This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Self-Mutating Network for Domain Adaptive Segmentation of Aerial Images

Kyungsu Lee
DGIST
Korea Republic of
ks_lee@dgist.ac.kr

Haeyun Lee
DGIST
Korea Republic of
haeyun@dgist.ac.kr

Jae Youn Hwang*
DGIST
Korea Republic of
jyhwang@dgist.ac.kr

Abstract

The domain-adaptive semantic segmentation of aerial images using a deep-learning technique is still challenging owing to the domain gaps between aerial images obtained in different areas. Currently, various convolutional neural network (CNN)-based domain adaptation methods have been developed to decrease the domain gaps. However, they still show poor performance for object segmentation when they are applied to images from other domains. In this paper, we propose a novel CNN-based self-mutating network (SMN), which can adaptively adjust the parameter values of convolutional filters as a response to the domain of an input image for better domain-adaptive segmentation. For the SMN, the parameter mutation technique was devised for adaptively changing parameters, and a parameter fluctuation technique was developed to randomly convulse the parameters. By adopting the parameter mutation and fluctuation, adaptive self-changing and fine-tuning of parameters can be realized for images from different domains, resulting in better prediction in domain-adaptive segmentation. Meanwhile, the results of the ablation study indicate that the SMN provided 11.19% higher Intersection over Union values than other state-of-the-art methods, demonstrating its potential for the domain-adaptive segmentation of aerial images.

1. Introduction

Aerial imagery has been widely utilized for urban planning, autonomous vehicles, and digital map generation in the field of remote sensing. In particular, building segmentation is crucial for digital map generation using aerial imagery. To segment buildings from aerial images, many convolutional neural networks (CNNs) [18]-based segmentation methods have been developed so far [26, 34, 37]. CNN-based segmentation methods outperform classical segmentation methods in the segmentation of buildings from aerial images. However, in general, aerial imagery has a variety of domains depending on time, countries, and aviation providers. Aerial images from various domains typically exhibit different resolutions, locations, and styles (Fig. 1). Although CNNs have shown powerful capabilities in a wide range of fields, they exhibit severe performance degradation when CNNs are applied to unobserved scenes and objects in other domains for various applications [7, 11]. That is, when a network trained with an aerial image of one main is applied to segment buildings from an aerial image of another domain, severe performance degradation of the network has been observed [5, 24, 38]. Therefore, domain adaptation (DA) has recently gained attention as an unmet need technique for resolving the limitations of CNNs.

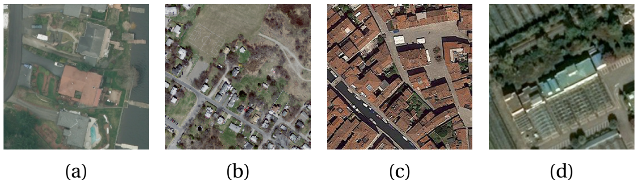


Figure 1: Sample images in different domain: (a) Inria Dataset, (b) Massachusetts Dataset, (c) WHU Dataset, and (d) Our Urban Dataset

To address this problem, a few deep learning techniques have been developed, such as transfer learning [24, 30, 38] and Generative Adversarial Networks (GANs) [10]. The transfer learning technique that reuses deep learning networks optimized in one domain has been introduced for many applications [24, 30, 38]. Despite the improved performance of the transfer learning technique for a building segmentation task, the transfer learning for the DA of aerial images did not show sufficient performance owing to a domain gap. Meanwhile, GANs have gained great attention as a novel technique that can be applied to DA [5, 36]. By using the GAN, style transfer for DA was realized. In addition, GAN-based layers were developed to decrease the domain gap during training [2, 8, 9]. However, because the images translated by GAN are not perfectly recognized as an image in the same domain, generalized feature extractions cannot be fully achieved [31]. That is, because the domain gap still ex-

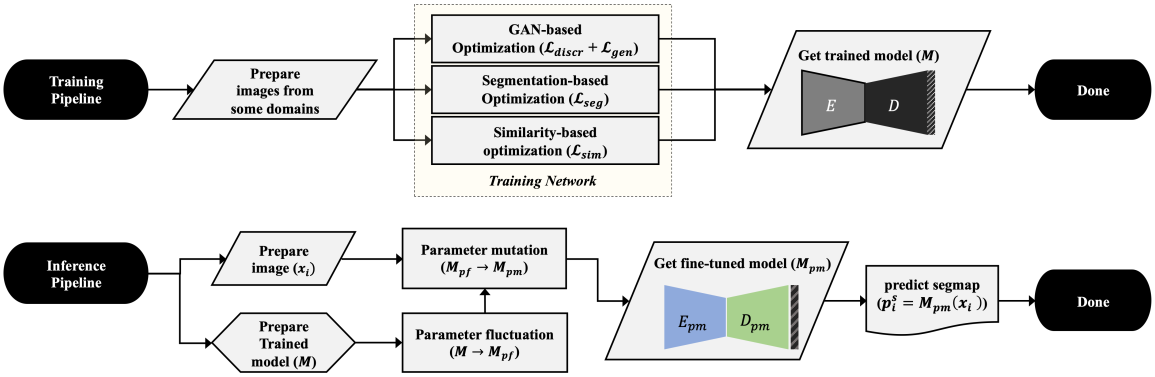

Training Pipeline
Prepare images from some domains
GAN-based Optimization ($\mathcal{L}_{discr} + \mathcal{L}_{gen}$)
Segmentation-based Optimization ($\mathcal{L}_{seg}$)
Similarity-based optimization ($\mathcal{L}_{sim}$)
Training Network
Get trained model (M)
E
D
Done
Inference Pipeline
Prepare image (x_i)
Prepare Trained model (M)
Parameter mutation ($M_{pf} \to M_{pm}$)
Parameter fluctuation ($M \to M_{pf}$)
Get fine-tuned model (M_{pm})
E_{pm}
D_{pm}
predict segmap ($p_i^s = M_{pm}(x_i)$)
Done

Figure 2: Flowchart of a self-mutating network. Encoder (E) and Decoder (D) of the network are optimized by three paths of loss functions simultaneously. In the inference step, the optimized network is fine-tuned by two steps of (1) parameter fluctuation and (2) parameter mutation.

ists after the DA, the performance of DA still needs to be improved to reduce the domain gap.

In this paper, we demonstrate a new GAN-based parameter self-mutating network to improve the building segmentation of aerial images from different domains. Here, rather than developing a new method to reduce the domain gap from different domains, we propose a network that is capable of converting the inherent properties of the network itself as a response to the domain of an input image. Therefore, we designed a GAN-based network called a self-mutating network (SMN). The SMN adjusts the parameter values of the convolution filters themselves as a response to the domain of an input image in the prediction step using two novel techniques: (1) *parameter mutation*, in which the parameters of a network are changed as the domain of inputs alongside the adaptation process of GAN, and (2) *parameter fluctuation*, in which the parameters of a convolutional network are finely oscillated to add randomness, resulting in an increase in the entropy of the network. Fig. 2 shows the flowchart of the SMN algorithm. The main contributions of this study are summarized as follows:

- We developed a novel GAN-based self-mutating network for the precise domain-adaptive semantic segmentation of aerial images from different domains. The network fine-tunes the parameters of the model on every testing image using two techniques:

- *Parameter mutation*: A fine-tuning technique for parameters of convolutional filters. GAN-based optimization within the prediction step adjusts the values of the parameters in response to the domain of an input aerial image.

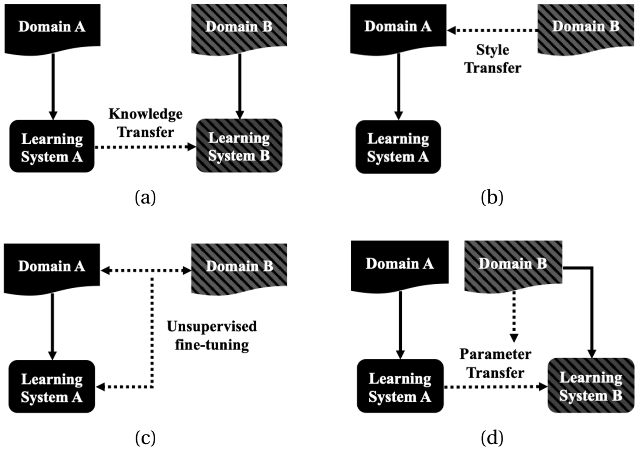
- *Parameter fluctuation*: A technique of adding randomness to the parameters of convolutional filters. While parameters are randomly vibrating, and the entropy of the network increases.

2. Related Work

The following works related to our approach are reviewed; DA and segmentation task. Fig. 3 describes the general and our DA methods.

2.1. Domain Adaptation (DA)

In the early work of GAN [10], it was simply applied for the various purposes of achieving enough number of datasets [21], extracting various features from aerial images [4], and applying attentions to the baseline networks



(a) (b)

(c) (d)

Figure 3: Different methodologies of Domain Adaptation. (a) Transfer learning that reuses the pre-trained networks for the other domains; (b) Style transfer method that translates the styles of the source domain to the target domain; (c) Unsupervised fine-tuning method that fine-tunes the architecture using source and target domain; (d) Parameter adaptation that translates parameter according to the domain. In this paper, a self-mutating network is developed based on (d) parameter adaptation, and fine-tunes itself according to the input domain.

[27]. Recently, by using GAN, the style transfer that the images in one domain are translated into another target domain has been developed [2, 8]. Moreover, GAN-based layers have been developed to help decrease domain gaps while training models [9, 29], as well as the purposes of data generation and gathering attention [27]. Dundar *et al.* matched the synthetic baseline of computer graphics images for DA. Y. Ganin *et al.* conducted literature studies on DA and applied them to diverse domains. In addition, Y. Yanchao *et al.* utilized the Fourier transform while translating the domains of images as DA [39].

2.2. Segmentation of buildings in aerial images

From the early studies on the semantic segmentation of buildings in aerial images, classical computer vision-based algorithms utilizing the color, shape, and boundaries of buildings have been developed to segment buildings in aerial images [14, 16, 25]. However, they have frailties that they cannot be applied to general aerial images with poor performance. From the first deep neural network [20] for a segmentation task, many deep learning networks have been developed in these fields. Ivanovsky *et al.* applied a simple CNN-based network in which the architecture was based on the encoder–decoder architecture [3] in the field of semantic segmentation of objects in aerial images [13]. In addition, an advanced architecture with modular architectures, called pyramid pooling layers, is applied to specify the exact shapes and locations of buildings [17, 41]. Furthermore, attention-based CNN architectures have been proposed to achieve an accurate segmentation map [27]. Wang *et al.* [32] introduced non-local blocks to achieve non-locality aerial images.

2.3. Domain-adaptive aerial semantic segmentation

With the drastic increase in the number of aerial images and datasets of aerial images, the transfer learning of domain adaptation has been developed to utilize the deep learning network, which is optimized in one domain into other domains. For transfer learning, a novel framework of the GAN has been introduced recently [5, 6, 19, 24, 40, 30]. Benjdira *et al.* studied cross-domain segmentation by changing the images from the source domain to the target domain, including resolution, captured image sensors, and captured locations of aerial images [5]. Furthermore, Benjdira *et al.* applied a GAN-based network for unsupervised semantic segmentation of aerial images as a DA [6]. Li *et al.* evaluated the performance of the utilization of GAN in the field of transfer learning for aerial images and concluded that the GAN-based networks have an acceptable performance to be utilized in the field of aerial images [19]. Onur *et al.* generated the same image as an input using a GAN, except for the spectral distribution. Na *et al.* devised a segmentation network based on a domain-adaptive transfer attack scheme [24].

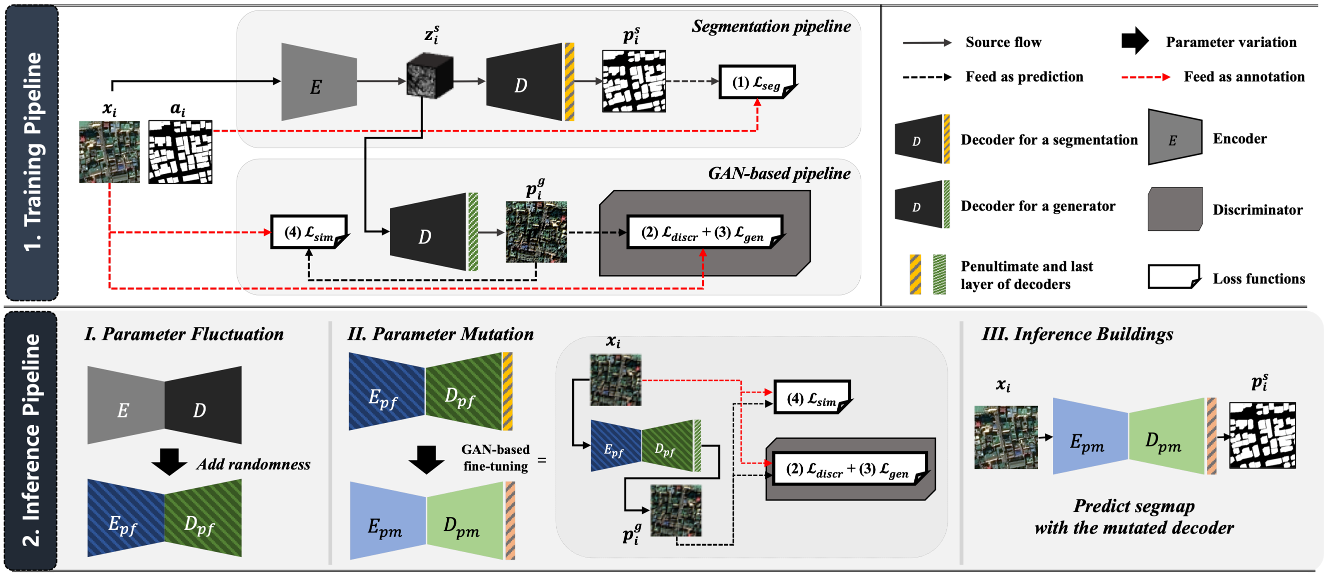


Figure 4: The pipeline of self-mutating network. In the training pipeline, losses are utilized to optimize overall CNN-based components including an encoder(E), a discriminator (S), and a decoder(D) for object segmentation and image generation. Each loss function is explained with the associated equations in the methods. In the inference pipeline, randomness is added to the parameters of E and D by parameter fluctuation. Parameter mutation is then performed while optimizing the losses of (2)∼(4) that are GAN-based optimizations. Finally, the inference of the buildings is performed with the mutated parameters of E and D, which are fine-tuned as the domain of an input image.

3. Methods

In this section, we provide an overview of our proposed network, that is, the SMN, and describe the novelty of an SMN in terms of parameter adaptation. The two major novelties applied are parameter fluctuation and parameter mutation. Fig. 4 illustrates the overall architecture of the SMN.

3.1. Architecture overview

The proposed architecture is ultimately configured to accept aerial images as input and generates a segmentation map (segmap) of the building objects as output. To this end, the proposed network consists of the encoder-decoder pipeline, which is mainly used in a segmentation task [3]. In addition to the basic baseline, the proposed network has an additional GAN-based structure [10] for the parameter mutation.

Fig. 5 illustrates the detailed structures of the proposed SMN. The SMN adopted encoder-decoder architecture, and in addition, a non-local block is embedded in the tail of the network to apply the non-locality [33]. The SMN includes four trainable CNN structures of an encoder, a decoder for a segmap, a decoder for a generator, and a discriminator. Note that the decoder for a segmap and the decoder for a generator share variables that have the same parameter values of the convolutional filters, except for the penultimate, and last layers. The encoder and decoder for segmentation are used to segment buildings, and the discriminator and decoder for the generator are utilized for the GAN-based optimization, which fine-tunes the parameters during the inference of the building. The details of each structure are provided in *supplemental material.*

3.2. Training phase of self-mutating network

To generate a segmap, a conventional encoder-decoder architecture of SegNet are utilized. They are here optimized with a cross-entropy loss as follows:

$$\nabla_{\theta_e,\theta_d} L_{seg} = \nabla_{\theta_e,\theta_d} \frac{1}{N} \sum_{i=1}^{N} \Big[\big(G\big) \log\big(D(E(x_i)\big) + \big(1 - G\big) \log\big(1 - D(E(x_i))\big) \Big] \tag{1}$$

where θ_e and θ_d are the trainable variables of the encoder and decoder for a segmap, respectively, x_i is an input image, and E, D, and G are an encoder, a decoder for a segmap, and the corresponding ground truth of x_i, respectively.

In addition, the proposed architecture has a pipeline that can fine-tune parameters through GAN as shown in the lower part of Fig. 5 similar to an auto-encoder. The following general loss functions of Eq. 2 and 3 for GAN are utilized to optimize the encoder, decoder for a generator, and discriminator.

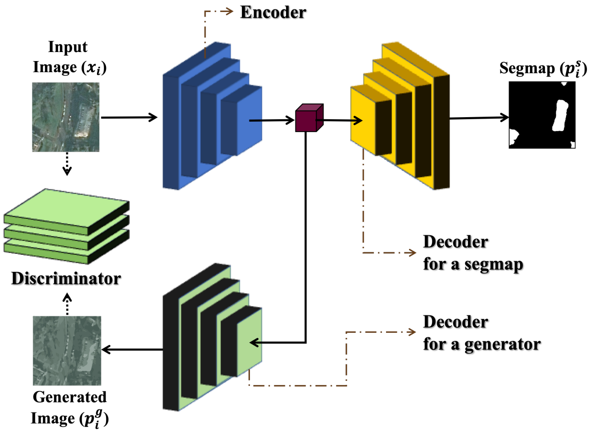


Figure 5: Detailed architecture and modules of self-mutating network. Here, the decoder for segmap and the decoder for the generator share variables except for the penultimate and the last layer.

$$\nabla_{\theta_s} L_{disc} = \nabla_{\theta_s} \frac{1}{N} \sum_{i=1}^{N} \Big[\log\big(S(x_i)\big) + log\big(1 - S(D(E(x_i)))\big) \Big] \tag{2}$$

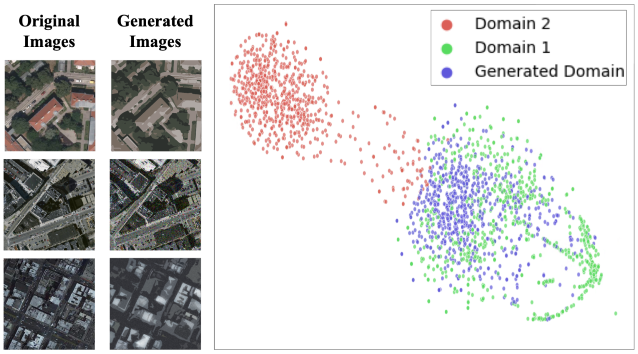
$$\nabla_{\theta_e,\theta_d} L_{gen} = \nabla_{\theta_e,\theta_d} \frac{1}{N} \sum_{i=1}^{N} \Big[log\big(1 - S(D(E(x_i)))\big) \Big] \tag{3}$$

where S indicates discriminator, and θ_s is the trainable variables of the discriminator. In general, when an image from a different domain, which has not been trained, inputs to the SMN, an image similar to the input image is not created. However, if parameters are finely adjusted through the GAN-based architecture for creating an image similar to an original image, the segmap can be accurately predicted for the image of a new domain. Therefore, the GAN-based structure was adopted here.

However, since the GAN-based optimization is slow to train, the loss function using structural similarity is applied to generate a similar image as follows:

$$L_{sim} = -\frac{1}{N} \sum_{i=1}^{N} \big[\mathrm{sim}\big(x_i, D(E(x_i))\big) \sum_{h,w,c}^{H,W,C} \big((D(E(x_i))) - x_i\big)^2\big] \tag{4}$$

where *sim* is the function to measure the structural similarity [35] between two images, and H, W, and C indicate the height, width, and channel of an image, respectively. When only the L_2 loss is used to generate an image that is exactly the same as an input image, the performance of the proposed network was degraded. Therefore, the weight factor of a structural similarity is added to lower the weights of L_2 loss, . When the generated image is similar to an input image, only the generator loss is then applied to the optimization of the network. The optimization process is illustrated in Fig. 4.



(a) (b)

Figure 6: (a) Sample images of generated images and (b) T-SNE plot of Domain 1, Domain 2, and a generated domain from Domain 1. The images in the generated domain can be recognized as Domain 1.

When the SMN is optimized, the encoder and decoder for a generator generate a similar image that is not completely identical but is in the same domain. Fig. 6(a) shows the examples of generated images through the generator including the encoder and decoder. Fig. 6 (b) illustrates the T-SNE [12] plot. The T-SNE plot demonstrates that the generated images belong to a domain that is almost identical to that of the input images.

3.3. Inference phase of the self-mutating network

As illustrated in Fig. 4, the prediction of buildings by SMN is carried out with the following steps: (1) Parameter fluctuation is applied to the encoder(E) and decoders(D), resulting in the fluctuated E and D. (2) The parameter mutation is then applied to fine-tune the fluctuated E and D while optimizing the GAN-based pipeline, and then, get the mutated E and D. (3) The prediction of a building is performed with the mutated E and D. When a raw image is given as input, the parameter vibrates through the parameter fluctuation method, and the values of the parameter change within a certain space range. Subsequently, the raw image is fed into the encoder and decoder for a generator, and a new image is generated through the compression and decompression processes. Simultaneously, the discriminator determines whether the generated image is similar to the raw input image. Here, as the generated image is optimized to be more similar to the input image, the parameter values of the encoder and decoder change. Subsequently, after finishing the changes in the parameter values, the raw image is fed into the structure of the encoder and decoder for a segmap, and the segmap of the building corresponding to the input image is predicted.

3.4. Parameter Fluctuation

Parameter fluctuation is a method of adding randomness to the parameters of E and D in the prediction step. Because of the gradient-vanishing problem, deeper layers cannot be mutated by parameter mutation. Therefore, parameter fluctuation is proposed to solve this problem.

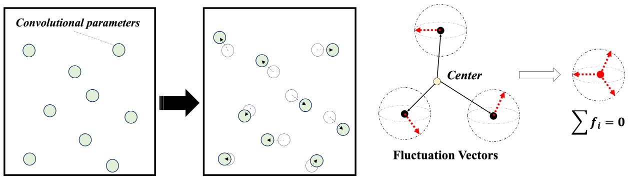
In the SMN, all parameters have a 3×3 size of convolution filters, and all parameters are mapped as vectors to variables in 9-dimensions. Subsequently, random vectors are added to each parameter to change the spatial position of the parameters. Here, the parameter vectors are denoted as v_i, the center point of v_i is defined as c, and the fluctuation vectors are defined as f_i. The following conditions should be satisfied, but the length of the fluctuation vectors does not exceed a constraint constant (λ_1):

$$\sum [f_i] = 0 \tag{5}$$

$$||f_i|| \leq \lambda_1 ||v_i - c|| \tag{6}$$

That is, the center of the vectors should not be changed even after adding random factors. The range of the random factors is set so as not to exceed λ_1% multiplied by the distance between each vector (v_i) and the center vector (c), thus ensuring that random elements do not exceed the constrained range of the optimized parameters. When the conditions in Eqs. 5 and 6 are met, a similar segmentation performance could be achieved in the same domain. The mathematical proof is given in the *Supplemental material.* The example of the parameter fluctuation is illustrated in Fig. 7. The sum of the all fluctuation vectors should be a zero-vector to maintain the center position of all parameters.

Algorithm.1 represents the algorithm of parameter fluctuation. Since the parameter fluctuation adds random elements to the parameters, the entropy of the network becomes increased, and the accuracy of segmentation from other domains become improved.



(a) (b)

Figure 7: (a) Schematic illustration of the parameter fluctuation. Randomness is added to the parameters of convolution filters. (b) Sum of fluctuation vectors. Here, the red dotted arrows represent a fluctuation vector for a convolutional parameter, and the sum of all fluctuation vectors should be a zero vector.

Algothim 1: The procedures and conditions of the parameter fluctuation.

```
Generates $4N$ vectors $(s_i)$ on the surface of randomly and uniformly distributed zero-mean Gaussian noise in 3D-coordinates.
    Here, the range $(r)$ of vectors should be less than $0.25\lambda_1 min(||v_i - c||)$
Select randomly one pair of vectors of $s_{r_{1,2}}$, add $\theta_{\text{rand}}$ and $\phi_{\text{rand}}$ to $s_{r_1}$, and add $-\theta_{\text{rand}}$ and $-\phi_{\text{rand}}$ to $s_{r_2}$, simultaneously          .
Select randomly 4 vectors of $s_{r_{1,2,3,4}}$ and add them as $f_i = s_{r_1} + s_{r_2} + s_{r_3} + s_{r_4}$ while preserving $\sum f_i = 0$
Add the fluctuation vector $(f_i)$ as $\hat{v}_i = v_i + f_i$
    Here, $\sum \hat{v}_i = 0$
return Fluctuated vectors $(\hat{v}_i)$
```

3.5. Parameter Mutation

In general, CNN-based networks that are optimized only for one domain cannot generate an acceptable performance when predicting images of another domain; therefore, a parameter mutation technique is proposed. The parameter mutation is a GAN-based fine-tuning technique that changes the parameters according to the input domain. As the input images are compressed and decompressed, new images are generated. However, if the encoder and decoder are not fully optimized in the source domain, the generated images are not sufficiently similar to the input images. Therefore, by fine-tuning the parameters of the encoder and decoder, similar images are generated with the input images during the prediction phase. Thus, the parameters of the SMN are fine-tuned according to the input domain, and the SMN can then generate a segmentation map of buildings with an improved performance.

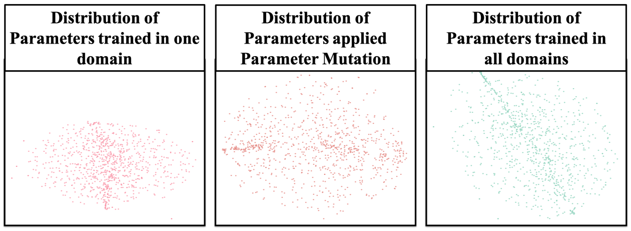


Figure 8: The T-SNE results of parameters of SMN optimized in one domain before (left) and after (middle) applying parameter mutation and optimized in all domains (right).

Note here that the parameters of the encoder and decoder are not fully optimized but are optimized while satisfying the following conditions:

$$\begin{aligned} &sim\big(x_i, D(E(x_i))\big) \geq \lambda_2 \\ &L_{sim} \leq \lambda_3 \end{aligned} \tag{7}$$

As the purpose of limiting the degree of optimization, (1) to overcome an overfitting problem because a fully optimized deep learning network does not guarantee a high segmentation performance and (2) to achieve a faster prediction time during the prediction phase, only a certain point of similarity is secured. Therefore, the parameter mutation is stopped if the conditions are satisfied. The parameter selection process is illustrated in *supplemental material*.

Fig. 8 illustrates the distributions of the parameters of the SMN (a) before and (b) after parameter mutation, and (c) the parameters trained to all domains in a 2D-space by applying the T-SNE method. Fig. 8 demonstrates that the parameter mutation can guarantee an acceptable performance in other domains, not limited to a particular domain of aerial images because the parameters are distributed to all domains as the parameters of the fully optimized network.

4. Experiments and Results

4.1. Dataset and Training Environment

To evaluate the performance of an SMN compared to other state-of-the-art networks, that is, FDA [39], DDA [5], DATA [24], and TreeUNet [40], which have been studied for domain adaptive semantic segmentation, a k-fold cross validation was applied to all datasets. In addition, the intersection over union (IoU) [28] of the buildings is utilized as the evaluation metric, and because the background IoU values do not impact the performance, only the IoU values of the buildings are utilized. The codes for the deep learning networks are implemented using the public library of Tensorflow version 1.13 [1], and the server for the testing is constructed with two Intel Xeon CPUs and four NVIDIA Titan-Xps GPUs.

To demonstrate the performance of the self-mutating network, four datasets, i.e., WHU [15], Inria [22], Massachusetts Buildings [23], and our Urban Dataset (OUD), were applied. Images in each domain have different characteristics of resolutions, locations, time, and architecture styles, as well as detailed information including differences, numbers of images, methods used to construct the datasets, and construction of the CNN models, as illustrated in the *supplemental material*. In addition, Fig. 9 illustrates that the T-SNE results after applying principal component analysis (PCA) and the characteristics of the images differ depending on the domains.

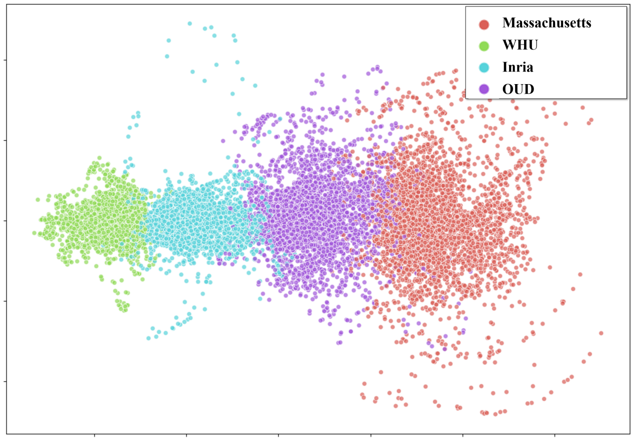


Figure 9: T-SNE results after applying PCA to aerial images of all domains including Massachusetts, WHU, Inria, and OUD datasets.

4.2. Experimental Results

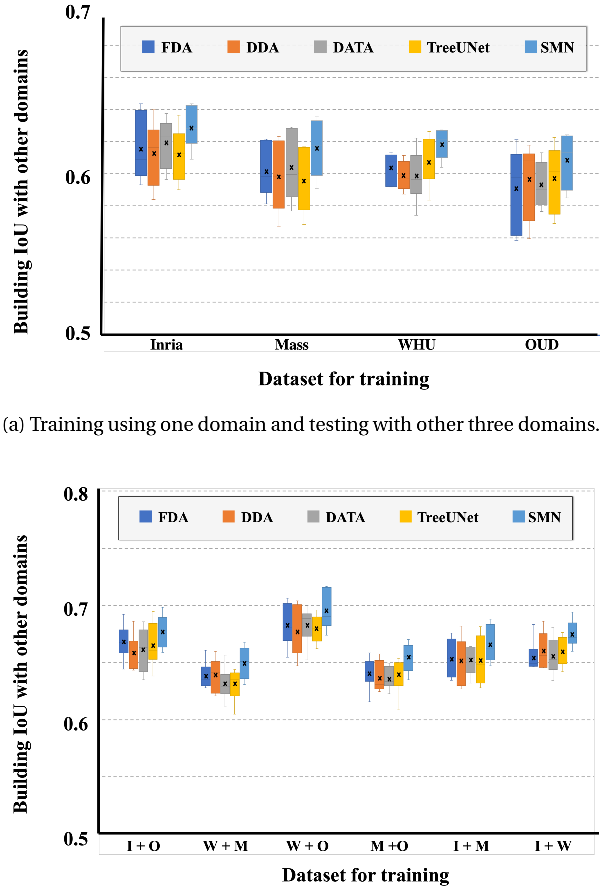
To evaluate the performance of SMN, we performed an ablation study of the SMN, compared the performance of SMN and state-of-the-art networks for the domain adaptive segmentation.

Table 1 shows the segmentation results of the ablation study of the self-mutating network. Here, I, M, W, and O indicate the Inria, Massachusetts, WHU, and OUD datasets, respectively. In addition, PM and PF indicate the parameter mutation and parameter fluctuation, respectively, and *SAME* indicates that only the l2 loss function is utilized in Eq. 4 instead of the structural weight factor. Note that the SAME generates exactly the same images as the input images, whereas Eq. 4 generates similar images that are completely different from the input images but are in the same domain.

In the ablation study (Table. 1), the GAN-based parameter mutation improved the segmentation performance, although the network predicted the images in different domains. Despite the poor performance of the arbitrary usage of the parameter fluctuation, the filter fluctuation helped improve the segmentation performance when it was utilized with the parameter mutation. Note that the case (PM) of generating similar but not the same images by applying a similar weight factor helped improve the accuracy compared to the case (SAME) of generating the same images through a GAN. Here, PM and PF seemed to generate better predictions by providing variations from the optimal point. However, since the improvement of performance of models with only PF or PM was not significant, it seems that the model is not at the optimal point but a local optimum. When PF and PM are used simultaneously, the performance of the model is significantly improved. So, it can be considered that the model with both PF and PM is at the optimal point with the fine-tuned parameters.

Table 1: Ablation study of the SMN. Here, the images of the illustrated domain are utilized as a trainset, and images in other three domains are utilized as a test set. For example, case-M indicates that the Massachusetts dataset is used as the training set and the other three domains are used as the test sets. The best performance marked as **bold** and the second as underline.

Structure	Building IoU			
	I	M	W	O
Baseline[3]	49.2	49.5	49.5	49.5
Baseline + *SAME*	51.0	50.8	50.5	49.7
Baseline + PF	52.9	52.4	52.1	51.2
Baseline + PM	52.8	54.0	51.2	54.9
Baseline + PF + *SAME*	52.2	51.6	53.2	52.5
Baseline + PF + PM (ours)	**62.9**	**61.6**	**61.9**	**60.9**



(a) Training using one domain and testing with other three domains.

(b) Training using two domains and testing with other two domains.

Figure 10: Segmentation results of SMN compared to other state-of-the-art networks. The illustrated dataset is used as the training set, and the other datasets of different domains are utilized as the test sets.

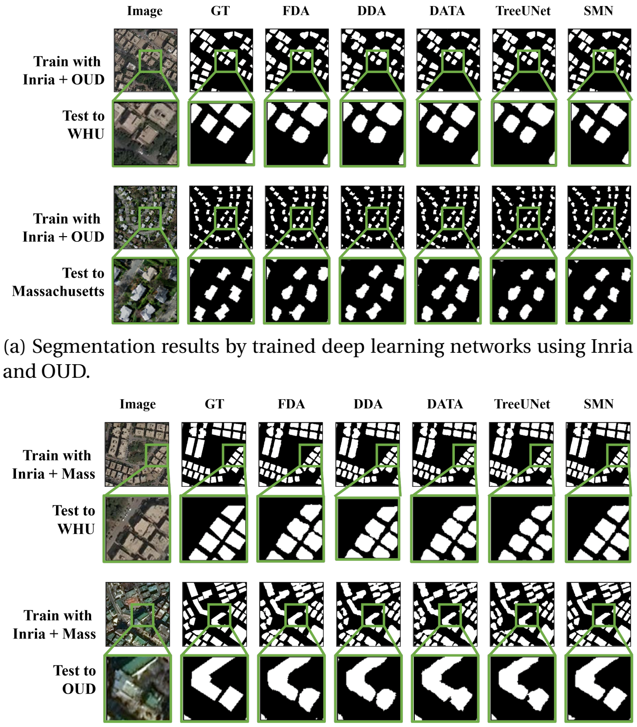
Table 2: Quantitative comparisons of different methods in terms of average of mean IoU. The first and the second best performance are in **bold** and underlined, respectively. The detailed quantitative comparisons are illustrated in the *supplemental material*.

Trainingset	FDA	DDA	DATA	TreeUNet	SMN
Inria	0.6158	0.6128	0.6193	0.6120	**0.6290**
Massachusetts	0.6018	0.5984	0.6038	0.5959	**0.6161**
WHU	0.6040	0.5986	0.5984	0.6075	**0.6190**
Ours	0.5913	0.5965	0.5930	0.5973	**0.6086**

Fig. 10(a) represents the IoU values of the predicted building by the FDA, DDA, DATA, TreeUNet, and SMN. The indicated dataset was utilized as a training set, and images in other domains were utilized as a test set. The SMN shows higher IoU values of 0.643, 0.635, 0.627, and 0.624 in the case of Inria, Mass, WHU, and OUD, respectively. Furthermore, the SMN shows the best IoU of 0.643, which is an 8.48% improvement over the others. Fig.10(b) represents the IoU values of predicted buildings of images from the other two domains by FDA, DDA, DATA, TreeUNet, and SMN trained with two domains. Here, the two datasets indicated were used as a training set, and images from the other two domains were used as a test set. The SMN shows the higher IoU values of 0.698, 0.668, 0.717, 0.670, 0.692, and 0.686 in the case of I+O, W+M, W+O, M+O, I+M, and I+W, respectively than other state-of-the-arts. Furthermore, the SMN shows the best IoU of 0.717, which is 11.19% higher than the others. As shown in Fig.10, the SMN shows the highest IoU values in every domain, and the mean IoU value of the SMN is higher than those of other state-of-the-art networks. Furthermore, as shown in Fig.11, the predicted segmaps of buildings of aerial images using SMN showed the most similar results to the ground truths in all tasks of aerial images.

5. Conclusions

In this paper, we proposed a novel deep learning architecture for the domain adaptive semantic segmentation of buildings in aerial images. The proposed network, denoted as a self-mutating network, changes the values of the trained parameters by using two novel approaches, that is, GAN-based parameter mutation and a mathematical methodology of the parameter fluctuation according to the domains of the input images. The experimental results demonstrate the feasibility that the proposed deep learning based on the vanilla network produces higher IoU values of buildings in aerial images, compared to other state-of-the-art models with 11.19% higher IoU values. In addition, to obtain much higher performance, we conducted an additional experiment using a state-of-the-art network-based SMN in the *supplemental material.* The main contribution of this study is the proposal of a novel utilization of a GAN within the prediction time as well as the mathematical parameter fluctuations. With the novel performance of the segmentation results for the DA of aerial images, the proposed self-mutating network can be used as a novel framework in this field. It is acceptable that the fine-tuning network in the prediction phase requires a large calculation time, as illustrated in the supplementary document, but it should be further improved.



(a) Segmentation results by trained deep learning networks using Inria and OUD.

(b) Segmentation results by trained deep learning networks using Inria and Mass.

Figure 11: Segmentation results corresponding to Fig. 10(b). The first two domains are utilized as the trainingset and the domain next to the arrow is utilized as the predictions. I, M, W, O indicate the Inria dataset, Massachusetts dataset , WHU dataset, Our Urban Dataset. The segmentation results related to Fig. 10(a) and Other combinations related to Fig. 10(b) are illustrated in the *supplemental material.*

6. Acknowledgement

This research was supported by "the 2021 Joint Research Project of Institutes of Science and Technology", "the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2020R1A2B5B01002786)", and the "Technology Innovation Program (1415173934) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea)". Also, Dabeeo Inc. partially supported this research. * indicates a corresponding author.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2810, 2018.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[4] Laila Bashmal, Yakoub Bazi, Haikel AlHichri, Mohamad M AlRahhal, Nassim Ammour, and Naif Alajlan. Siamese-gan: Learning invariant representations for aerial vehicle image categorization. *Remote Sensing*, 10(2):351, 2018.

[5] Bilel Benjdira, Adel Ammar, Anis Koubaa, and Kais Ouni. Data-efficient domain adaptation for semantic segmentation of aerial imagery using generative adversarial networks. *Applied Sciences*, 10(3):1092, 2020.

[6] Bilel Benjdira, Yakoub Bazi, Anis Koubaa, and Kais Ouni. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing*, 11(11):1369, 2019.

[7] JS Blundell and DW Opitz. Object recognition and feature extraction from imagery: The feature analyst® approach. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36(4):C42, 2006.

[8] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 2018.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Yiyou Guo, Jinsheng Ji, Xiankai Lu, Hong Huo, Tao Fang, and Deren Li. Global-local attention network for aerial scene classification. *IEEE Access*, 2019.

[12] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:857–864, 2002.

[13] Leonid Ivanovsky, Vladimir Khryashchev, Vladimir Pavlov, and Anna Ostrovskaya. Building detection on aerial images using u-net neural networks. In *2019 24th Conference of Open Innovations Association (FRUCT)*, pages 116–122. IEEE, 2019.

[14] Mohammad Izadi and Parvaneh Saeedi. Automatic building detection in aerial images using a hierarchical feature based image segmentation. In *2010 20th International Conference on Pattern Recognition*, pages 472–475. IEEE, 2010.

[15] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018.

[16] Konstantinos Karantzalos and Demertre Argialas. A region-based level set segmentation for automatic detection of man-made objects from aerial and satellite images. *Photogrammetric Engineering & Remote Sensing*, 75(6):667–677, 2009.

[17] Jun Hee Kim, Haeyun Lee, Seonghwan J Hong, Sewoong Kim, Juhum Park, Jae Youn Hwang, and Jihwan P Choi. Objects segmentation from high-resolution aerial images using u-net with pyramid pooling layers. *IEEE Geoscience and Remote Sensing Letters*, 16(1):115–119, 2018.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] Xue Li, Muying Luo, Shunping Ji, Li Zhang, and Meng Lu. Evaluating generative adversarial networks based image-level domain transfer for multi-source remote sensing image segmentation and object detection. *International Journal of Remote Sensing*, 41(19):7343–7367, 2020.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Dongao Ma, Ping Tang, and Lijun Zhao. Siftinggan: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro. *IEEE Geoscience and Remote Sensing Letters*, 16(7):1046–1050, 2019.

[22] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

[23] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.

[24] Younghwan Na, Jun Hee Kim, Kyungsu Lee, Juhum Park, Jae Youn Hwang, and Jihwan P Choi. Domain adaptive transfer attack-based segmentation networks for building extraction from aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[25] Ali Özgün Ok. Robust detection of buildings from a single color aerial image. *Proceedings of GEOBIA*, 6, 2008.

[26] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[27] Xuran Pan, Fan Yang, Lianru Gao, Zhengchao Chen, Bing Zhang, Hairui Fan, and Jinchang Ren. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. *Remote Sensing*, 11(8):917, 2019.

[28] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[29] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Residual parameter transfer for deep domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4339–4348, 2018.

[30] Onur Tasar, SL Happy, Yuliya Tarabalka, and Pierre Alliez. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[31] Hristina Uzunova, Jan Ehrhardt, and Heinz Handels. Memory-efficient gan-based domain translation of high resolution 3d medical images. *Computerized Medical Imaging and Graphics*, 86:101801, 2020.

[32] Shengsheng Wang, Xiaowei Hou, and Xin Zhao. Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. *IEEE Access*, 8:7313–7322, 2020.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2606–2615, 2017.

[35] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[36] Dennis Wittich and Franz Rottensteiner. Adversarial domain adaptation for the classification of aerial images and height data using convolutional neural networks. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2019.

[37] Guangming Wu, Xiaowei Shao, Zhiling Guo, Qi Chen, Wei Yuan, Xiaodan Shi, Yongwei Xu, and Ryosuke Shibasaki. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing*, 10(3):407, 2018.

[38] Michael Wurm, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, and Hannes Taubenböck. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, 150:59–69, 2019.

[39] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[40] Kai Yue, Lei Yang, Ruirui Li, Wei Hu, Fan Zhang, and Wei Li. Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 156:1–13, 2019.

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.