IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 6, 2019, accepted July 5, 2019, date of publication July 15, 2019, date of current version October 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928882

Dense Refinement Residual Network for Road Extraction From Aerial Imagery Data

KARUNA KUMARI EERAPU[1], BALRAJ ASHWATH[1], SHYAM LAL[1], (Senior Member, IEEE), FABIO DELL'ACQUA[2], (Senior Member, IEEE), AND A. V. NARASIMHA DHAN[1]

[1]Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal 575025, India
[2]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

Corresponding authors: Karuna Kumari Eerapu (karuna.eerapu5023@gmail.com), Balraj Ashwath (balrajashwath98@gmail.com), Shyam Lal (shyamfec@nitk.edu.in), Fabio Dell'Acqua (fabio.dellacqua@unipv.it), and A. V. Narasimha Dhan (dhan257@gmail.com)

This work was supported in part by the Young Faculty Research Fellowship Award under the Visvesvaraya Ph.D. Scheme for Electronics and IT Government of India, in part by the Digital India Corporation (formerly Media Lab Asia), and in part by the Ministry of Communications and Information Technology (MCIT), Government of India, under Grant DIC/MUM/GA/10(37)D.

ABSTRACT Extraction of roads from high-resolution aerial images with a high degree of accuracy is a prerequisite in various applications. In aerial images, road pixels and background pixels are generally in the ratio of ones-to-tens, which implies a class imbalance problem. Existing semantic segmentation architectures generally do well in road-dominated cases but fail in background-dominated scenarios. This paper proposes a dense refinement residual network (DRR Net) for semantic segmentation of aerial imagery data. The proposed semantic segmentation architecture is composed of multiple DRR modules for the extraction of diversified roads alleviating the class imbalance problem. Each module of the proposed architecture utilizes dense convolutions at various scales only in the encoder for feature learning. Residual connections in each module of the proposed architecture provide the guided learning path by propagating the combined features to subsequent DRR modules. Segmentation maps undergo various levels of refinement based on the number of DRR modules utilized in the architecture. To emphasize more on small object instances, the proposed architecture has been trained with a composite loss function. The qualitative and quantitative results are reported by utilizing the Massachusetts roads dataset. The experimental results report that the proposed architecture provides better results as compared to other recent architectures.

INDEX TERMS Dense convolutions, dense blocks, DRR Net, IOU, loss function, residual connections.

I. INTRODUCTION

The topographical map of any geographical location can be built by capturing high-resolution aerial images using Aircraft, Helicopters, Unmanned aerial vehicle (UAVs) , etc. Information about presence and location of topographical features such as roads, dams, buildings, bare land, etc., is essential for applications like urban planning, disaster assessment, traffic management, and map updating. This information is usually collected by extracting the objects of interest (topographical features) from aerial images. Among all objects, road information is primary in many applications. Thus, segmentation of roads serves as a basis to update maps for global positioning system (GPS) -based navigation devices and also for the majority of the above-mentioned applications. In high-resolution aerial images, roads do not possess a continuous regular shape and they frequently appear very narrow as they take small number of pixels across. In order to get higher levels of accuracy, all kinds of diversified roads have to be extracted while preserving connectivity in dense and scattered environments. Hence, reliable road extraction from aerial imagery data is a challenging problem in the field of computer vision.

Amo *et al.* extracted road pixels by initially using a region growing technique and then refining the results are by applying region competition techniques. The major limitation of the introduced method was, the requirement of user seed selection for region growing technique [1]. Hu *et al.* presented an automated method based on Bayes decision rule to distinguish road pixels and to track road networks. The main drawback of this work is over-segmentation of roads in the process of classification of road pixels [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Yungang Zhu.

Sahar *et al.* proposed a technique to achieve correct segmentation of road regions using extended Kalman filter and Particle filter. It is mentioned that the proposed technique has to be fine-tuned before it can segment roads in complex situations [3]. Jiangye *et al.* attempted to circumvent the need for post-processing by extracting the roads in three stages using the so-called locally excitatory globally inhibitory oscillator networks (LEGION) . The major pitfall of this approach is the determination of optimal parameters in road segmentation and grouping stage for various types of images [4]. Das *et al.* calculated the spectral contrast and linear trajectory features by training support vector machines. From these calculated features, road regions were segmented without the need for parameter tuning. This process could extract roads of greater width. However, narrow roads covered by shadows were extracted improperly [5]. Cem Unsalan *et al.* attempted to extract roads in all kinds of environments using graph-based approaches in a probabilistic way. The disadvantage of this method is that it extracted roads only from images of predefined spatial resolution [6].

In [1]–[6] roads were extracted in more than two stages by calculating road features in an unsupervised way. Due to the computation of features from the smaller context of training data and also due to dependency on previous stage outputs, the final predictions would not lead to satisfactory results. However, techniques based on deep Convolutional Neural Networks (CNNs) extract the objects in a supervised way by considering the larger context of input data. The major benefit of deep CNNs is their ability to calculate features by learning from the high volume of input data. The Deep-CNN-based approaches extract the objects through semantic segmentation with the aim of perceiving *what is in the image and where it is located*.

Badrinarayan *et al.* introduced an encoder-decoder based architecture for semantic segmentation. The max-pooled encoder feature maps are transferred to the decoder through pooling indices. However, by considering only maximum values of encoder feature maps, there might be a possibility of losing fine details associated with small objects. This may result in inefficient segmentation of small objects especially in the case of high-resolution images [7]. Ronneberger *et al.* introduced skip connections as an alternative to pooling indices for transferring the learned features to the corresponding resolution level of the decoder. This results into a higher number of feature maps; hence, the complexity of the decoder increases [8]. The architectures in [7], [8] share a common point of using convolutional filters for feature learning and pooling layers to exploit semantics. The use of pooling layers reduces the spatial resolution of feature maps. Preserving spatial resolution is important for retaining fine details of objects. Fisher Yu *et al.* introduced another type of convolution called as dilated/atrous convolutions in order to preserve the spatial resolution while avoiding pooling [9]. Chen *et al.* introduced a semantic segmentation architecture by utilizing a distinct dilation filters in spatial pyramid pooling (SPP) [10] for aggregation of multi-scale context. The resulting architecture produced a segmentation map with one-eighth input resolution [11]. Chen *et al.* placed an additional decoder to maintain the spatial resolution of above mentioned architecture [12]. Yang *et al.* introduced a semantic segmentation architecture by providing parallel and cascade connections among various dilation filters [13]. However, it is observed that the obtained receptive field due to the usage of dilation filters in [9]–[13] is not sufficient to preserve the spatial connectivity of roads during extraction.

Huang *et al.* introduced the idea of dense convolutions that iteratively reuse the learned features at later resolutions [14]. Jegou *et al.* extended the concept of dense convolutions to semantic segmentation by utilizing them in the paths of encoder and decoder. However, due to the usage of dense convolutions together with skip connections in the up-sampling path, the model demands more memory during training [15]. Pohlen *et al.* and Samy *et al.* exploited the benefits of operating at full resolution by processing up-sampling and down-sampling streams concurrently. This increases both localization and classification accuracy. However, the introduced techniques are computationally intensive as they operate at full resolution [16], [17]. Zhang *et al.* proposed a model named as ResUNet utilizing residual connections in U-Net. The resulting model failed however to segment small roads in parking lots [18]. Filin *et al.* attempted to refine the predictions of ResUNet model by further processing the road pixels in order to fill the gaps in between them [19]. Tao Sun *et al.* introduced a model for generation of road maps by stacking two U-Nets. The introduced model needs further post-processing operations to extract road center lines and to connect disjoint roads [20]. Kim *et al.* placed SPP at the end of the encoder of U-Net to aggregate multi-scale contextual information. The major limitation is the increased depth of feature maps due to the usage of greater number of filters. This lead to increased computational complexity [21]. Aich *et al.* introduced a technique called Depth to Space (D2S) to reduce the computational complexity by excluding the decoder. This is however not well suited for segmentation of small objects as there is no learning path for up-sampling [22].

In this work we propose an efficient architecture that is inspired by the effectiveness of dense convolutions for feature learning [14], [15] and residuals to achieve progress in learning ability of network [23], [24] at full resolution. The main contributions of the paper are as follows:

(i) A novel semantic segmentation architecture is proposed based on dense convolutions and residual connections. The proposed architecture operates at full resolution and is composed of multiple DRR modules.

(ii) Each module of the proposed architecture learns features at different resolutions to extract affluent semantics and also endeavors to obtain predictions.

(iii) The modules are constrained to refine the predictions by stacking them.

The organization of this research paper is as follows: The detailed explanation of proposed architecture along with its internal modules is given in Section II. The description of the dataset used for training of all models, including the particulars of hyperparameters utilized is presented under Section III. An elaborate discussion about simulation results of all architectures are described in Section IV. Finally Section V concludes this work.

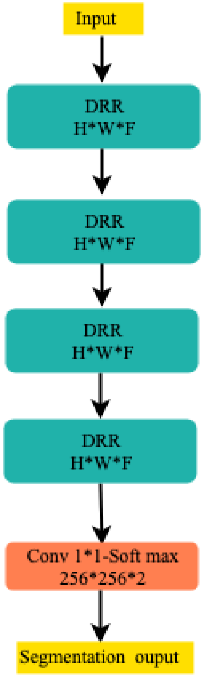


FIGURE 1. The proposed architecture for semantic segmentation of aerial imagery data.

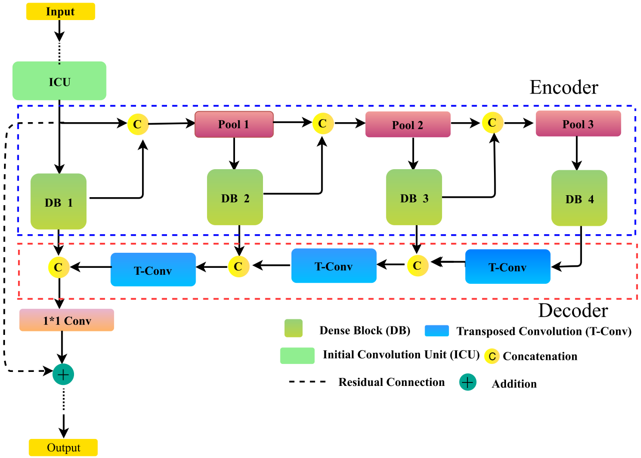


FIGURE 2. Dense Refinement Residual (DRR) module.

II. PROPOSED ARCHITECTURE

The proposed dense refinement residual network for semantic segmentation of aerial images is presented in Fig. 1. The DRR Net is primarily composed of dense refinement residual (DRR) module(s), and the structure of DRR module(s) is presented in Fig. 2. In the proposed DRR architecture, each DRR module inherently contains down-sampling (encoder) and up-sampling (decoder) paths. In the encoder of the DRR module, features are extracted at different resolutions by utilizing dense convolutions. Similarly, in the decoder transposed convolutions are used at multiple scales to learn the up-sampling of feature maps together with learned features of the encoder. Residual connections are employed in each DRR module to provide a supervisory signal to successive DRR modules. This supervisory signal is formed by adding initial features on which each DRR module operates and its corresponding learned features. The successive DRR modules of the proposed architecture attempt to improve the predictions of antecedent(s) by operating on their features. The architecture effectively reuses the features through dense, residual connections and also by stacking of individual DRR modules. This leads to an increase in longevity of feature propagation. The resolution at which each DRR module of proposed architecture operates is given by $H * W * F$[1]. The final stage utilizes 1*1 convolutions as softmax in order to produce individual class probabilities. The detailed functionality of DRR module utilized in the proposed architecture is described in the following section.

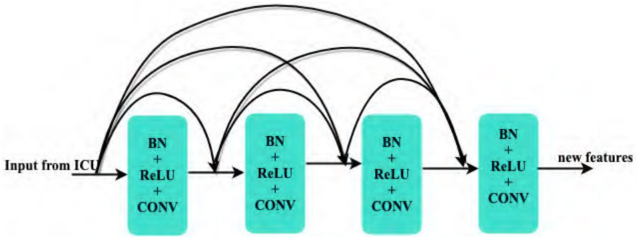


FIGURE 3. Dense Block (DB) .

A. DENSE REFINEMENT RESIDUAL (DRR) MODULE

The dense refinement residual module of the proposed architecture extracts and up-samples the fine-grained features from the input data. The initial convolution unit (ICU) of first DRR module attempts to learn the initial features from input by applying a sequence of normal convolutions. In the later DRR modules, ICU learns the intermediate feature maps from its preceding DRR module. The structural diagram of dense blocks (DBs) employed in DRR module(s) is represented in Fig. 3. This structural module of DBs function on the features of ICU. In each layer of DB, batch normalization (BN) [25], rectified linear unit (ReLU) and convolution (CONV) operations are performed by taking all the possible direct connections from its preceding layers. The number of such layers (L) used is 4 with each layer having a growth rate (The number of convolutional filters used k) of 16. Moderate values are chosen for k and L to ensure a sufficient amount of information is added to the next layer in each DB and also to the successive DBs. Further, these values also

[1] H, W are height and width of the input image respectively and F represents the number of filters used in initial convolution unit of DRR module

help to maintain a constant depth dimension across all DRR module(s). The learned features of DBs are then passed to successive DBs after pooling. After each level of learning at dense blocks, the feature maps are concatenated with preceding learned features and are also spatially reduced by a factor of two. DB1 predominantly focuses on initial features, and its output is linked with them. The resulting feature maps are max pooled before feeding them to the successive dense blocks. DB2 attempts to learn a different set of feature maps based on DB1 output and initial features. Features extracted out of DB3 are based on the cumulative knowledge of the outputs of DB2, DB1 and initial features. Finally, DB4 extracts high-level features by making use of the collective knowledge accumulated by DRR module up to that point. Feature maps at this level are down-sampled by a factor of eight. From Fig. 2 it can be seen that the up-sampling process begins at the higher level features of DB4. The up-sampling of feature maps for remaining resolutions is achieved by considering feature maps of preceding dense blocks and learned features of corresponding dense blocks. Residual connections in DRR modules provide a deep supervision to subsequent modules by transferring the combined initial and learned features. Thus, the strength of feature propagation increases due to the effective utilization of feature maps in the encoder and decoder of DRR module.

To summarize, the highlights of the proposed architecture are given as follows:

(1) Each DRR module of the proposed architecture learns diverse features at various scales with the help of dense blocks.

(2) In DRR module, dense blocks at consecutive pool path learn new features based on collective knowledge accumulated by the network.

(3) In an up-sampling path of DRR module, transposed convolutions are used instead of dense blocks which results in a great reduction in the number of parameters without comprising the prediction accuracy.

(4) Predictions are refined by stacking multi-scale context successively at full resolution.

(5) The proposed DRR Net provides a guided learning path to successive DRR modules with establishment of residual connections in each module.

(6) The depth of feature maps remain constant, though multiple DRR modules are appended sequentially. Thus avoiding feature map explosion.

(7) The proposed architecture provides competitive results with a tenfold reduction in the number of parameters as compared to other existing semantic segmentation architectures.

(8) The proposed architecture provides increased flexibility to append or efface number of DRR modules based on computational budget and accuracy.

B. REFINEMENT STAGE

Let N denote the number of DRR modules and X_i denote the initial features extracted from initial convolution unit (ICU). Let X_{ij} represent the features of dense blocks at different resolutions at i^{th} DRR module and j^{th} pool respectively. Similarly, Y_i represent the predictions or segmentation maps of i^{th} DRR module, where $i \in [1, N]$ and $j \in [0, 3]$

Let X'_{ij} denote the up sampled features learned at different resolutions, where $i \in [1, N]$ and $j \in [1, 3]$

Thus, X_{11}, X_{12}, X_{13} are the features learned at Pool 1, Pool 2 and Pool 3 respectively in first DRR module.

$X'_{11}, X'_{12}, X'_{13}$ are the features up sampled at Pool 1, Pool 2 and Pool 3 respectively in first DRR module.

Learned features from DB 1, DB 2, DB 3 and DB 4 can be given as

$$\left.\begin{aligned} X_{10} &= H\{X_i\} \\ X_{11} &= H\{X_{10}, X_i\} \\ X_{12} &= H\{X_{11}, X_{10}, X_i\} \\ X_{13} &= H\{X_{12}, X_{11}, X_{10}, X_i\} \end{aligned}\right\} \tag{1}$$

Here, H represent batch normalization, ReLU and convolution operations performed in the layers of dense blocks at different scales.

Further, up-sampled feature maps at Pool 3, Pool 2, and Pool 1 respectively are given as

$$\left.\begin{aligned} X'_{13} &= F\{X_{13}\} \\ X'_{12} &= F\{X'_{13}, X_{12}\} \\ X'_{11} &= F\{X'_{12}, X_{11}\} \end{aligned}\right\} \tag{2}$$

Here, F represents transposed convolution operation for up-sampling of feature maps.

The output from first DRR module is given as

$$Y_1 = F'\{X'_{11}, X_{10}\} + Xi \tag{3}$$

Here F' define the non-linearity applied due to 1*1 convolutional filters. In the same way if multiple DRR modules (consider number of modules (N) as 4) are connected consecutively its corresponding outputs are given as

$$\left.\begin{aligned} Y_2 &= F'\{X'_{21}, X_{20}\} + Y_1 \\ Y_3 &= F'\{X'_{31}, X_{30}\} + Y_2 \\ Y_4 &= F'\{X'_{41}, X_{40}\} + Y_3 \end{aligned}\right\} \tag{4}$$

Finally, substituting Y_3, Y_2, Y_1 values recursively, Y_4 can be written as

$$\left.\begin{aligned} Y_4 = F'\{X'_{41}, X_{40}\} + F'\{X'_{31}, X_{30}\} + \\ F'\{X'_{21}, X_{20}\} + F'\{X'_{11}, X_{10}\} + Xi \end{aligned}\right\} \tag{5}$$

From equation (5), the successive DRR module operates on output of previous DRR module(s) and also on initial features at which it operated. It can be concluded that the learned features are effectively reused in the path of encoder, decoder, and also at various modules.

1) Without Residual

If multiple DRR modules (N = 4) are connected consecutively without residual connections, the corresponding outputs are given as.

$$\left.\begin{aligned} Y_1 &= F'\{X'_{11}, X_{10}\} \\ Y_2 &= F'\{X'_{21}, X_{20}\} \\ Y_3 &= F'\{X'_{31}, X_{30}\} \\ Y_4 &= F'\{X'_{41}, X_{40}\} \end{aligned}\right\} \tag{6}$$

From equation (6), it can be observed that the successive DRR module has no information on the initial features on which the previous DRR module has been operated (i.e Y_{i+1} does not depend on Y_i, $1 \leq i \leq N-1$)

III. TRAINING AND IMPLEMENTATION

The proposed DRR architecture has been trained and evaluated by utilizing the Massachusetts roads dataset published in [26]. Each image is composed of 1500*1500 pixels covering an area of 500 square km at a resolution of 1.2 m/pixel.

A. IMAGE DATASET

In this work, we consider aerial images that contain less than 50 per cent of white noise. Each resulting image is divided into thirty-six patches of size 256*256 pixels by padding with zeros instead of taking random crops. Thus, we generated 49,680 training, 1008 validation and 3528 test images including masks. The dataset was enlarged by applying horizontal, vertical flips and also brightness variations of different degrees at the time of training. The proposed DRR Net and state-of-the-art architectures were trained using TensorFlow [27] as a deep learning framework with an NVIDIA Tesla k80 GPU with 11GB on-board memory. The initial learning rate was set to 0.0002 and decayed exponentially by a factor of 0.994. The weights of convolution filters were initialized with Xavier initialization [28]. The optimal weights of filters are calculated during backpropagation by using the Adam optimizer [29]. The optimizer has an exponential decay rate value of 0.99 for first-order momentum ($\beta 1$) and 0.999 for second-order momentum ($\beta 2$) respectively. All models are trained for 24,8400 number of iterations with a batch size of 2. The inference of all trained models is performed using an Intel central processing unit (CPU).[2]

B. COMPOSITE LOSS FUNCTION

A binary cross entropy loss function (BCE) calculates the loss based on prediction probabilities of each pixel. The BCE loss value is high for false predictions and low for true predictions. Since the data set is highly skewed (it contains ~96% background pixels and ~4 % road pixels) the model bias towards background pixels frequently results into higher loss values. Hence, during the training phase, the semantic segmentation architectures take a long time to learn and also to converge. The Jaccard index or Intersection over Union (IOU) for semantic segmentation is evaluated by considering the overlap of pixels between the predicted image and its mask. This reduces the bias towards the most frequent classes and it is also a useful metric for evaluating the performance of semantic segmentation. The Lovasz softmax loss (LZS) is proposed in [30] as a mean to optimize the mean Intersection over Union by considering a collection of pixel predictions. The combination of binary cross entropy and Lovasz softmax loss is utilized in experiments to improve the pixel-wise classification accuracy of intended objects.

$$L_{composite} = L_{BCE} + L_{LZS} \tag{7}$$

where L_{BCE} is binary cross entropy loss and L_{LZS} is Lovasz softmax loss. Following the definition of cross entropy mathematical expression for L_{BCE} is written as

$$L_{BCE} = \frac{-1}{N}\sum_{i=1}^{N}[Y_i \cdot log(p(\widetilde{Y_i})) + (1-Y_i) \cdot log(1-p(\widetilde{Y_i}))] \tag{8}$$

Here, Y_i represents the actual class label values, $p(\widetilde{Y_i})$ denotes the predicted class probabilities after applying the softmax layer, and N denotes the total number of training samples in the dataset. Following [30], the L_{LZS} is given by

$$L_{LZS} = \frac{1}{|C|}\sum_{c\in C} \triangle \overline{J} c E(c) \tag{9}$$

Here $\triangle \overline{J} c$ is the loss surrogate to the Jaccard index of class c, $E(c)$ is the vector of errors $[0, 1]^p$ and $|C|$represents the number of classes.

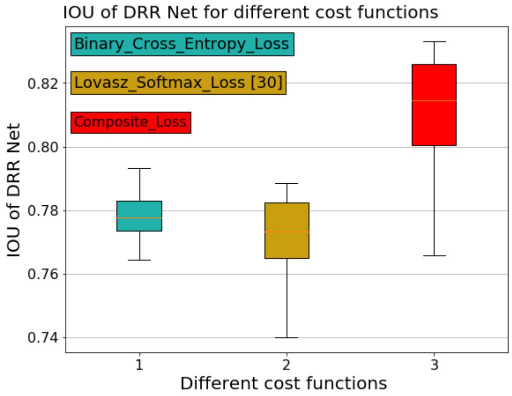


FIGURE 4. Box plot of Intersection Over Union of proposed model for different loss functions.

The proposed model has been trained separately with binary cross entropy loss function, Lovasz softmax loss function and also with composite loss function to observe its combination effect. When trained with BCE only, the proposed model took a long time before showing an improvement. When trained with Lovasz softmax loss function, the proposed model showed better performance at earlier iterations but did not maintain the same at later iterations. However, the model trained with combination of loss functions maintained its progress over the iterations. The IOU values of DRR Net when trained with individual loss functions and composite loss function is reported in Fig. 4. It can be observed

[2] Intel Xeon Processor E5-2650 v4@2.20 GHz

that the proposed architecture trained with composite loss function ($BCE + LZS$) yields better IOU values as compared to other two loss functions. This is due to an uplift in the margin for correct predictions while minimizing the errors that penalize IOU most of the times.

IV. SIMULATION RESULTS AND DISCUSSION

The proposed model and some of the semantic segmentation architectures are trained with the same hyper parameters and the loss function is considered as the composite loss function. The number of training iterations is the same for all models. The proposed DRR Net does not depend on any pre-trained weight set and it is instead trained end-to-end. To perform comparative analysis, quality metrics such as IOU, Road accuracy, Precision and Recall values are evaluated at the end of every group of 12,420 iterations and also at the end of the training phase. All training images of the dataset can be fed to the model in 12,420 number of iterations. The quality metrics are obtained by considering test images as input. Few of the considered test images are shown in Fig. 8a, 9a, 10a and 11a. Road accuracy and mean IOU values are considered to measure the variability of these performance metrics. Figs. 5 and 6 represent the boxplots of road accuracy and mean IOU values of semantic segmentation architectures.

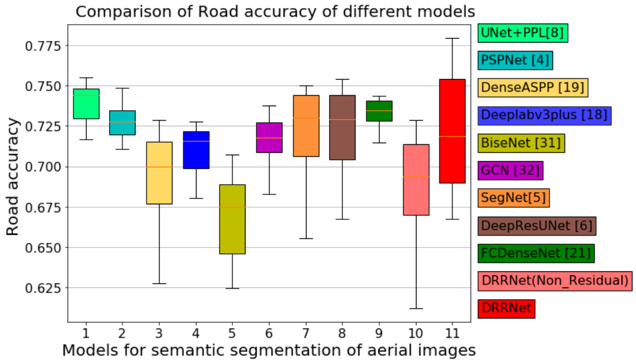


FIGURE 5. Box plot of Road accuracy of models.

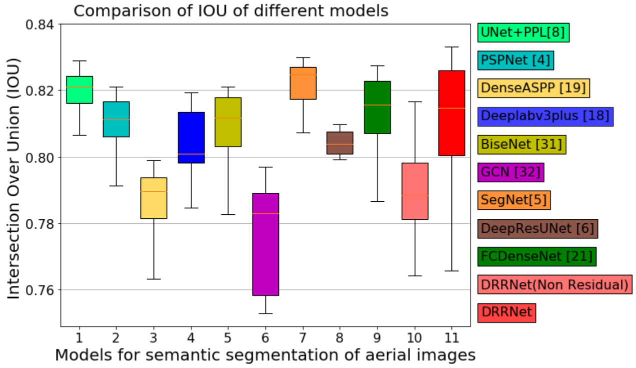


FIGURE 6. Box plot of Intersection Over Union of different models.

From Fig. 5 one can observe that the proposed DRR Net produces a wide range of road accuracy values. Additionally, it can be observed that the proposed model provides a 11.19% improvement over its initial value to reach a maximum value. This is comparable with other models and implies that the proposed model has good learning ability when compared with other architectures. Another measure to quantify a semantic segmentation technique is IOU or Jaccard index. IOU estimates the percentage of pixel overlap between semantic map and its corresponding ground truth. Fig. 6 reports that the proposed DRR model and the model given in [32] exhibit the same higher level of IOU variability. The proposed model reaches a maximum IOU value from an initial overlap of 76.57 per.comcent between predicted and ground truth image. In addition to this one can observe that the models Deep LabV3+ [12], FC-DenseNet [15] and BiseNet [31] possess a narrow range of IOU values. Further, box lengths of the remaining models is observed to be smaller. Table 1 lists the parameters of models and their corresponding performance metrics. The performance metric are evaluated by inferring the models at the end of the training. From Table 1, considering the number of trainable parameters the descending order of models is given by UNet+PPL[3] [21], PSPNet [10], DeepLabV3+ [12], BiseNet [31], DenseASPP [13], GCN [32], SegNet [7], FC-DenseNet [15], Deep ResUNet [18] and DRR Net. The order implies that UNet+PPL [21] model requires maximum number of trainable parameters while the proposed model has least number of trainable parameters. Thus, the road accuracy of DRR Net is significantly superior to other models which also showed discrimination in the corresponding Precision and Recall values. The parameters of models together with proposed DRR Net are represented in Bar graph which is shown in Fig. 7 and it reveals that the proposed DRR Net have far fewer parameters (2.63 million) compared to other models.

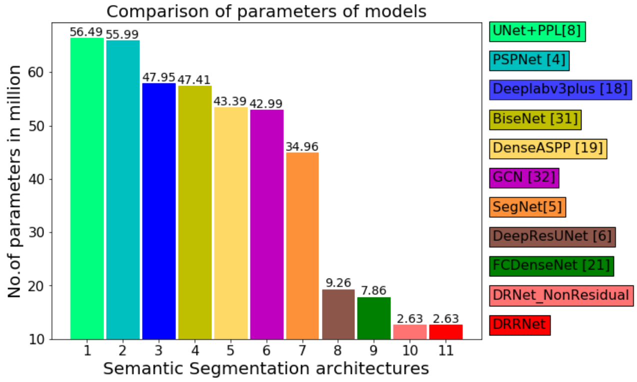


FIGURE 7. Bar graph for comparison of parameters of different models.

A. COMPUTATIONAL COMPLEXITY ANALYSIS

In this section, an elaborate discussion of computational complexity of all architectures including the proposed architecture is presented. In Table 2, the total training time

[3] Modified version of original architecture

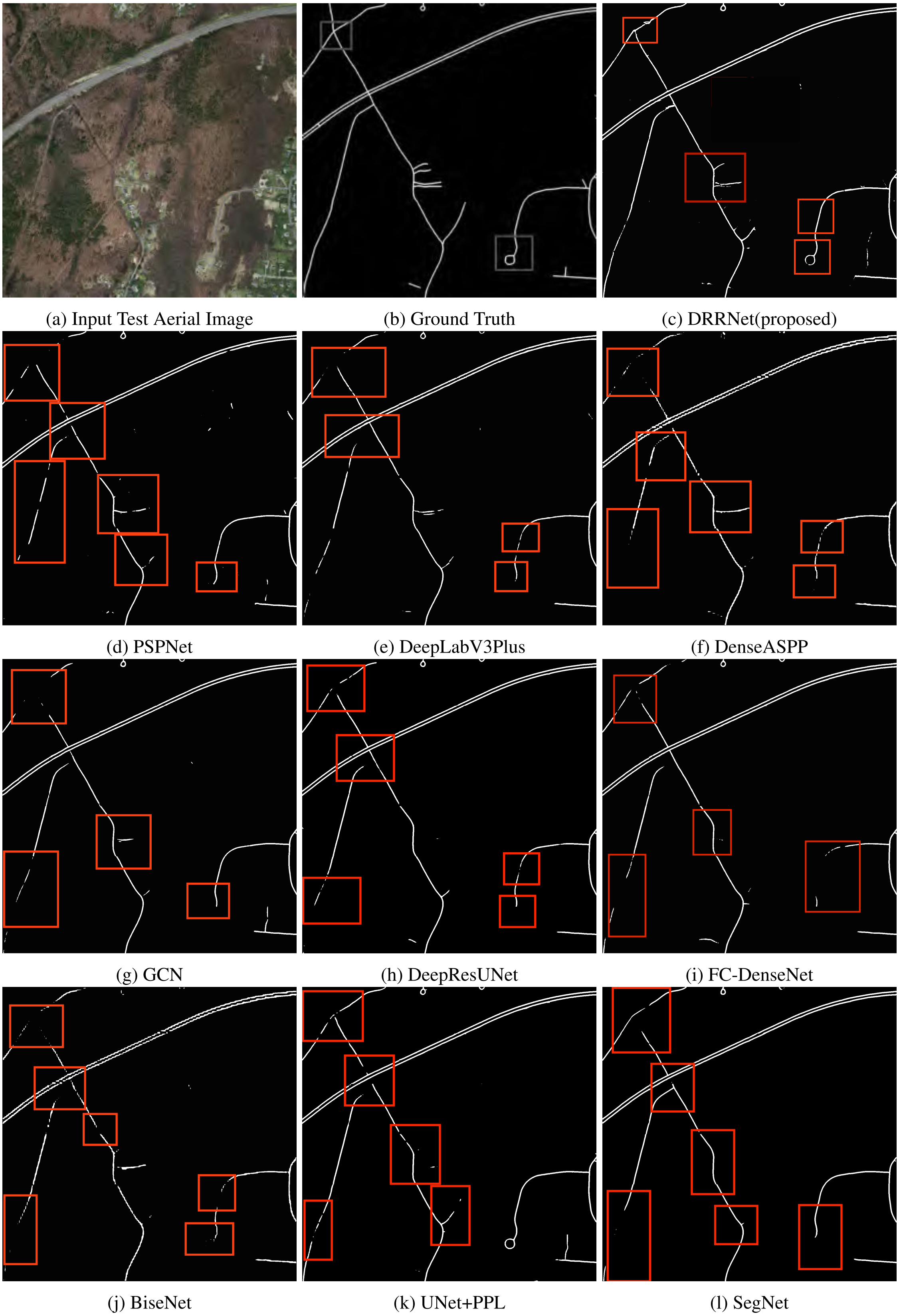
TABLE 1. Comparison of quality metrics of semantic segmentation of aerial images.

Model	PreTrained	Mean IOU	Road Accuracy	Precision	Recall	# Parameters (in million)
PSPNet [10]	√	0.820	0.7483	0.9827	0.9795	55.90
DeepLabV3Plus [12]	√	0.819	0.7276	0.9839	0.9800	47.95
BiseNet [31]	√	0.797	0.7073	0.9815	0.9765	47.41
Dense ASPP [13]	√	0.799	0.7384	0.9820	0.9771	43.39
GCN [32]	√	0.821	0.7376	0.9844	0.9801	42.99
FC-DenseNet [15]	×	0.829	0.7434	0.9846	0.9812	9.26
DeepResUNet [18]	×	0.8304	0.7520	0.9841	0.9812	7.85
UNet+PPL [21]	×	0.829	0.7434	0.9846	0.9812	56.49
SegNet [7]	×	0.822	0.7320	0.9841	0.9812	34.96
DRR Net(Non-Residual (N=4))	×	0.816	0.7287	0.9848	0.9808	2.63
DRR Net (N=4)	×	**0.833**	**0.7794**	**0.9805**	**0.9792**	**2.63**

TABLE 2. Comparison of FLOPS, Training and average Test run time of all models.

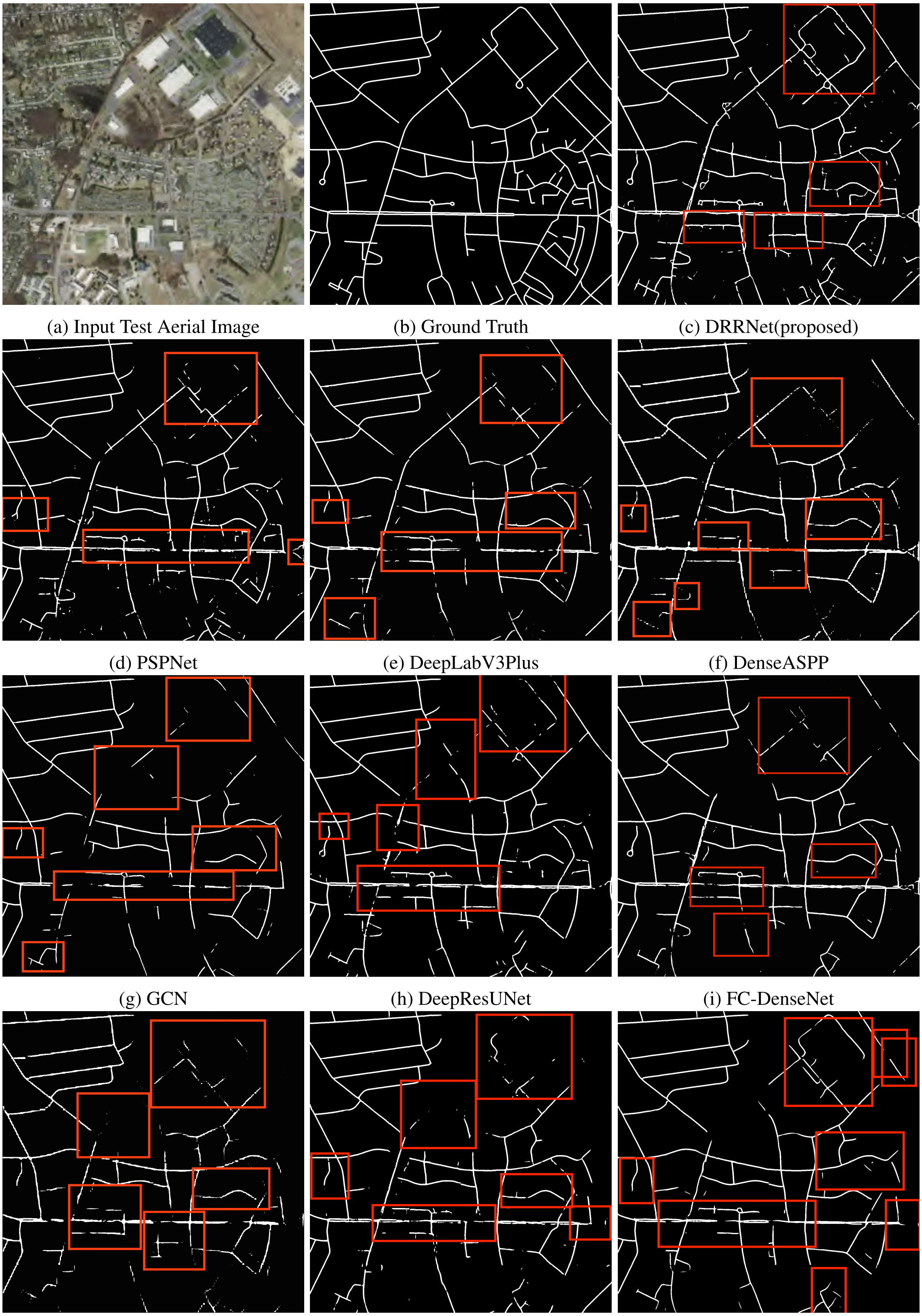
Model	Training time per image(sec)	Total Training time(Hours)	Average Test run time(sec)	FLOPS(in billion)
PSPNet [10]	0.32	48	0.41	62.5
DeepLabV3Plus [12]	0.26	37	0.31	32.8
BiseNet [31]	0.34	46.35	0.32	20.4
Dense ASPP [13]	0.19	28	0.15	22.1
GCN [32]	0.375	52	0.395	20.9
FC-DenseNet [15]	0.51	70.63	4.17	52.3
DeepResUNet [18]	0.39	57.81	0.094	77.6
UNet+PPL [21]	0.65	89.5	1.30	144.1
SegNet [7]	0.54	75.4	0.916	90.0
DRR Net(Non-Residual (N=4))	0.63	86.8	4.66	80.1
DRR Net (N=4)	**0.63**	**86.8**	**4.66**	**80.1**

(per-image and also for all images of the dataset), the average test run time and the number Floating point operations (FLOPS) of all models are presented. The total training time is defined as the time taken to train individual architectures. The average test run time is defined as the average time required to infer the trained model over the total number of test images. It can be observed that, pre-trained network architecture based models such as Dense ASPP [13], Deeplabv3+ [12], PSPNet [10], BiseNet [31] and GCN [32], require comparatively less training time than other models. The proposed DRR model and the model presented in [15] are built with dense convolutions. Because of the concatenation of features from the specified number of convolutional layers, these models need longer training time as contrary to other models. In the DeepResUNet model proposed in [18], the training time is considerably reduced due to presence of residual connections. Due to the concatenation of feature maps after pooling with different scales, the UNet+PPL [21] model demands increased training time as contrary to other models. The total time allocated to train



(a) Input Test Aerial Image
(b) Ground Truth
(c) DRRNet(proposed)
(d) PSPNet
(e) DeepLabV3Plus
(f) DenseASPP
(g) GCN
(h) DeepResUNet
(i) FC-DenseNet
(j) BiseNet
(k) UNet+PPL
(l) SegNet

FIGURE 8. **Predicted images of semantic segmentation models of Fig. 8(a).**



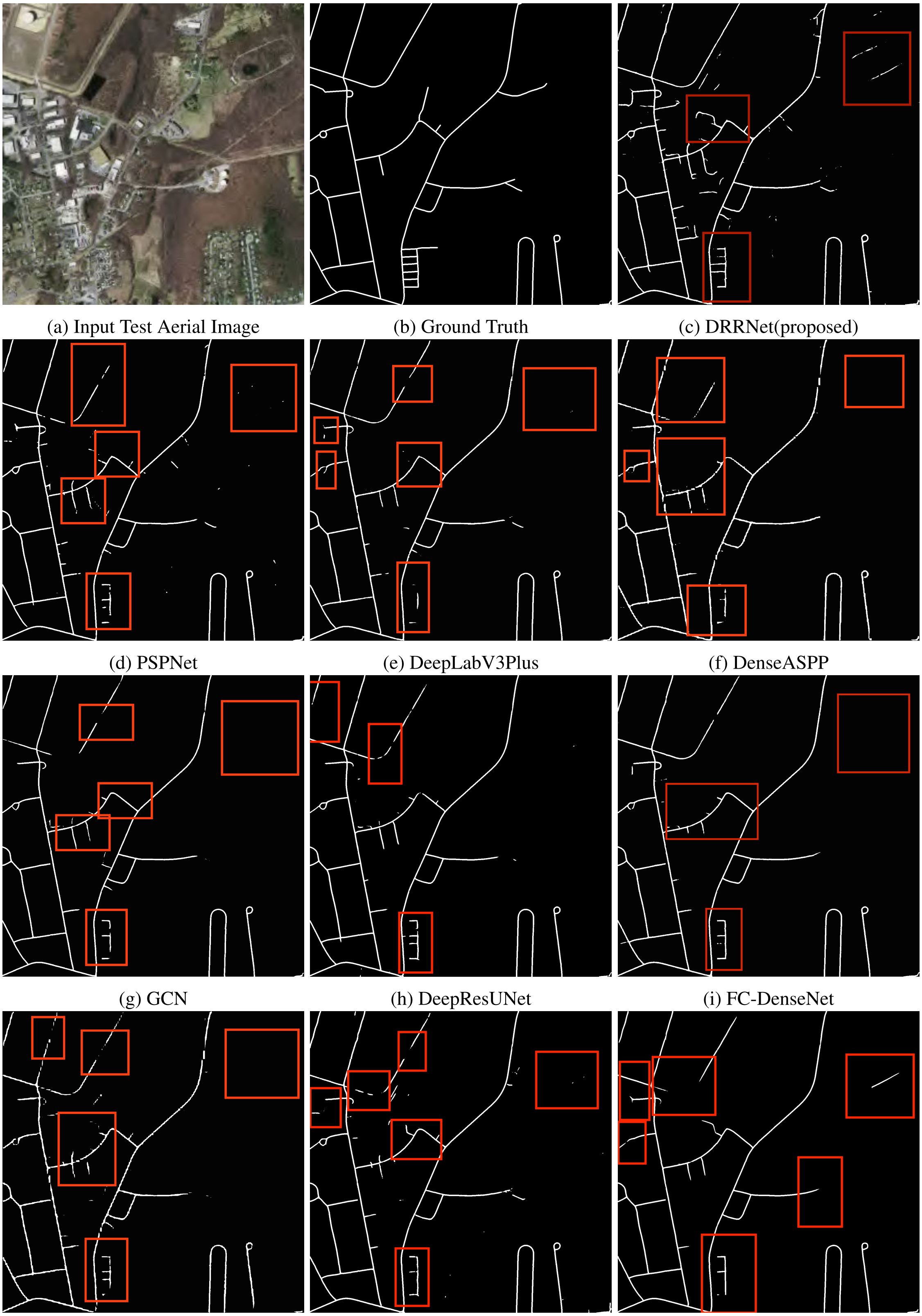
(a) Input Test Aerial Image (b) Ground Truth (c) DRRNet(proposed)

(d) PSPNet (e) DeepLabV3Plus (f) DenseASPP

(g) GCN (h) DeepResUNet (i) FC-DenseNet

(j) BiseNet (k) UNet+PPL (l) SegNet

FIGURE 9. Predicted images of semantic segmentation models of Fig. 9(a).



(a) Input Test Aerial Image

(b) Ground Truth

(c) DRRNet(proposed)

(d) PSPNet

(e) DeepLabV3Plus

(f) DenseASPP

(g) GCN

(h) DeepResUNet

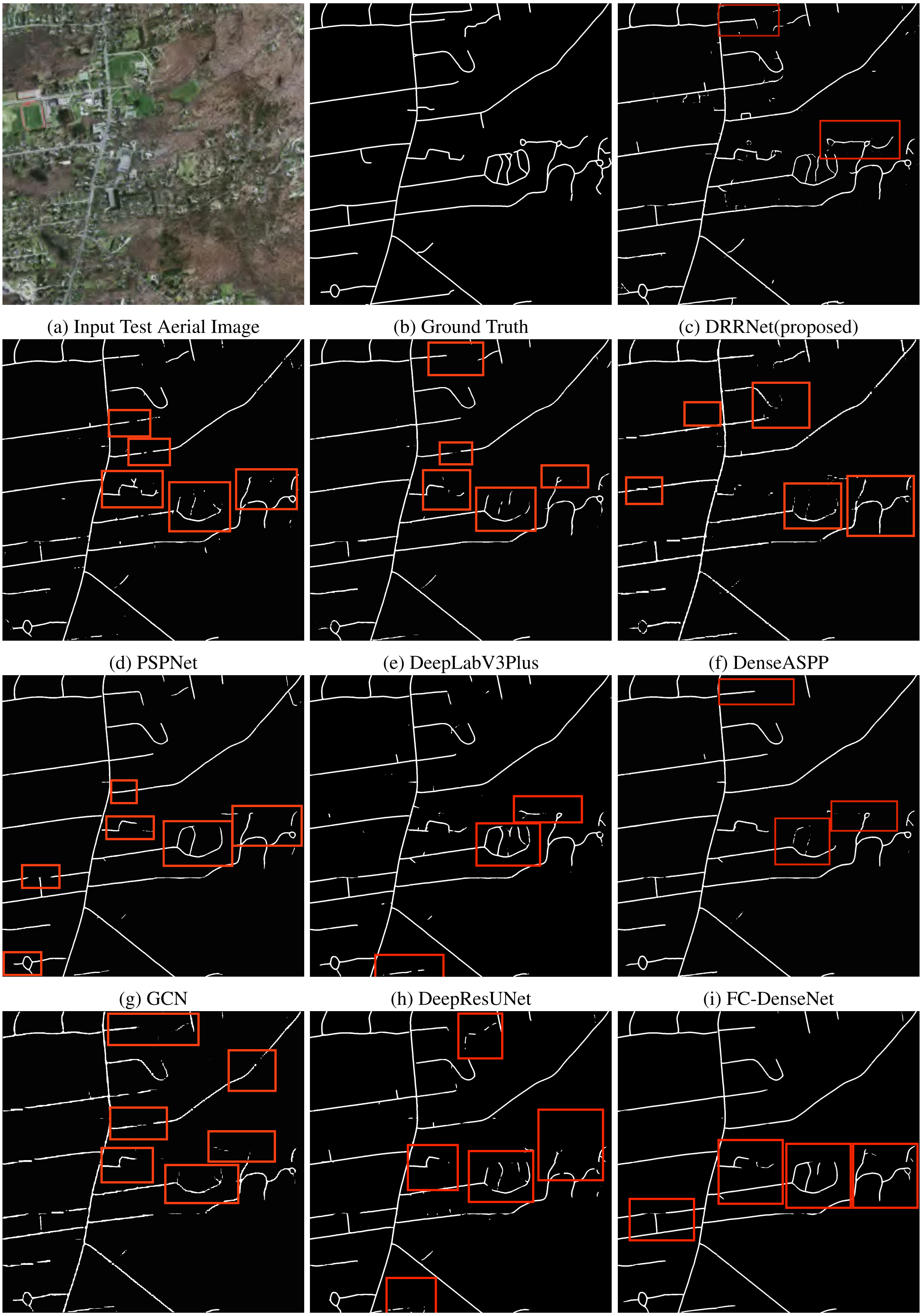
(i) FC-DenseNet

(j) BiseNet

(k) UNet+PPL

(l) SegNet

FIGURE 10. Predicted images of semantic segmentation models of Fig. 10(a).



(a) Input Test Aerial Image

(b) Ground Truth

(c) DRRNet(proposed)

(d) PSPNet

(e) DeepLabV3Plus

(f) DenseASPP

(g) GCN

(h) DeepResUNet

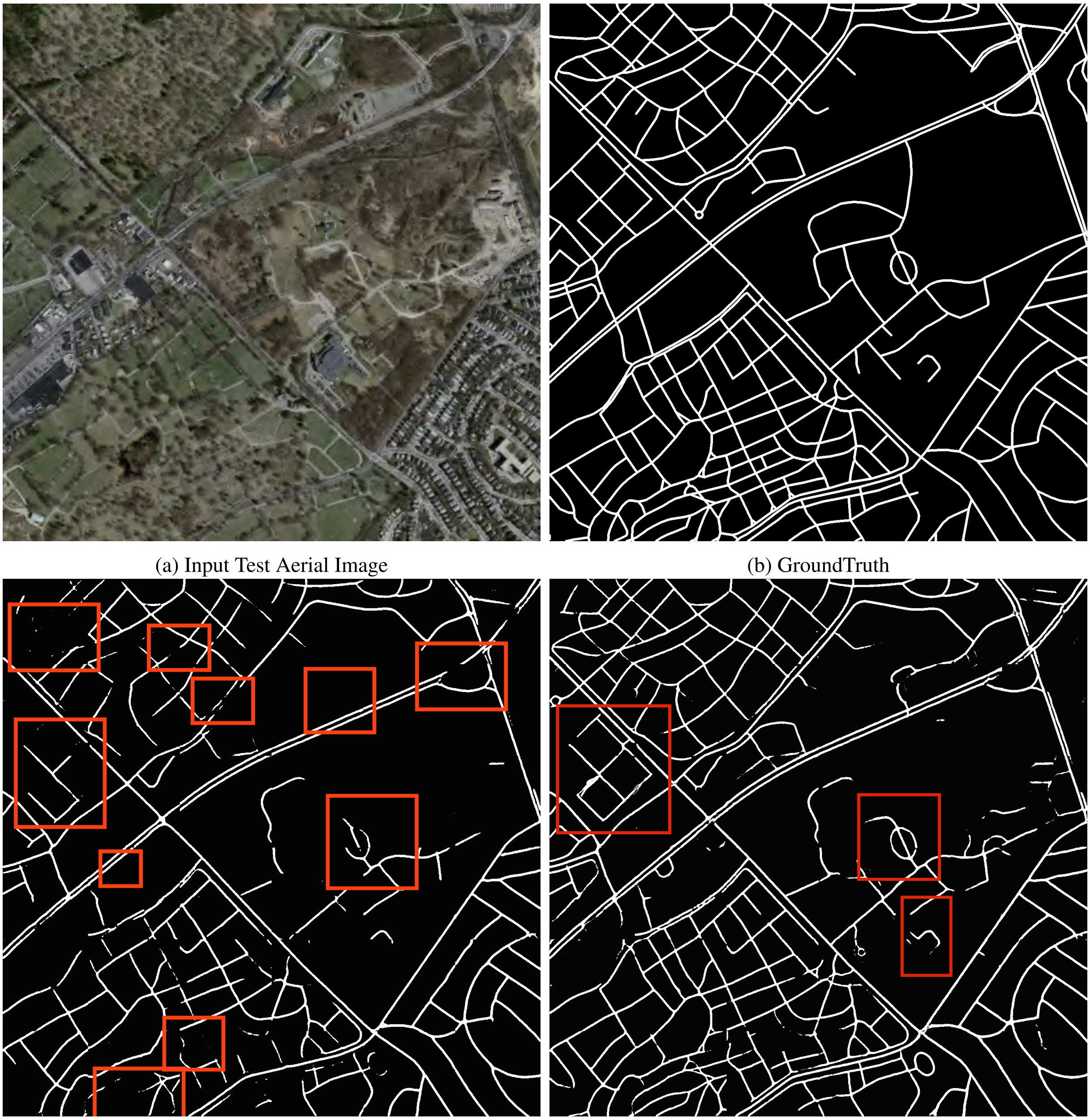
(i) FC-DenseNet

(j) BiseNet

(k) UNet+PPL

(l) SegNet

FIGURE 11. Predicted images of semantic segmentation models of Fig. 11(a).



(a) Input Test Aerial Image

(b) GroundTruth

(c) DRRNet(W/o Residual)

(d) DRRNet(Proposed)

FIGURE 12. Predicted images of DRR Net with and without residual connection.

all models is 678 hours. Referring to Table 2, from the average test run time of all models, FC-DenseNet [15], DRR Net requires a longer time to load the learned weights during a forward pass from dense convolutions of the trained model. Due to dense connectivity, the features and gradients have to flow through multiple paths during forward and backward propagations. This leads to an increase in training and testing times of DRR Net though the number of parameters is less. The increasing order of computational complexity of models (in terms of FLOPs) is BiseNet [31], GCN [32], DenseASSP [13], DeepLabV3+ [12], FC-DenseNet [15], PSPNet [10], DeepResUNet [18], DRR Net (proposed), SegNet [7] and U-NetPPL [21]. The proposed DRR Net ranks third in increasing computational complexity order.

For the test images in Fig. 8a, 9a, 10a and 11a, the segmentation maps produced by proposed and the state-of-the-art architectures along with ground truth images are presented in Fig. [8b - 8l], [9b - 9l], [10b - 10l] and [11b - 11l] respectively. To highlight the performance of the DRR Net, its predicted images are compared with state-of-the-art-models by highlighting some parts of the image with red color boxes. Fig. [8b - 8l] represents the segmentation output of all models

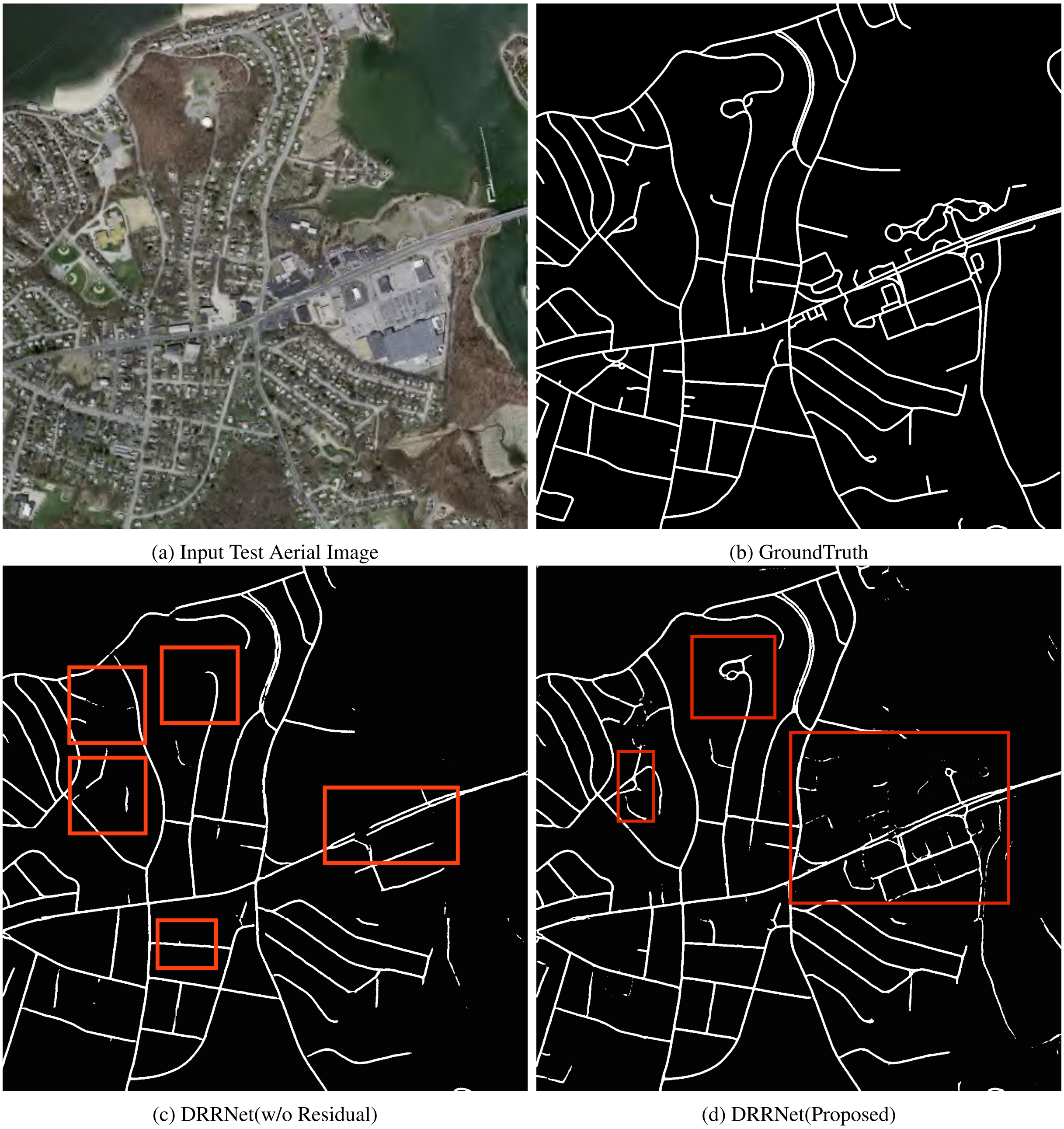


FIGURE 13. Predicted images of DRR Net with and without residual connection.

for the test input aerial image of Fig. 8a. From these predicted images, it can be seen that the DRR Net extracts round-shaped roads and also the intersections of roads without any gap. The predicted images of the test input images Fig 9a, 10a and 11a are shown in Fig. [9b - 9l], [10b - 10l] and [11b - 11l]. These images reveal that the proposed DRR Net differentiates the parallel, smaller and diverse-shaped road regions clearly from other regions. The segmentation maps for another set of input test aerial images are presented in Section VI. To show the importance of residual connections in the proposed DRR Net the model has been trained by removing the residual connection. Due to the removal of residual connections, there is no sharing of initial features of each module to successive DRR modules of the proposed architecture. This leads to a reduction in the learning ability of network. The prediction results of the proposed architecture with and without residual connections are presented in Fig. 12 and 13 along with input and ground truth images. From these predicted images it can be seen that the DRR Net clearly distinguished the road pixels better than the DRR Net without residual connections. The quality metrics of the proposed Net with and without residual are quantified in Table 1. These values reveal

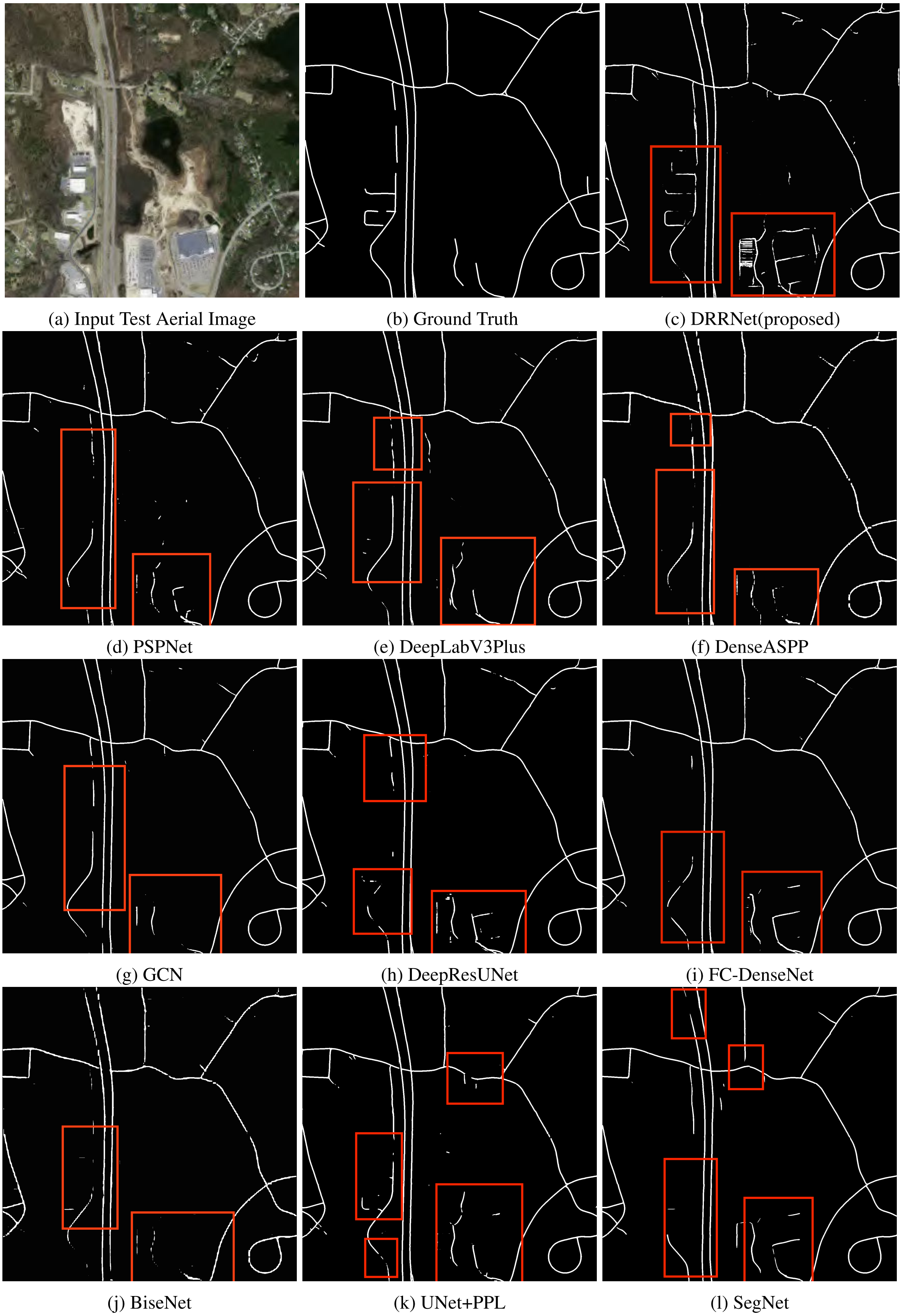
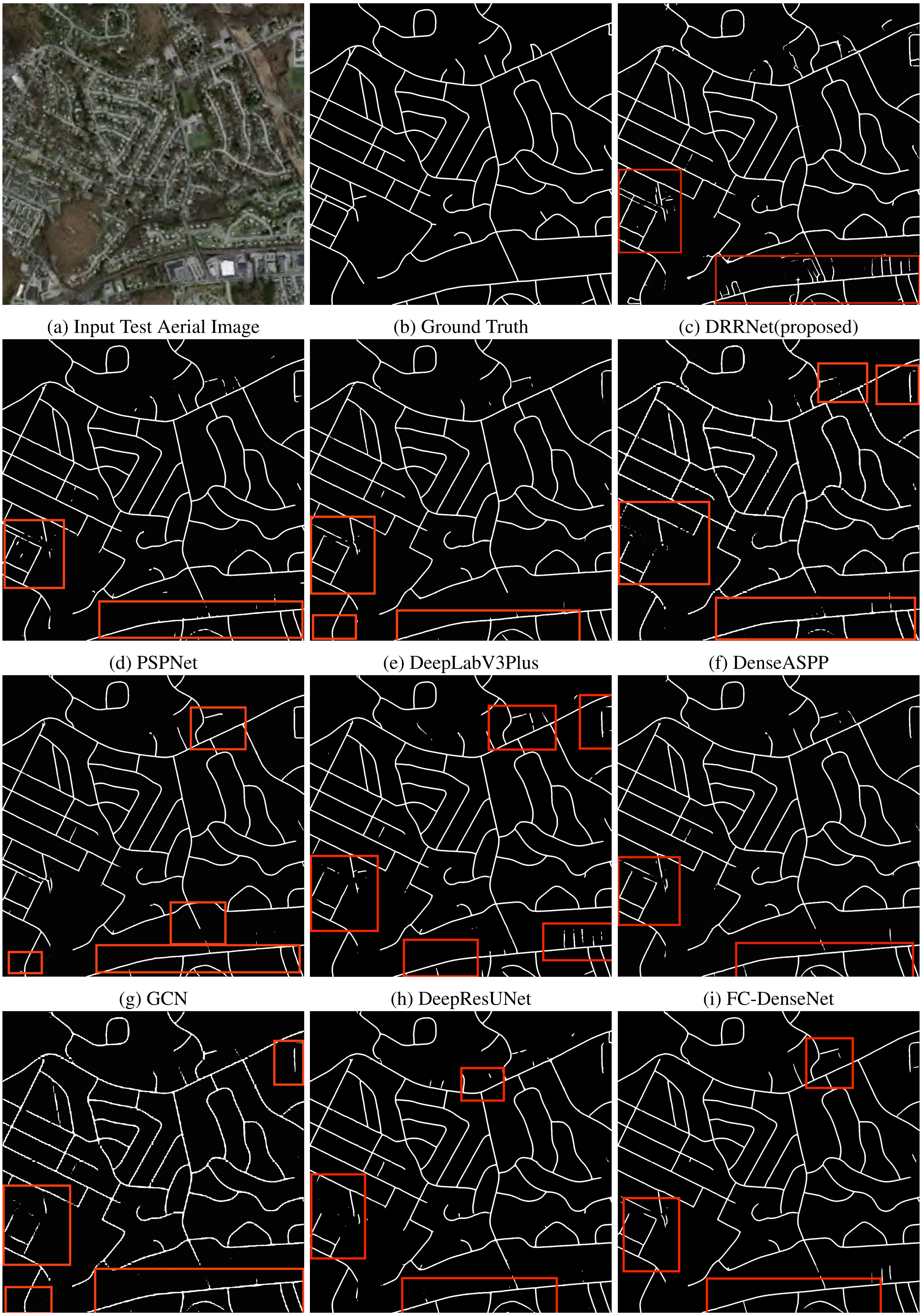

(a) Input Test Aerial Image
(b) Ground Truth
(c) DRRNet(proposed)
(d) PSPNet
(e) DeepLabV3Plus
(f) DenseASPP
(g) GCN
(h) DeepResUNet
(i) FC-DenseNet
(j) BiseNet
(k) UNet+PPL
(l) SegNet

FIGURE 14. **Predicted images of semantic segmentation models of Fig. 14(a).**



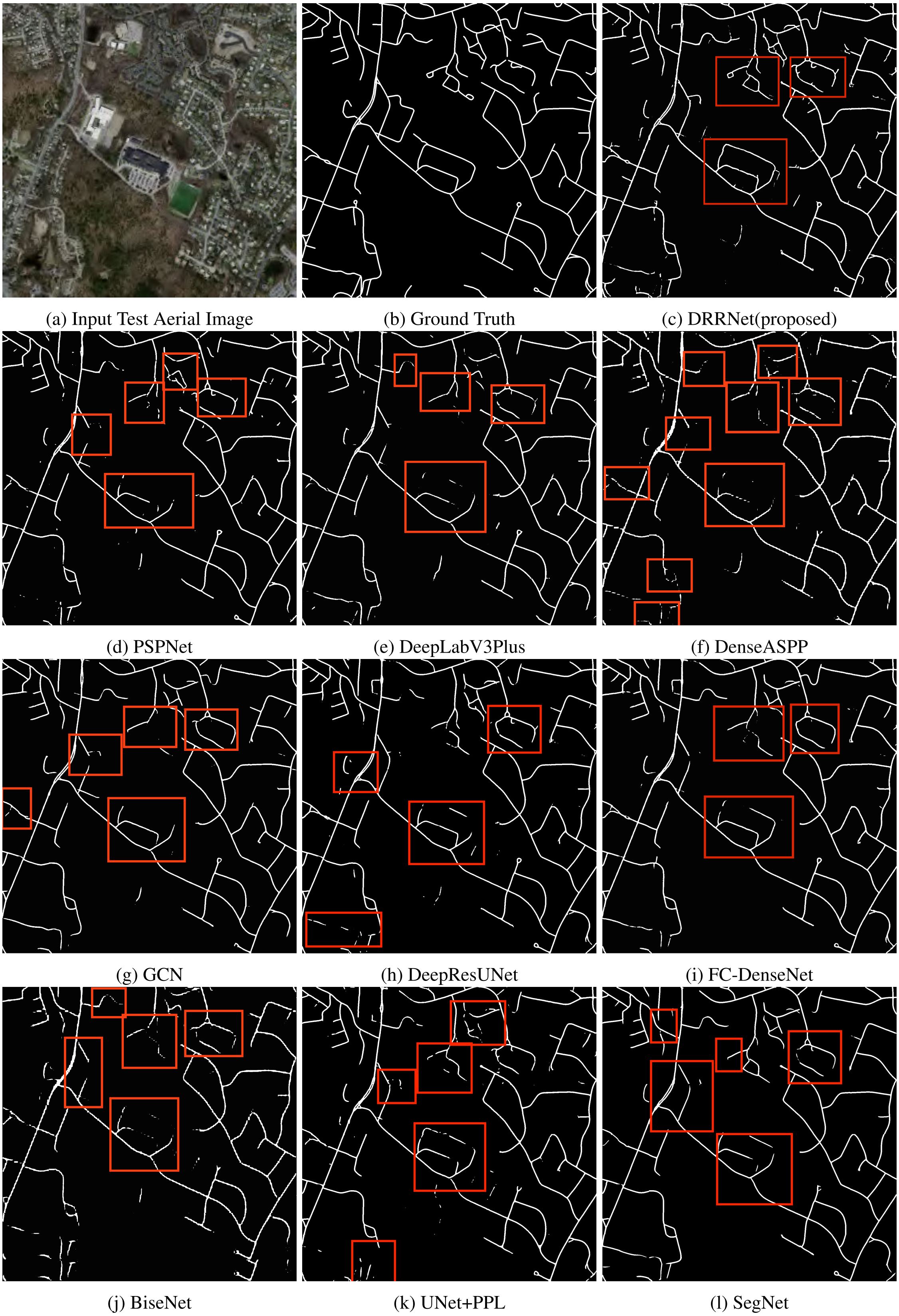
(a) Input Test Aerial Image (b) Ground Truth (c) DRRNet(proposed)

(d) PSPNet (e) DeepLabV3Plus (f) DenseASPP

(g) GCN (h) DeepResUNet (i) FC-DenseNet

(j) BiseNet (k) UNet+PPL (l) SegNet

FIGURE 15. Predicted images of semantic segmentation models of Fig. 15(a).

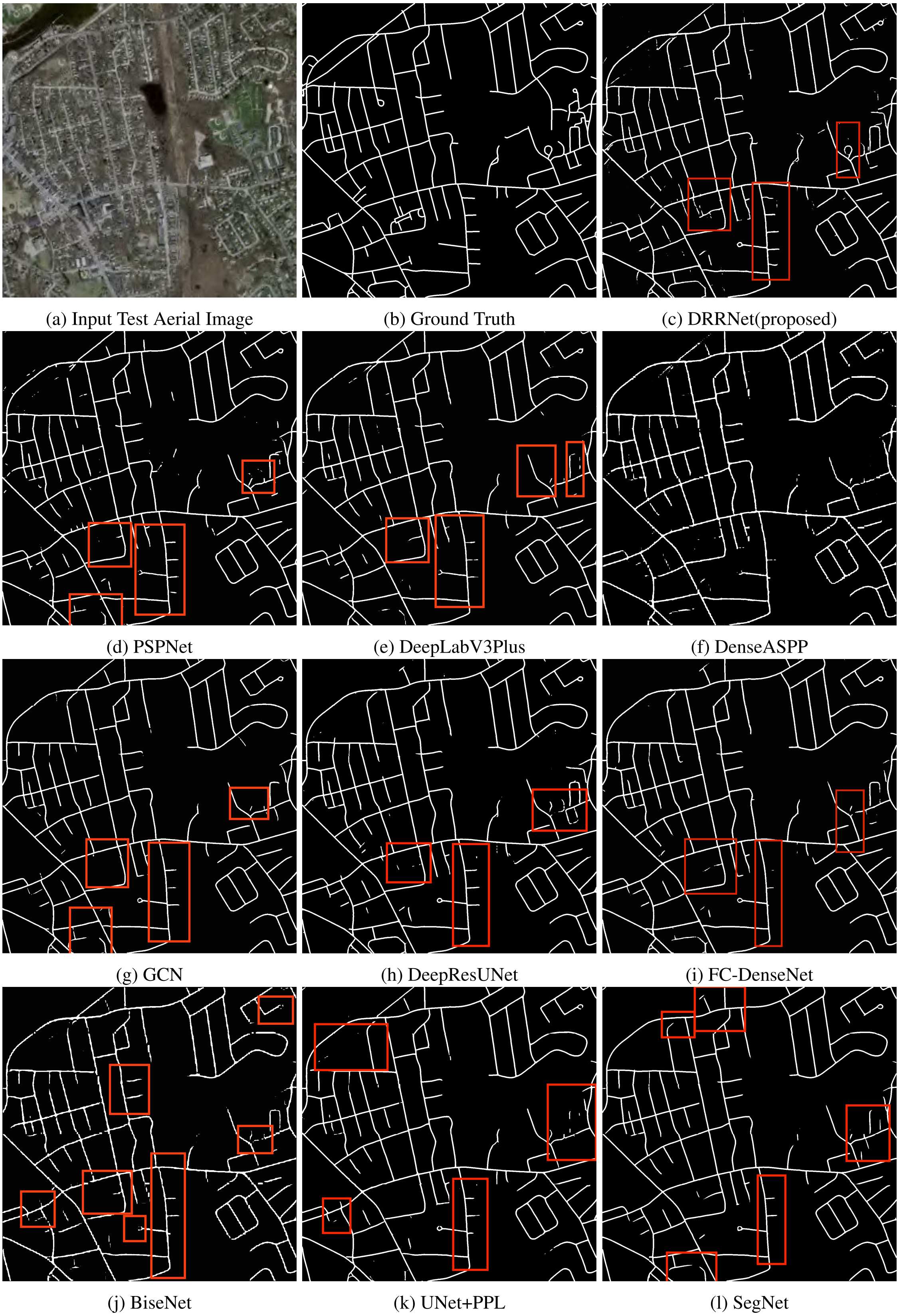

(a) Input Test Aerial Image (b) Ground Truth (c) DRRNet(proposed)

(d) PSPNet (e) DeepLabV3Plus (f) DenseASPP

(g) GCN (h) DeepResUNet (i) FC-DenseNet

(j) BiseNet (k) UNet+PPL (l) SegNet

FIGURE 16. Predicted images of semantic segmentation models of Fig. 16(a).



(a) Input Test Aerial Image

(b) Ground Truth

(c) DRRNet(proposed)

(d) PSPNet

(e) DeepLabV3Plus

(f) DenseASPP

(g) GCN

(h) DeepResUNet

(i) FC-DenseNet

(j) BiseNet

(k) UNet+PPL

(l) SegNet

FIGURE 17. Predicted images of semantic segmentation models of Fig. 17(a).

that the residual connections play a vital role in producing better IOU, road accuracy, precision and recall values. The computational complexity of the two models remains the same in terms of training time, test time and FLOPS, which are presented in Table 2. After observing all predicted images of DRR Net, it is clear that the model precisely differentiated pixels of smaller, curved and parallel roads from background pixels. In addition to this the proposed architecture provided good separation of roads when background pixels are the majority in number.

V. CONCLUSION

In this paper, a semantic segmentation architecture named DRR Net is proposed to segment roads in high-resolution aerial imagery data. The proposed DRR model was able to precisely segment roads and achieve prominent results on Massachusetts roads dataset as compared with state-of-the-art semantic segmentation architectures. The qualitative and quantitative results showed that the DRR Net could segment all kinds of roads including variable-extent roads and also non-labeled roads. A comparison of the proposed architecture has been done with the diversified semantic segmentation architectures based on normal convolutions, atrous convolutions, global convolutions and dense convolutions. Among all other models, the proposed model showed remarkable performance in all aspects including background-dominant scenarios.

The distinctive performance of the proposed architecture can be attributed to the iterative reuse of collective knowledge acquired at various scales through dense, residual connections and the connectivity of DRR modules. It can be noted that the iterative reuse leads to an increase in receptive field for pixels of less frequent classes (road pixels). The proposed architecture achieved a ∼2.74% increase in road accuracy with a contemporary tenfold reduction in the number of parameters. Moreover, the proposed architecture offered good discrimination of roads in all scenarios. Additionally, the proposed DRR Net architecture can also be used to segment other kinds of objects like buildings, dams, trees, etc.

APPENDIX

See Figs. 14–17.

ACKNOWLEDGMENT

(Karuna Kumari Eerapu and Balraj Ashwath contributed equally to this paper.)

REFERENCES

[1] M. Amo, F. Martínez, and M. Torre, "Road extraction from aerial images using a region competition algorithm," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1192–1201, May 2006.

[2] J. Hu, A. Razdan, J. C. Femiani, M. Cui, and P. Wonka, "Road network extraction and intersection detection from aerial images by tracking road footprints," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4144–4157, Dec. 2007.

[3] S. Movaghati, A. Moghaddamjoo, and A. Tavakoli, "Road extraction from satellite images using particle filtering and extended Kalman filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2807–2817, Jul. 2010.

[4] J. Yuan, D. Wang, B. Wu, L. Yan, and R. Li, "LEGION-based automatic road extraction from satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4528–4538, Nov. 2011.

[5] S. Das, T. T. Mirnalinee, and K. Varghese, "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3906–3931, Oct. 2011.

[6] C. Unsalan and B. Sirmacek, "Road network detection using probabilistic and graph theoretical methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4441–4453, Nov. 2012.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," 2015, *arXiv:1511.00561*. [Online]. Available: https://arxiv.org/abs/1511.00561

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[9] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 346–361.

[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*. [Online]. Available: https://arxiv.org/abs/1802.02611

[13] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[15] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1175–1183.

[16] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," 2017, *arXiv:1611.08323*. [Online]. Available: https://arxiv.org/abs/1611.08323

[17] M. Samy, K. Amer, K. Eissa, M. Shaker, and M. ElHelw, "Nu-Net: Deep residual wide field of view convolutional neural network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 267–2674.

[18] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[19] O. Filin, A. Zapara, and S. Panchenko, "Road detection with EOSResUNet and post vectorizing algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 211–215.

[20] T. Sun, Z. Chen, W. Yang, and Y. Wang, "Stacked u-nets with multi-output for road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 187–1874.

[21] J. H. Kim, H. Lee, S. J. Hong, S. Kim, J. Park, J. Y. Hwang, and J. P. Choi, "Objects segmentation from high-resolution aerial images using U-Net with pyramid pooling layers," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 115–119, Jan. 2019.

[22] S. Aich, W. van der Kamp, and I. Stavness, "Semantic binary segmentation using convolutional networks without decoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–1824.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.

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

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[26] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 210–223.

[27] 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.

[28] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[30] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4413–4421.

[31] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.

[32] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4353–4361.

KARUNA KUMARI EERAPU received the B.E. and M.Tech. degrees in electronics and communication engineering from Jawaharlal Nehru Technological University, in 2009 and 2011, respectively. She is currently pursuing the Ph.D. degree with the National Institute of Technology Karnataka, Surathkal, India. Her current research interests include image processing, machine learning, deep learning, semantic segmentation, and classification.

BALRAJ ASHWATH was born in Bengaluru, India, in 1998. He is currently pursuing the bachelor's degree in electronics and communication engineering (ECE) with the National Institute of Technology Karnataka (NITK), Surathkal. His research interests include artificial intelligence, computer vision, deep learning, pattern recognition, and machine learning. His current research interest includes applying deep learning techniques on different computer vision problems.

SHYAM LAL received the Ph.D. degree in digital image processing from the Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India, in 2013. He has been an Assistant Professor with the Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, India, since 2013. He has more than 16 years of teaching and research experience. He has supervised three Ph.D. students, and five Ph.D. students are currently working under his supervision in the area of medical and remote sensing image processing. He has published more than 65 research papers in the area of digital image processing, medical image processing, and remote sensing at international/national journals and conferences. He has supervised three doctoral students in the area of image processing. His research interests include digital image processing, histopathology image processing, medical image processing, remote sensing image processing, application of deep learning, and optimization algorithms in digital image processing in general. He is a Senior Member of the IEEE, a Life Member of ISTE, New Delhi, India, a Life Member of IAENG, Hong Kong, and a Life Member of IACSIT, Singapore. He received the Early Career Research Award (Young Scientist) from the Science Engineering and Research Board, Department of Science and Technology, Government of India, in 2017, and the Young Faculty Research Fellowship Research Grant under the Visvesvaraya Ph.D. Scheme for Electronics & IT, MEITY, Government of India, in 2019. He has been a Guest Editor of IJSISE (Inderscience Publishers) and an Editorial Member of the *Open Access Journal of Biomedical Engineering and its Applications* (Lupine Publishers, USA).

FABIO DELL'ACQUA received the five-year degree *(cum laude)* (Hons.) in electronics engineering and the Ph.D. degree in remote sensing from the University of Pavia, Italy, in 1996 and 1999, respectively. In 2000, he was an Associate Researcher with the Division of Informatics, University of Edinburgh, U.K. In 2001, he obtained a permanent position as an Assistant Professor with the Department of Electronics, University of Pavia, Italy, where he has been an Associate Professor of remote sensing with the Department of Electrical, Computer and Biomedical Engineering, since 2015. He teaches courses in remote sensing at the University of Pavia. He has established strong links with companies with business in remote sensing applications. His research interests include radar data processing and radar/optical data fusion for risk-related applications. In this area, he is/has been participating to, or leading, several research projects both at national and international levels. From 2011 to 2015, he organized yearly editions of an International Summer School on Data Fusion in Aerospace Applications, which attracted up to 40 students from around the world. In 2014, he co-founded a university spin-off company, named Ticinum Aerospace, to exploit commercially his research results in the use of EO data for risk management. In 2016, he started leading an H2020 MSCA-RISE exchange project, "EOXPOSURE," on the use of remote sensing for analyzing environmental disease spread factors. He is a Life Member of the Technical and Scientific Board of the Lombardy Aerospace Industry Cluster. Currently, his publication records include 55 journal papers, over 160 conference papers, and 15 contributions to books. According to Scopus, he has currently (2019) authored 156 papers, with a total of 1737 citations and a Hirsch index of 21 (excluding self-citations). According to Google Scholar, he has authored 249 papers with a total of 2882 citations and a Hirsch index of 24.

A. V. NARASIMHA DHAN received the B.E. degree in electronics and communication engineering from Andhra University, in 2005, the M.Tech. degree in signal processing from IIT Guwahati, India, in 2007, and the Ph.D. degree from the Indian Institute of Science, India, in 2012. He is currently an Assistant Professor with the Department of Electronics and Communication Engineering, National Institute of Technology, Karnataka, India.

• • •