HF-FCN: Hierarchically Fused Fully Convolutional Network for Robust Building Extraction

Tongchun Zuo, Juntao Feng, Xuejin Chen*
Department of Electronic Engineering and Information Science,
University of Science and Technology of China

ACCV '16

Introduction

➢ Goal

- Propose a robust and efficient system for building extraction in remote sensing image.

➢ Background

- High resolution satellite images are easily accessible.
- Building rooftop is essential for a wide range of technologies, such as, urban planning, automated map making, 3D city modelling, disaster assessment, military reconnaissance.

➢ Challenges

- Extract arbitrary-size buildings with largely variant appearances or occlusions with high efficiency.

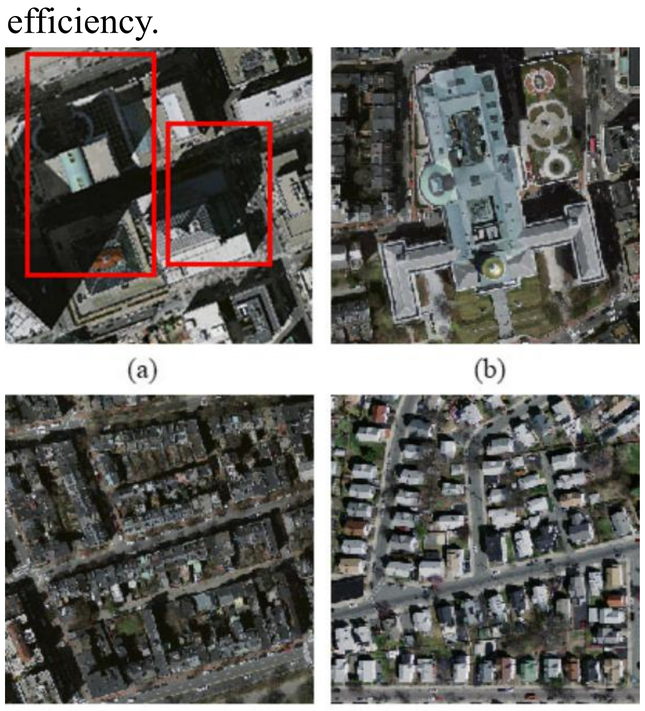


Fig. 1. Examples of aerial images with different type of challenges.
(a) Occlusions in red boxes. (b) Variant appearances. (c) Low contrast. (d) A large number of tiny buildings.

➢ Our ideas

- Design a new scheme to effectively integrate multi-level semantic information and spatial information.
- Deploy a neural network architecture to process arbitrary images with much complex scenes than training data.

➢ Our contributions

- A new architecture is developed for building extraction, which has a strong ability in processing appearance variations, varying building sizes and occlusions. The overall accuracy exceeds the state-of-the-art algorithms.
- Our approach leads to a notable reduction of computation cost compared with previous solutions.

Our Algorithm

➢ Network architecture

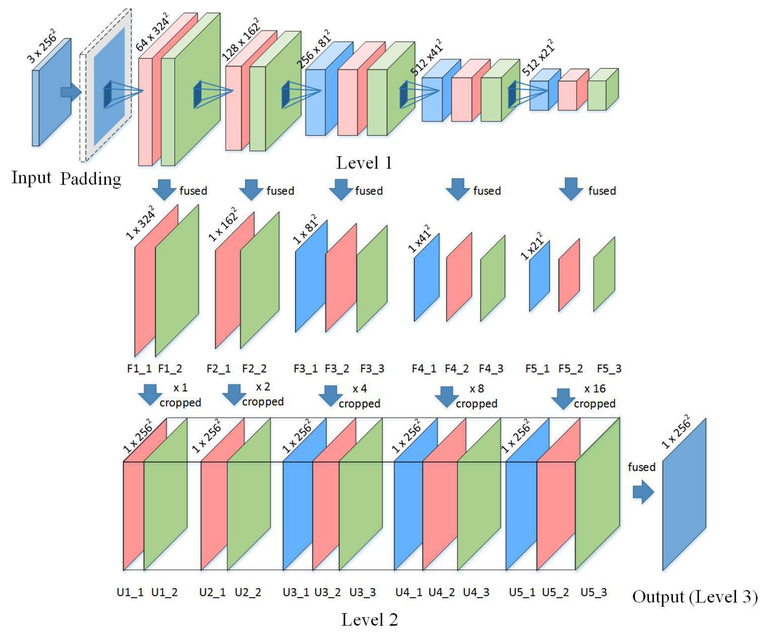


Figure 2. Our network architecture. F1_1 means the fusion of feature maps generated from its corresponding convolutional layer conv1_1, U1 _1 means the upsampling of F1_1, and so forth.

➢ Loss function

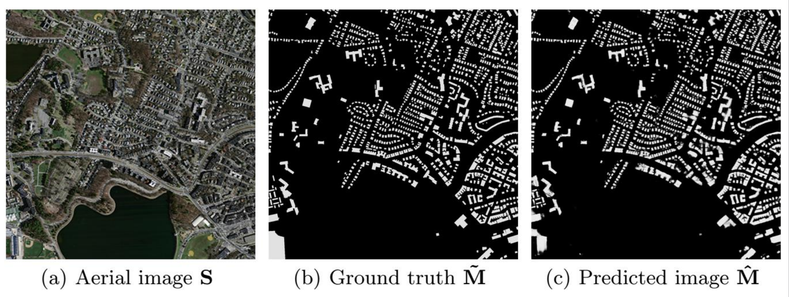


Figure 3. An example of the predicted image.

$$\mathcal{L} = -\frac{1}{|\mathbf{S}|}\sum_{s_j \in \mathbf{S}} [\tilde{m}_j \log \hat{m}_j + (1-\tilde{m}_j)\log(1-\hat{m}_j)]$$

➢ An example

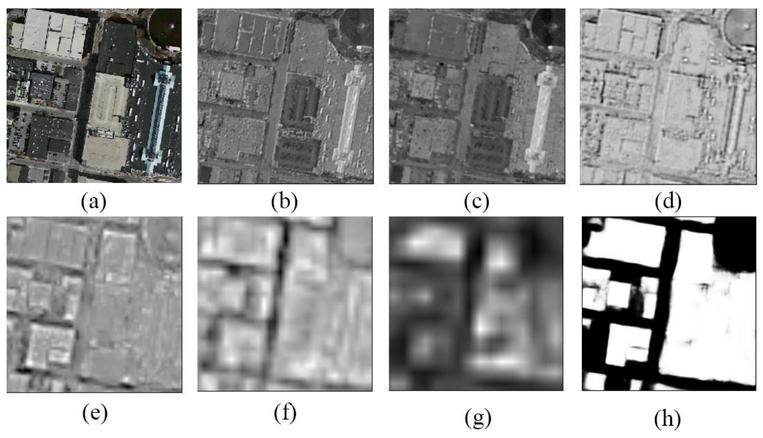


Figure 4. (a) Input aerial image. (b - g) Feature maps generated from U1_1, U1_2, U2_1, U3_3, U4_2, U5_2, respectively. (h) Predicted labelling map.

➢ Dataset

	Training	Validation	Testing
Number	75938	2500	10
Resolution	256*256	256*256	1500*1500

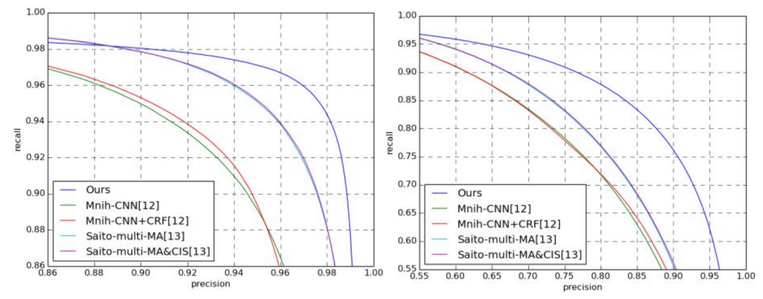
Table 1. The dataset is created from Massachusetts Buildings Dataset proposed publicly by Minh [12].

➢ Training strategies

- HF-FCN is fine-tuned from an initialization with the pre-trained VGG16 Net model and trained in an end-to-end manner.

Our Results

➢ Whole performance



(a) Slack parameter ρ = 3 (b) Slack parameter ρ = 0

Figure 5. The relaxed precision-recall curves from different methods with two slack parameters.

Table 2. Performance comparison with [12, 13]. Recall here means recall at breakeven points. Time is computed in the same computer with a single NVIDIA Titan 12GB GPU.

	Recall ($\rho = 3$)	Recall ($\rho = 0$)	Time (s)
Mnih-CNN [12]	0.9271	0.7661	8.70
Mnih-CNN+CRF [12]	0.9282	0.7638	26.60
Saito-multi-MA [13]	0.9503	0.7873	67.72
Saito-multi-MA&CIS [13]	0.9509	0.7872	67.84
Ours (HF-FCN)	**0.9643**	**0.8424**	**1.07**

Table 3. Recall at the selected regions of the test images.

Image ID	01	02	03	04	05	06	07	mean
Mnih-CNN+CRF [12]	0.784	0.869	0.769	0.653	0.893	0.764	0.800	0.784
Saito-multi-MA&CIS [13]	0.773	0.915	0.857	0.789	0.945	0.773	0.830	0.851
Ours (HF-FCN)	**0.874**	**0.964**	**0.899**	**0.901**	**0.986**	**0.840**	**0.851**	**0.911**

➢ Challenging instances' performance

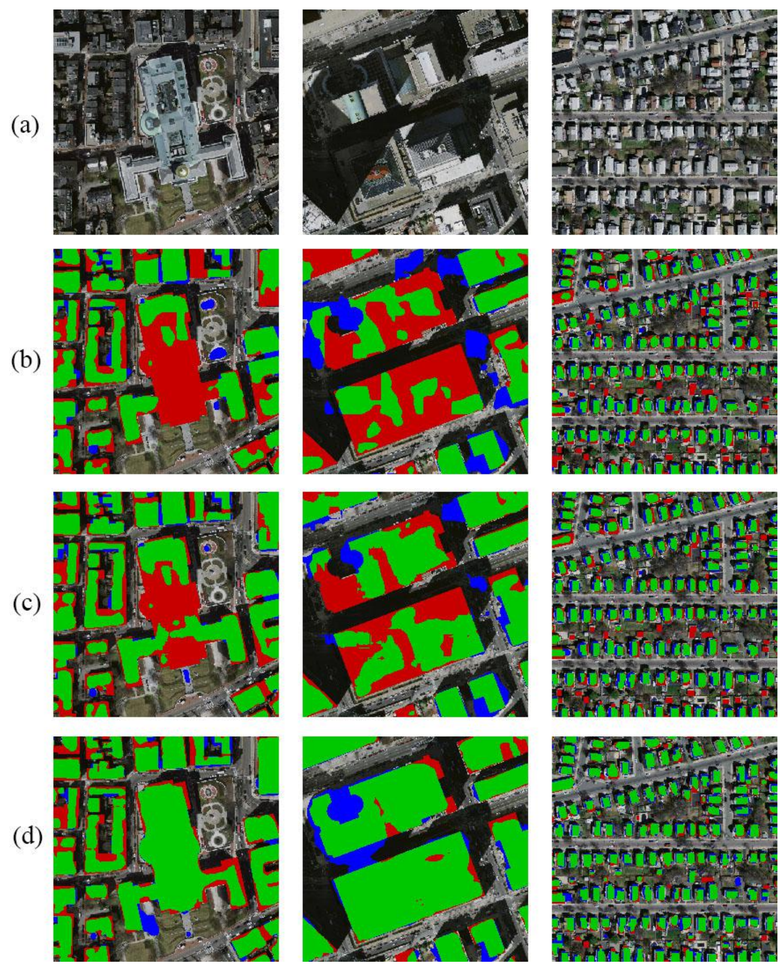


Figure 6. (a) Input images. (b) Results of Mnih-CNN+CRF [12]. (c) Results of Saito-multi-MA&CIS [13]. (d) Our results. Correct results (TP) are shown in green, false positives (FP) are shown in blue, and false negatives (FN) are shown in red.