Dual-HRNet for Building Localization and Damage Classification

Jamyoung Koo
SI Analytics, Republic of Korea
jmkoo@si-analytics.ai

Junghoon Seo
Satrec Initiative, Republic of Korea
sjh@satreci.com

Kwangjin Yoon
SI Analytics, Republic of Korea
yoon@si-analytics.ai

Taegyun Jeon
SI Analytics, Republic of Korea
tgjeon@si-analytics.ai

Abstract

Satellites imagery has assisted the humanitarian assistance and disaster relief (HADR) efforts when a disaster strikes. To speed up analysis of overhead imagery it is important to develop such an automated system. In this study, we propose Dual-HRNets for the task of both localizing buildings and classifying their damage level on the satellite imagery. Our network, termed Dual-HRNets, consists of two HRNets with some modifications. Since the goal of this study is to predict the location of buildings on the pre-disaster image as well as classify the damage level in accordance with the post-disaster image, our Dual-HRNets take a pair of pre- and post-disaster images over each HRNet. Each intermediate stage of Dual-HRNet, the features of each HRNet are fused by a fusion block. Our experiments present a good performance of Dual-HRNets for the task of both localization and classification. Our approach with Dual-HRNet achieved the 5th place in xView2 challenge among more than 3,5000 competitors. The code is avaiable at https://github.com/SIAnalytics/dual-hrnet

1 Introduction

Most recently satellites have assisted the humanitarian assistance and disaster recovery (HADR) efforts when a disaster strikes. However, covering the wide devastated areas across the satellites imagery by human observers requires a huge effort and time. Hence, automated system that can analyze the overhead images to assess the damages has currently received a lot of attention, such as computer vision, machine learning and deep learning.

In this paper, we proposed a deep neural network, which is based on HRNet [1, 2], for the task of localizing buildings and scoring the degree of damage to the building before and after disasters. Images taken from the satellite have a large variance since buildings, land use patterns, and disasters likely to occur are different around the world. Therefore, it is important for a model to predict accurate damage assessment using a pair of before and after disaster images, because images taken under normal circumstances represent the specific local context in the area of interest. In late 2019, the xView2 challenge, held by the Defence Innovation Unit, has started and focused on automating the process of assessing building damage after a natural disaster (wildfire, landslides, volcanic eruptions, earthquakes, tsunamis and flooding) using a pair of pre- and post-disaster images. This challenge has encouraged researchers to develop computer vision algorithms that can accurately analyze overhead imagery by localizing and categorizing various types of building damage. For this challenge, researchers have to predicts not only the location of building using the pre-disaster image but also the category of building damage assessment against the post-disaster image. For the category of damage assessment, Joint Damage Scale (JDS) is used to provide a consistent unified assessment

scale for building damage in satellite imagery across multiple hazard types, structure categories, and geographical locations. In Table 1, the description of each JDS category is presented. Thus, an algorithm is required to predict the JDS label for the damage classification task.

Table 1: Description of Joint Damage Scale (JDS)

Score	Label	Visual description of the structure
0	No damage	Undisturbed. No sign of water, structural damage, shingle damage, or burn marks
1	Minor damage	Building partially burnt, water surrounding the sturcture, volcanic flow nearby, roof elements missing, or visible cracks.
2	Major damage	Partial wall or roof collapse, encroaching volcanic flow, or the structure is surrounded by water or mud.
3	Destroyed	Structure is scorched, completely collapsed, partially or completely covered with water or mud, or no longer present.

For the task of both localization and classification, we proposed a deep neural network termed Dual-HRNets, which consists of two HRNets [1, 2] with some modifications. By taking advantage of HRNet that is known for the high performance in semantic segmentation task, the proposed Dual-HRNets segments buildings from images with their proper damage level, *i.e.,* JDS. A brief description of Dual-HRNets is depicted in 1 and detailed explanation of our method is given in the section 2.

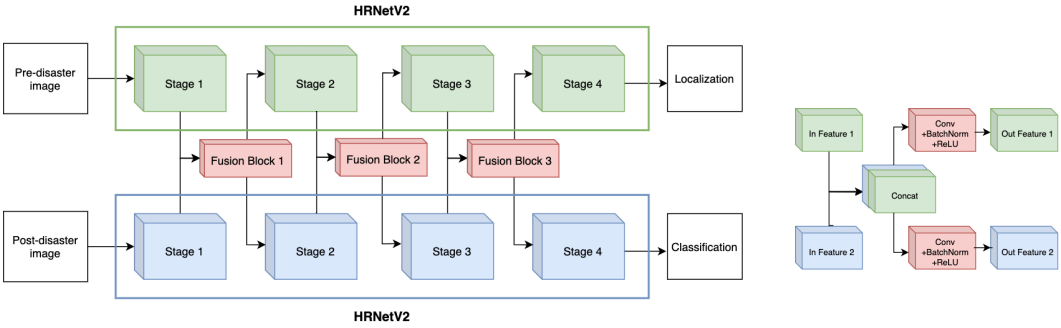
We take the Dual-HRNet to participate in the xView2 challenge and ranked the fifth place in the challenge.

2 Approaches

Problem Formulation Let training dataset and testing dataset be $D_{train} = \{(X_i^{train}, Y_i^{train})\}_{i=1}^{N_{train}}$ and $D_{test} = \{(X_i^{test}, Y_i^{test})\}_{i=1}^{N_{test}}$, respectively. Each input data X_i is a pair of image data (X_i^{pre}, X_i^{post}). $X_i^{pre} \in \mathbb{R}^{H\times W\times 3}$ is an optical satellite image with an RGB channel taken before the disaster, while $X_i^{post} \in \mathbb{R}^{H\times W\times 3}$ is taken after the disaster. X_i^{pre} and X_i^{post} are taken of the same area, both of which have spatial dimensions H and W. Each target data Y_i is a pair of label data of semantic segmentation $(\vec{y^*}_i^{loc}, \vec{y^*}_i^{cls})$. Each label in localization label data $\vec{y^*}_i^{loc} \in \{0,1\}^{H\times W}$ is assigned 0 if the joint damage scale of the data is 0, and 1 otherwise. Each label in classification label data $\vec{y^*}_i^{cls} \in \{1,2,3\}^{H\times W}$ is assigned joint damage scale itself if the joint damage scale of the data is not zero. If the joint damage label is 0, the label is marked *ignore* label. In reality, X_i^{pre} and X_i^{post} cannot be perfectly aligned due to numerical errors and physical limitations of satellite systems. Considering this alignment error, all Y_i are labeled to be aligned with X_i^{pre}.

Let the neural network parameterized with θ be denoted as $f_\theta : \mathbb{R}^{H\times W\times 3} \times \mathbb{R}^{H\times W\times 3} \mapsto \{0,1\}^{H\times W} \times \{1,2,3\}^{H\times W}$. f_θ receives a pair of image X_i, and produces both localization prediction $\vec{\tilde{y}}_i^{loc}$ and classification prediction $\vec{\tilde{y}}_i^{cls}$ i.e. $f_\theta(X_i) = (\vec{\tilde{y}}_i^{loc}, \vec{\tilde{y}}_i^{cls})$. In this work, our ultimate goal is to find out a good f_θ that makes less difference between Y_i^{test} and $f_\theta(X_i^{test})$ for all $i \in \{1, 2, \cdots, N_{test}\}$ using D_{train}.

Network Architectures We exploit HRNet[2, 3] for localization and classification. HRNet maintains high-resolution representations by connecting high-to-low resolution convolutions in parallel and repeatedly conducting multi-scale fusions across parallel convolutions. HRNet achieved high performance of semantic segmentation by high-resolution representations and spatially precise. Our network consists of two HRNets and fusion block which is shown in Figure 1. The network is an end-to-end trainable and inference. Both a pre-disaster image and a post-disaster image are feed into the network then it conducts localization and classification at ones. We fuse two features of each HRNet by a simple fusion block. The fusion block consists of concatenation and two convolution block(conv+batchNorm+ReLU). We tried more complicated fusion block but it didn't work well.



(a) Dual-HRNet (b) Fusion Block

Figure 1: Proposed Dual-HRNet is consists of (a) two HRNets and (b) fusion block for localization and classification.

Training Objectives One of golden standards on choice of training objectives (a.k.a loss functions) for semantic segmentation is cross-entropy loss [4, 5, 6]. It is well known that the standard cross-entropy is vulnerable to extreme class imbalance [7]. Thus, in the early stage of our work, we tried to use weighted cross-entropy [8, 9, 10, 7] with various class weighting strategies. However, our Dual-HRNets with none of them seemed to correctly converge in training phase. We infer that this observation derived from the fact that xBD dataset used in xView2 challenge has much more serious class imbalance [11] than the other ordinary semantic segmentation datasets e.g. Pascal VOC [12] and COCO [13]. Therefore, this difficulty needs to be taken into account when dealing with xBD datasets rather than other datasets.

To solve this problem, we borrow Lovasz-softmax loss [14] as surrogate loss for maximizing Jaccard index J of class c:

$$J_c(\vec{y^*}, \vec{\tilde{y}}) = \frac{|\{\vec{y^*} = c\} \cap \{\vec{\tilde{y}} = c\}|}{|\{\vec{y^*} = c\} \cup \{\vec{\tilde{y}} = c\}|}, \tag{1}$$

given ground truth labels $\vec{y^*}$ and predicted labels $\vec{\tilde{y}}$. The induced training objective L_{IoU} is defined as $L_{IoU}(\vec{y^*}, \vec{\tilde{y}}) = 1 - J_c(\vec{y^*}, \vec{\tilde{y}})$. Using submodularity of the set function L_{IoU}, Lovasz-softmax loss $\hat{L}$ was designed as a differentiable surrogate function of L_{IoU}. See the original work [14] for more details on its mathematical motivation and implementation.

In our case, the most attractive factor about Lovasz-softmax loss is that we can treat class-wise IoU as loss function, so we can train the model with the same importance between classes regardless of the heavy-tailed nature of specific class. It is noteworthy that Lovasz-softmax loss was originally proposed as a surrogate loss of dataset-IoU (or image-IoU), not to solve class imbalance problems. To best of our knowledge, this is the first work to use Lovasz-softmax loss in aspect of solving class imbalance problem.

Our final objective function L is the weighted sum of the two lovasz-softmax losses: the localization loss and the classification loss:

$$L = \hat{L}(\vec{y^*}^{loc}, \vec{\tilde{y}}^{loc}) + \alpha * \hat{L}(\vec{y^*}^{cls}, \vec{\tilde{y}}^{cls}). \tag{2}$$

where α is a hyper-parameter deciding loss weight between localization loss and classification loss. Note that the classification loss $\hat{L}(\vec{y^*}^{cls}, \vec{\tilde{y}}^{cls})$ is not computed where label of $\vec{y^*}^{cls}$ is *ignore* label.

3 Experiments

Details of Training Phase We train on an input resolution of 512 x 512 by random cropping. To reduce overfitting, we adopt standard data augmentation including random horizontal flipping, random vertical flipping, random rotation and random scaling (between 0.8 to 1.2). We use the SGD optimizer with the base learning rate of 0.01, the momentum of 0.9 and the weight decay of 0.0005. The poly learning rate policy with the power of 0.9 is used for dropping the learning rate for 250

epochs. The model trained with a batch-size of 32 (on 16 GPUs) with syncBN. Two HRNetW32(32 indicates the width of the high-resolution convolution) are initialized with ImageNet pretrain and fusion block are randomly initialized.

Details of Inference Phase We infer on a original input resolution of 1024 x 1024 to make use whole semantic information. There is no test augmentation and no model ensemble. To keep only classification of building, we set classification output in background by localization output.

Evaluation Our model was evaluated by official xView2 evaluation server. In Table 2, We achieve 86.570%, 72.577% for localization and classification F1 score respectively. Combined F1 score for ranking metric of xView2 is 76.775%(30% localization F1 score + 70% damage classification F1 score). Finally, we achieve 77.862% on the hold-out dataset and ranked the 5th place in the challenge. We have also shown some qualitative results in Figure 2. The first and second row show competitive results in both small and large building. The third row show inconsistent result in partial damaged building because our model can't distinguish individual buildings.

Table 2: Our Results on xView2 Challenge

Dataset Type	F1 score(%)	Localization(%)	Classification(%)
Public	76.775	86.570	72.57
Hold-out	77.862		

4 Conclusion

In this paper, we present Dual-HRNet for task of both localization and classification using before and after disasters satellite imagery. The fusion for each HRNet is simple but achieve competitive performance. We ranked 5th place in xView2 Asses Building Damage challenge. In future work, we seek to instance semantic segmentation for consistent damage score of partial damaged building.

References

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[2] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *CoRR*, vol. abs/1908.07919, 2019.

[3] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv preprint arXiv:1904.04514*, 2019.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

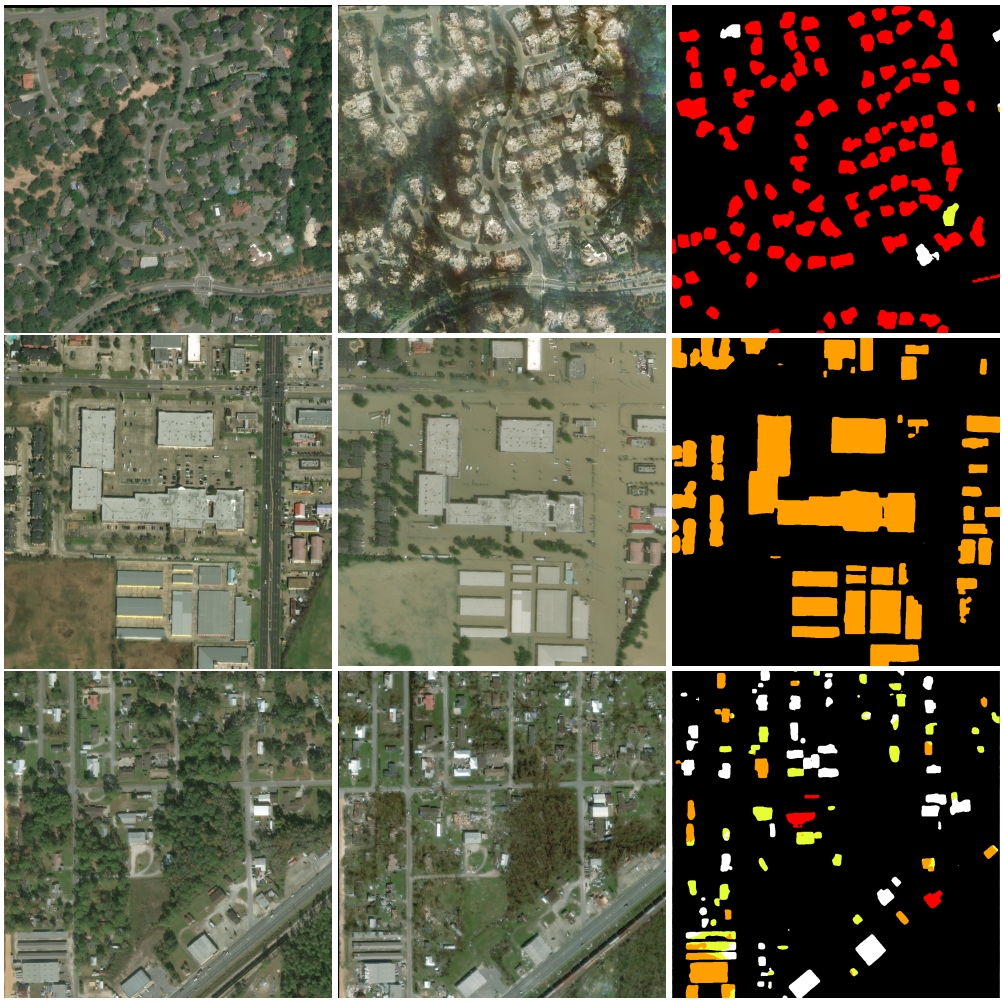
[6] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 11–19, 2017.

[7] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

[8] 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*, pp. 234–241, Springer, 2015.

[9] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

[10] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi, *et al.*, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 415–423, Springer, 2016.



(a) Pre-disaster images (b) Post-diaster images (c) results of our approache

Figure 2: **Qualitative results of Dual-HRNet.** White, yellow, orange and red regions represent no damage, minor damage, major damage and destroyed respectively.

[11] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, "Creating xbd: A dataset for assessing building damage from satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 10–17, 2019.

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[14] M. Berman, A. Rannen 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4413–4421, 2018.