Deep Learning-based Building Footprint Extraction with Missing Annotations

Jian Kang, *Member, IEEE,* Ruben Fernandez-Beltran, *Senior Member, IEEE,* Xian Sun, *Senior Member, IEEE,* Jingen Ni, *Senior Member, IEEE,* and Antonio Plaza, *Fellow, IEEE*

***Abstract*—Most state-of-the-art deep learning-based methods for extraction of building footprints are aimed at designing proper convolutional neural network (CNN) architectures or loss functions able to effectively predict building masks from remote sensing (RS) images. To properly train such CNN models, large-scale and pixel-level building annotations are required. One common approach to obtain scalable benchmark datasets for segmentation of buildings is to register RS images with auxiliary geospatial information data, such as those available from OpenStreetMaps (OSM). However, due to land-cover changes, urban construction, and delayed geospatial information updating, some building annotations may be missing in the corresponding ground-truth building mask layers. This will likely introduce confusion in the training of CNN models for discriminating between background and building pixels. To solve this important issue, we first formulate the problem as a long-tailed classification one. Then, we introduce a new joint loss function based on three terms: 1) logit adjusted cross entropy (LACE) loss, aimed at discriminating between building and background pixels from a long-tailed label distribution; 2) weighted dice loss, aimed at increasing the F_1 scores of the predicted building masks; and 3) boundary alignment loss, which is optimized for preserving the fine-grained structure of building boundaries. Our experiments, conducted on two benchmark building segmentation datasets, validate the effectiveness of our newly proposed loss with respect to other state-of-the-art losses commonly used for extracting building footprints. The codes of this paper will be publicly available from https://github.com/jiankang1991/GRSL_BFE_MA.**

***Index Terms*—Building extraction, semantic segmentation, deep learning, missing labels, remote sensing.**

I. Introduction

EXTRACTING building footprints from high-resolution remote sensing (RS) images has been a fundamental task within the field of intelligent image interpretation. Footprint maps of buildings play an important role in several different tasks, such as urban planning, disaster monitoring, change detection, and autonomous driving. Thus, accurately generating footprints of buildings is always an on-going and hot topic in the RS community. Nowadays, with the rapid development of satellite sensors, massive volumes of high-resolution RS images are available for developing effective building footprint extraction techniques. Moreover, such big data also foster the development of deep learning-based methods for extracting footprints of buildings in an end-to-end manner [1]–[5].

Fig. 1. An illustration of a satellite image containing buildings with missing annotations, where the cyan color denotes the available ground-truth footprints and the red color indicates the missing footprints.

One of the first convolutional neural network (CNN) architectures adopted for the extraction of building footprints was the fully convolutional network (FCN) [6], which replaces fully connected layers with convolutional layers to create building masks of the same size with respect to the input RS images. Liu *et al.* proposed an encoder-decoder CNN framework [with a spatial residual inception (SRI) module] for capturing and fusing the multi-scale features during the phase of building extraction [7]. Based on the feature pyramid network, Wei *et al.* developed a multi-scale aggregation FCN with polygon regularization for refining the boundaries of buildings [8]. In order to accelerate the computational performance of an encoder-decoder framework intended to process

This work was in part supported by the Ministry of Science, Innovation and Universities of Spain (RTI2018-098651-B-C54), the Valencian Government of Spain (GV/2020/167), FEDER-Junta de Extremadura (Ref. GR18060) and the European Union under the H2020 EOXPOSURE project (No. 734541). *(Corresponding author: Jingen Ni)*

J. Kang and J. Ni are with the School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: jiankang@suda.edu.cn; jni@suda.edu.cn).

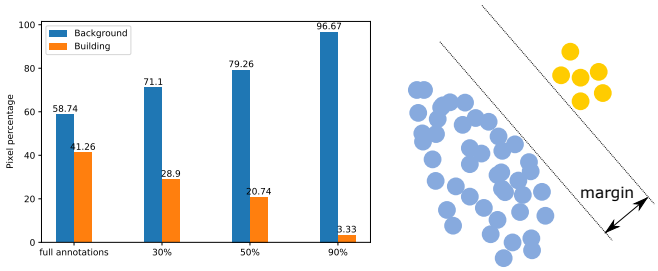
R. Fernandez-Beltran is with the Institute of New Imaging Technologies, University Jaume I, 12071 Castellón de la Plana, Spain (e-mail: rufernan@uji.es).

X. Sun is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunxian@mail.ie.ac.cn).

A. Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain. (e-mail: aplaza@unex.es).

very large input images, Li *et al.* introduced a multiple-feature reuse network (MFRN) that enabled the direct use of hierarchical features and achieved prominent building segmentation performance [9]. The feature pairwise conditional random field (FPCRF) integrated in CNN model was also used for preserving sharp boundaries and fine-grained building segments [10]. By combining a multi-scale feature extraction strategy and attention mechanisms, Zhu *et al.* proposed a multiple attending path neural network (MAP-Net) which could precisely generate multi-scale footprints of buildings and accurate polygons [11]. Rather than learning the building masks, PolygonCNN was also proposed for directly generating vector building polygons based on an encoder-decoder CNN framework, in an end-to-end manner [12]. Another perspective for designing deep learning-based approaches for the extraction of building footprints was based on the considered loss function. Although most of the above-mentioned methods exploit the cross entropy (CE) loss, there are some works aimed at optimizing the loss design for accurately predicting building regions and boundaries. For example, Yuan proposed to use the signed distance function, which calculates the distance from the pixels to their nearest points on the boundaries, to accurately capture the building shapes [13]. Wu *et al.* exploited the boundary loss as a regularizer of the region-based CE loss for extracting building segments and outlines [14]. Bokhovkin *et al.* introduced a differentiable (surrogate) loss for penalizing the misalignment of building boundaries [15].

All the above-mentioned methods for building footprint extraction require accurate building area annotations. There are of course unsupervised solutions like [16] that can achieve remarkable building footprint extraction from aerial remote sensing data in a efficient unsupervised way, but most building segmentation benchmark datasets are just constructed based on the geo-registration between RS images and some auxiliary geospatial information data, e.g., OpenStreetMaps (OSM), in order to avoid the expensive and time-consuming human labeling procedure. Nonetheless, under this scenario, missing annotations (Figure 1) may often appear in the corresponding ground-truth building mask layer due to several reasons, including land-cover changes, urban construction, delayed updating, or even low-quality volunteered geographic information (VGI). Logically, all these factors may result in the potential confusion of trained CNN models when discriminating between the background and building pixels. To relieve these issues, we first formulate the problem as a long-tailed classification one. Then, we introduce a new joint loss function that considers the possible existence of missing building annotations in the dataset. Our newly developed loss function includes three terms: 1) logit adjusted cross entropy (LACE) loss, aimed at discriminating between the building and background pixels from a long-tailed label distribution; 2) weighted dice loss, aimed at increasing the F_1 scores of the predicted building masks; and 3) boundary alignment loss, optimized for preserving the fine-grained structures of building boundaries. Our newly proposed loss is evaluated on two benchmark datasets, outperforming other state-of-the-art competitors. The contributions of this letter can be summarized as follows:



(a) (b)

Fig. 2. (a) The background and building pixel percentages of the Massachusetts Buildings dataset with the full, 70%, 50% and 10% building annotations. (b) LACE tries to set a pairwise label margin between the predicted scores between majority and minority classes.

1) To our best knowledge, this is the first paper in the literature that investigates the problem of deep learning-based building segmentation with missing annotations, approaching it from the perspective of designing an effective loss function to specifically deal with this issue.
2) We formulate the task as a long-tailed classification problem and then introduce a new joint loss function.
3) Compared with other state-of-the-art methods, the proposed loss function achieves the best performance on two widely used benchmarks.

The reminder of this letter is structured as follows. Section II describes the proposed joint loss for guiding the optimization of CNN models when the input dataset contain missing annotations. Section III describes the conducted experiments and analyzes the results. Finally, Section IV concludes the letter with some remarks and hints at plausible future research lines.

II. Proposed Approach

A. Notations

Let $\mathcal{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_N\}$ denote a building extraction dataset consisting of N images with binary masks, and $\mathcal{Y} = \{\mathbf{Y}_1, \cdots, \mathbf{Y}_N\}$ be the associated set of binary masks, where each element is either 0 or 1, i.e., $y_{ij} \in \{0, 1\}$. In this letter, we denote 1 as the building area and 0 as the background. $f(\cdot)$ represents the CNN model which maps the input image $\mathbf{X}_i$ to the predicted building mask $\mathbf{Y}'_i$.

B. The Proposed Joint Loss Function

1) logit adjusted cross entropy (LACE): When building annotations are missing, an imbalanced or long-tailed label distribution exhibits. As an illustrative example, we can randomly select 30%, 50% and 90% buildings from the well-known Massachusetts Buildings Dataset [17] and flip the associated building labels to background labels. Then, we calculate the pixel percentages of the two classes. As shown in Figure 2(a), as the number of missing building annotations increases, the long-tailed label distribution becomes more obvious. Therefore, the CE loss, a conventional loss used for training CNN models, will simply guide CNN models to classify every pixel with the majority label, i.e., background.

However, such models cannot generalize well in the testing phase. To cope with this issue, we adopt the LACE [18]:

$$L_{\text{LACE}} = -\sum_{ij} \log\Big(\frac{\exp\left(f_{y_{ij}}(x_{ij}) + \tau\log(\pi_y)\right)}{\sum_{y'_{ij}\in\{0,1\}}\exp\left(f_{y'_{ij}}(x_{ij}) + \tau\log(\pi_{y'})\right)}\Big) = \sum_{ij}\log\Big(1 + \sum_{y'_{ij}\neq y_{ij}}(\frac{\pi_{y'}}{\pi_y})^{\tau}\exp\left(f_{y'_{ij}}(x_{ij}) - f_{y_{ij}}(x_{ij})\right)\Big) \tag{1}$$

where τ denotes a temperature parameter and π_y is an estimate of the class prior $P(y)$, e.g., an empirical class frequency on the training set. Basically, LACE tries to set a pairwise label margin $\log\left(\frac{\pi_{y'}}{\pi_y}\right)^{\tau}$ between the predicting scores for y and y' as shown in Figure 2(b). Intuitively, by setting a margin between the predicted scores of rare and dominant classes, the trained CNN models are optimized in a way that the scores of the rare class are not overwhelmed by those from the dominant class. Thus, in the presence of a long-tailed label distribution, the learned classifier based on LACE can avoid scenarios in which most samples are categorized into the dominant class, achieving better precision scores. Note that, in this work, such dominant class corresponds to the image background (to relieve the requirement of a complete set of building footprint annotations.

2) Weighted Dice: In addition to the commonly utilized CE loss for binary segmentation problems, another region-based loss called Dice loss is also adopted. As opposed to the CE loss, which aims at optimizing the precision scores, minimizing the Dice loss is designed to increase the F_1 score:

$$L_{\text{Dice}} = 1 - \sum_{i}\frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{2}$$

where TP, FN and FP respectively denote true positives, false negatives and false positives on the predicted mask given the ground-truth. Considering the fact that we are dealing with a long-tailed label distribution, we utilize the weighted Dice loss, wherein the class-wise Dice losses are averaged in a weighted manner as follows:

$$L_{\text{WDice}} = \sum_{y}\pi_y L^{y}_{\text{Dice}}. \tag{3}$$

3) Boundary Alignment: Accurate building boundary generation is very important for footprint extraction. However, the two losses above are region-based and cannot penalize the boundary misalignment. In order to align the predicted building boundaries with the ground-truth, the boundary loss proposed in [15] is also adopted:

$$L_{\text{BD}} = \sum_{y}(1 - \frac{2\text{P}^y\text{R}^y}{\text{P}^y + \text{R}^y}) \tag{4}$$

where P^y and R^y denote the precision and recall scores of the boundary pixels with class y. To this end, when the benchmark datasets contain missing annotations, the proposed joint loss function for building footprint extraction is formulated as:

$$L_{\text{BD\_LACE\_WDice}} = L_{\text{BD}} + L_{\text{WDice}} + L_{\text{LACE}} \tag{5}$$

C. CNN model

In this letter, the standard U-Net[19] architecture is exploited for the building footprint extraction. U-Net fuses multi-level feature maps to simultaneously capture hierarchical semantics and preserve fine-grained shapes of objects in the predicted masks. We choose U-Net as the CNN backbone since it has been widely adopted as a benchmark CNN model for binary segmentation problems. However, it is worth noting that other CNN models with different architectures can be also combined with the proposed loss function for building footprint extraction.

III. Experiments

A. Experimental Setup

We evaluate the proposed loss function on two building segmentation datasets including: 1) Massachusetts Buildings dataset [17], and 2) ISPRS Potsdam dataset[1]. The training and test set of the Massachusetts Buildings dataset are the same as in the original paper. For the ISPRS Potsdam dataset, we select 2_13, 6_15 and 7_13 sets for testing and the others for training. To create the training datasets with missing annotations, we randomly select 30%, 50% and 90% of buildings in each image and flip the associated labels from building ($y = 1$) to background ($y = 0$). In order to feed the images into a graphical processing unit (GPU) with limited memory, we create image-mask pairs with the sizes of 300×300 pixels for the Massachusetts Buildings dataset and and 600×600 pixels for the ISPRS Potsdam dataset. In this way, the complete building footprint extraction can be conducted by sequentially processing each sub-image. The stochastic gradient descent (SGD) optimizer with initial learning rate set to 0.002 is adopted for minimizing the loss. The learning rate is decayed by 0.5 every 30 epochs. The parameters τ, $\pi_{y=1}$ and $\pi_{y=0}$ are set to 1, 0.9 and 0.1, respectively. For validation, we compare the proposed loss with other commonly exploited losses for binary segmentation, including: 1) CE [6], [11]; 2) Dice; 3) NR-Dice [20]; 4) CE_Dice[21]; 5) ELL [22]; 6) Weighted Dice (WDice); 7) BD_Dice [15]; and 8) BD_WDice. All the experiments are performed on an NVIDIA Tesla P100 GPU.

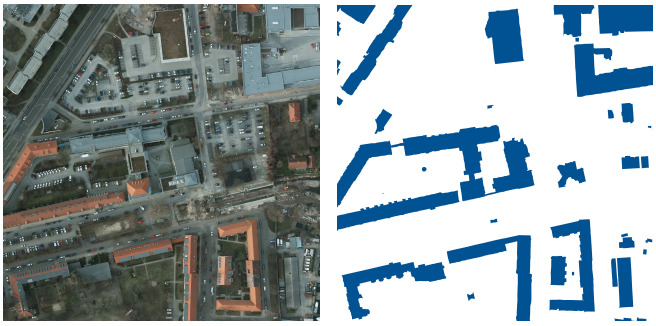
B. Experimental Results

1) Comparison with State-of-the-art Methods: In our evaluation, we compute the averaged Dice and IoU scores ($\%$) on the test sets based on the predicted masks, where the Dice score is defined as $2\text{TP}/(2\text{TP} + \text{FN} + \text{FP})$ and the IoU score as $\text{TP}/(2\text{TP} + \text{FN} + \text{FP})$. Table I displays the scores of these two metrics obtained by the U-Net trained on all the considered losses when there are 30%, 50% and 90% missing annotations for the buildings in the training sets. Compared with the other losses, the proposed one achieves the best performance on both datasets. For example, when there are 50% missing annotations in the ISPRS Potsdam buildings, an improvement of around 3% and 4% can be obtained by our BD_LACE_WDice loss (in terms of Dice and IoU scores)

[1] https://bit.ly/38rD6vG

TABLE I
DICE AND IOU METRICS (%) EVALUATED ON THE TWO BENCHMARK DATASETS WITH 30%, 50% AND 90% MISSING BUILDING ANNOTATIONS IN THE TRAINING SETS.

	Massachusetts						Potsdam					
	30%		50%		90%		30%		50%		90%	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
CE	66.44	49.93	6.37	3.33	–	–	82.05	70.19	60.61	43.61	–	–
Dice	79.72	66.40	76.25	61.81	40.48	25.93	88.26	78.99	78.99	65.92	–	–
NR-Dice	77.24	63.04	71.35	55.56	–	–	71.00	55.69	71.30	55.43	–	–
CE_Dice	77.09	62.87	71.35	55.56	–	–	80.41	67.64	66.26	49.95	1.61	0.81
ELL	79.31	65.84	73.46	58.27	–	–	85.00	74.09	77.65	63.47	–	–
WDice	80.38	67.32	78.50	64.80	48.63	32.62	88.52	79.41	85.68	75.09	72.09	56.94
BD_Dice	80.82	67.95	79.78	66.51	44.36	29.48	84.71	74.06	85.29	74.64	41.65	26.84
BD_WDice	80.68	67.73	78.88	65.24	52.24	35.96	83.53	72.05	85.43	74.64	–	–
LACE	74.91	60.01	75.37	60.60	9.22	4.86	88.33	79.14	86.73	76.59	37.80	23.95
LACE_WDice	80.17	67.00	78.89	65.24	57.23	40.35	88.30	79.11	85.75	75.14	57.03	40.97
BD_LACE_WDice	**81.19**	**68.45**	**80.06**	**66.85**	**57.55**	**40.92**	**88.72**	**79.93**	**88.21**	**79.07**	**73.95**	**59.05**

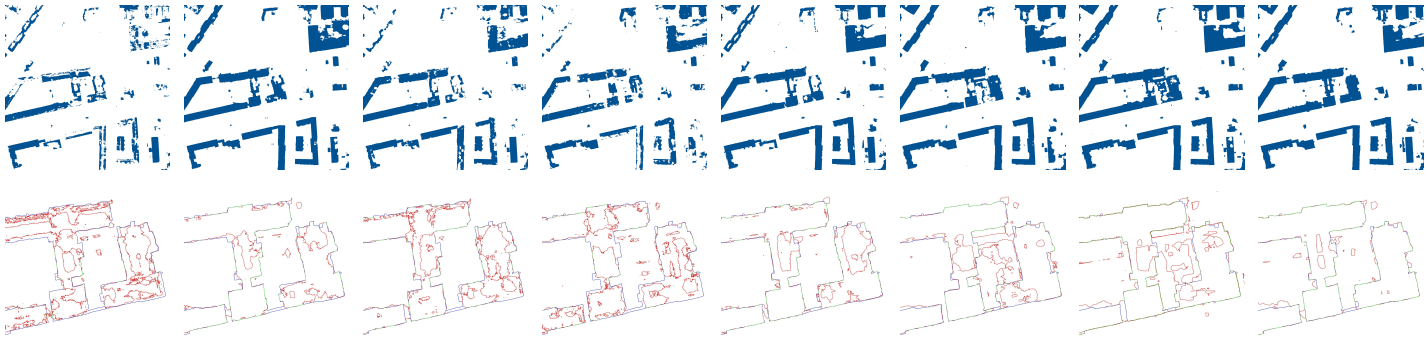


(a) (b)

Fig. 3. One test image from the ISPRS Potsdam dataset with its ground-truth building mask exploited for comparison among all the considered losses (a) ISPRS Potsdam image, (b) Ground-truth building mask.

with regards to other losses, e.g., BD_Dice or WDice. Although both NR-Dice and WDice are designed for the tackling segmentation problem with unbalanced label distribution, one plausible reason that explains why our joint loss outperforms those is the fact that it considers the penalization of boundary misalignments. In addition, the boundary loss is integrated in BD_Dice, while the long-tail label distribution affects its classification performance. When most of the buildings are not annotated (i.e., 90% case), it can be observed that several losses cannot successfully guide the trained UNets to predict the building regions on the test sets (noted with – in the table). Since almost 96% pixels are annotated as background [shown in Figure 2(a)], the correct classification contributes most to the training loss, so that the network just learns to produce the associated label. With the introduction of LACE, more emphasis can be made on the predictions of the building label. For a qualitative comparison, we select one test image from the ISPRS Potsdam dataset and utilize the trained U-Net through the compared losses and the proposed one to predict its building mask. Figure 3 displays the ground-truth building mask from the ISPRS Potsdam dataset. The top row of Figure 4 shows the predicted building masks based on the considered losses. The bottom row displays the associated building boundary comparison between the predicted masks and the ground-truth mask, where red, green and blue colors denote the FPs, TPs and FNs, respectively. It is important to note that the corresponding building boundaries have been generated by means of the Canny edge detector. From the results, it can be observed that some building areas cannot be correctly predicted based on the standard losses, such as CE and Dice, since the missing labels can confuse the discrimination of the U-Net on building and background pixels. Based on the proposed loss, the involvement of LACE can enforce the U-Net to emphasize more the labeled buildings, so that the building areas cannot be misclassified as background. Therefore, our BD_LACE_WDice achieves prominent segmentation performance, despite the fact that large amounts of buildings are unlabeled in the training datasets.

2) Ablation Study: For the ablation study, we conduct the same experiments as in the previous subsection, where LACE and LACE_WDice are both exploited as the losses to train the U-Net. The results are presented in Table I. For the Massachusetts Buildings dataset, without WDice and boundary terms, LACE only cannot achieve comparable performance as LACE_WDice and BD_LACE_WDice, while the segmentation results of the three losses are slightly different on the ISPRS Potsdam dataset. Since the spatial resolution of the Massachusetts Buildings dataset is about 1m^2 per pixel, many buildings just occupy only a few pixels, so that distinguishing between building instances is important for the metric evaluation. With the integration of the boundary loss, the building boundaries are penalized more than the other regions. Thus, BD_LACE_WDice can better recognize small building instances, achieving the best performance. For the ISPRS Potsdam dataset, the spatial resolution is about 5cm^2 per pixel. The fine-grained structure of the building regions is important for the metric evaluation. When large amounts of buildings are not annotated, the conventional losses (such as CE) cannot easily discriminate between background and building pixels. Therefore, the three losses (LACE, LACE_WDice and BD_LACE_WDice) outperform the conventional losses in the ISPRS Potsdam dataset, being their results comparable.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4. ISPRS Potsdam dataset: the top row displays the predicted building masks and the bottom row displays the generated building boundaries of one area, where red, green and blue colors denote the FP, TP and FN, respectively: (a) CE; (b) Dice; (c) NR-Dice; (d) CE_Dice; (e) ELL; (f) WDice; (g) BD_Dice; (h) BD_LACE_WDice. (Better to view zoom-in.)

IV. Conclusions

This letter presents a new joint loss function for extracting footprints of buildings using deep learning technology, under the assumption that there are many buildings that are not annotated. In order to solve this problem, we first investigate the label distribution when there are missing annotations at different levels. Then, we formulate the problem as a long-tailed classification one, and propose a joint loss function including: 1) LACE; 2) weighted Dice; and 3) boundary alignment loss to optimize the CNN model and better predict region and boundary pixels. Based on two building segmentation benchmark datasets, we validate the proposed loss function compared with other state-of-the-art approaches, and achieve the best performance when 30%, 50% and 90% buildings are missing in the training sets. The proposed joint loss function can be applied with any CNN architecture for binary segmentation problems when there are missing annotations. Robust deep learning techniques [23] considering missing annotations as noisy labels will be adopted in the future.

References

[1] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, p. 407, 2018.

[2] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and lidar data using coupled cnns," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, 2020.

[3] R. Huang, Y. Xu, D. Hong, W. Yao, P. Ghamisi, and U. Stilla, "Deep point embedding for urban classification using als point clouds: A new perspective from local to global," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 62–81, 2020.

[4] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[5] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[6] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, 2017.

[7] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, 2019.

[8] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, 2019.

[9] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sens.*, vol. 10, no. 9, p. 1350, 2018.

[10] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf)," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[11] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[12] Q. Chen, L. Wang, S. L. Waslander, and X. Liu, "An end-to-end shape modeling framework for vectorized building outline generation from aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 114–126, 2020.

[13] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, 2017.

[14] G. Wu, Z. Guo, X. Shi, Q. Chen, Y. Xu, R. Shibasaki, and X. Shao, "A boundary regulated network for accurate roof segmentation and outline extraction," *Remote Sens.*, vol. 10, no. 8, p. 1195, 2018.

[15] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," in *International Symposium on Neural Networks*. Springer, 2019, pp. 388–401.

[16] Y. Xu, W. Yao, L. Hoegner, and U. Stilla, "Segmentation of building roofs from airborne lidar point clouds using robust voxel-based region growing," *Remote Sensing Letters*, vol. 8, no. 11, pp. 1062–1071, 2017.

[17] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto, 2013.

[18] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," *arXiv preprint arXiv:2007.07314*, 2020.

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

[20] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, "A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images," *IEEE Trans. Med. Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[21] Y. Lin, D. Xu, N. Wang, Z. Shi, and Q. Chen, "Road extraction from very-high-resolution remote sensing images via a nested se-deeplab model," *Remote Sens.*, vol. 12, no. 18, p. 2985, 2020.

[22] K. C. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood, "3d segmentation with exponential logarithmic loss for highly unbalanced object sizes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 612–619.

[23] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. J. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Trans. Geosci. Remote Sens.*, 2020.