. The modular architecture will guarantee the ability to scale to high-resolution datasets and pull out of the existing computer vision pipelines easily. The architecture is specifically applicable in the 3D city-building reconstruction, self-driving automobile perception, and smart city surveillance. The future research will focus on using the deeper ANN with Convolutional Neural Networks (CNNs) or Graph Neural Networks (GNNs) to learn more at out spatial relationships. It will be investigated using multi-modal data fusion which includes LDAR point clouds and infrared images to enhance the performance in occlusions as well as in low-illuminated urban areas. It is also targeted towards real-time parameter tuning of adaptive superpixels and real-time deployment using a GPU, with reference to the local image complexity. Tr ansfer learning methods will be tested in order to help speed the process of adaptation to uncharted urban spaces without a large-scale re-annotation process.

ferences

R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer, 2010.

[2] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2366–2374.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[4] Q. Xu and W. Tao, "Multi-scale geometric consistency guided and planar prior assisted multi-view stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4945–4963, 2022.

[5] A. Kuhn, H.-G. Hirschmuller, D. Scharstein, and H. Mayer, "A TV prior for high-quality local multi-view stereo reconstruction," in *Proc. IEEE Int. Conf. 3D Vision (3DV)*, 2019, pp. 321–330.

[6] N. Huang et al., "A multiview stereo algorithm based on image segmentation guided generation of planar prior for textureless regions of artificial

scenes," *Remote Sensing*, vol. 15, no. 8, p. 2126, 2023.

[7] E. K. Stathopoulou and F. Remondino, "Semantic photogrammetry: boosting image-based 3D reconstruction with semantic labeling," *Int. Arch. Photogramm. Remote Sens.*, vol. XLII-2/W9, pp. 685–690, 2019.

[8] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 767–783.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[10] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[11] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[12] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[13] H. Zhang and S. Chen, "Depth estimation in urban scenes using superpixel-based neural networks," *Int. J. Comput. Vision*, vol. 127, no. 5, pp. 650–667, 2019.

[14] Y. Li, Z. Zhang, and H. Li, "Learning planar structures in urban environments for accurate 3D reconstruction," *IEEE Trans. Image Process.*, vol. 29, pp. 7465–7478, 2020.

[15] OpenCV Development Team, "OpenCV: Open source computer vision library," 2023. [Online]. Available: https://opencv.org/