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

1. Details of Datasets

To demonstrate the performance of the Self-Mutating Network, four datasets including our Urban dataset (OUD), an Inria dataset [4], a WHU dataset [3], and a Massachusetts Building dataset [5] are utilized. Images in each domain have different characteristics in terms of resolutions, locations, time, and architecture styles.

1.1. Our Urban Dataset

9,000 images of various areas in the three cities of Seoul, Suwon, and Daegu, which are ones of the most complex cities in Korea, are constructed for the OUD dataset. The data consisted of RGB images with a resolution of 0.61 m and the corresponding ground truths for binary classes. The dataset covered an area of 147.46 km^2 and was divided into an area of 110.59 km^2 for training and an area of 36.87 km^2 for testing. The images had a pixel size of 256 × 256.

1.2. Inria Dataset

As the public data, an Inria dataset[4] was utilized in the experiments. The Inria dataset covered the area of 405 km^2. The aerial image had a spatial resolution of 0.3m and covers five cities such as Austin, Chicago, Kitsap, Tyrol, and Vienna. Here the image sets were randomly cropped into the size of 256 × 256 from the original size of 5000 × 5000, and a total of 144,000 images were utilized for each city.

1.3. WHU Building Dataset

From the WHU Building Dataset, we used one of Satellite Dataset I. The dataset was collected from cities over the world and from various remote sensing resources including QuickBird, Worldview series, IKONOS, ZY-3, etc [3]. The WHU dataset contains 204 images of which size is 512 × 512 but randomly cropped so that a total number of 20,400 images are constructed for the WHU dataset. Here, the resolutions of the WHU dataset varies from 0.3 m to 2.5 m.

1.4. Massachusetts Buildings Dataset

The Massachusetts Building Dataset consists of 151 aerial images of the Boston area. The size of the images is 1500 × 1500 pixels for an area of 2.25km^2, and thus the entire dataset covers roughly 340km^2. The data was split into a training set of 137 images, a test set of 10 images, and a validation set of 4 images [5]. Here, we randomly cropped the original images to the small images of which size was 256 × 256, and a total of 16,577 images were utilized to the Massachusetts dataset. In the Massachusetts Building Dataset, the target maps were obtained by rasterizing building footprints obtained from the OpenStreetMap project. The Massachusetts Building Dataset is a large amount of high-quality building footprint data. The dataset covers mostly urban and suburban areas, including buildings with various sizes, individual houses, and garages, are included in the labels [5].

2. Architecture description

The SMN has four trainable CNN-based architectures; an encoder, a discriminator, a decoder for a segmap, and a decoder for a generator. Table 1 illustrates the detailed layers of SMN. The architecture of the discriminator is the same with a previously designed discriminator [2]. The baseline of the SMN is devised from the SegNet [1] for the encoder and the decoder. Note that, the decoders for a segmap and a generator share the values (Upsampling1 - Upsampling4). However, only two layers of two decoders are different (See Logit in Table 1).

3. Hyper-parameter selection

In Eq. 6 and Eq. 7, hyper-parameter values (λ_1, λ_2, and λ_3) are utilized to determine a degree of the optimization of GAN in a prediction step. The hyper-parameter values affect the performance and prediction speed of SMN. Therefore, the values of hyper-parameters, which can offer high accuracy and fast prediction time, were here determined for SMN. As shown in Fig. 1, λ_1, λ_2, and λ_3 were determined to be 0.01, 0.87, and 0.40 to achieve high ac-

Table 1: The detailed architecture of a self-mutating network. SegNet is utilized as a baseline architecture for the SMN. The layers of decoders for a segmap and a generator are the same except for the last layers. A discriminator for GAN-based architecture is the same with that of a basic GAN [2]. Here, the Conv + BN + ReLU includes batch normalization (BN) and an activation function (ReLU).

	name	**image size**	**operation** encoder		**channel**
	Input image	256×256			3
Encoder	Down 1	256×256	3×3 Conv + BN + ReLU		64
		256×256	3×3 Conv + BN + ReLU		64
		128×128	Maxpooling		64
	Down 2	128×128	3×3 Conv + BN + ReLU		128
		128×128	3×3 Conv + BN + ReLU		128
		64×64	Maxpooling		128
	Down 3	64×64	3×3 Conv + BN + ReLU		256
		64×64	3×3 Conv + BN + ReLU		256
		64×64	3×3 Conv + BN + ReLU		256
		64×64	3×3 Conv + BN + ReLU		256
		32×32	Maxpooling		256
	Down 4	32×32	3×3 Conv + BN + ReLU		512
		32×32	3×3 Conv + BN + ReLU		512
		32×32	3×3 Conv + BN + ReLU		512
		32×32	3×3 Conv + BN + ReLU		512
		16×16	Maxpooling		512
Bridge	Bottom	16×16	3×3 Conv + BN + ReLU		512
		16×16	3×3 Conv + BN + ReLU		512
	name	**image size**	**operation** decoder for a segmap	decoder for a generator	**channel**
Decoder	Upsampling 1	32×32	Deconvolution + 3×3 Conv + BN		512
		32×32	3×3 Conv + BN + ReLU		512
		32×32	3×3 Conv + BN + ReLU		512
	Upsampling 2	64×64	Deconvolution + 3×3 Conv + BN		256
		64×64	3×3 Conv + BN + ReLU		256
		64×64	3×3 Conv + BN + ReLU		256
	Upsampling 3	128×128	Deconvolution + 3×3 Conv + BN		128
		128×128	3×3 Conv + BN + ReLU		128
		128×128	3×3 Conv + BN + ReLU		128
	Upsampling 4	256×256	Deconvolution + 3×3 Conv + BN		64
		256×256	3×3 Conv + BN + ReLU		64
		256×256	3×3 Conv + BN + ReLU		64
	Logit	256×256	3×3 Conv + BN + ReLU		64
		. 256×256	3×3 Conv	3×3 Conv + tanh	2/64
		256×256	SoftMax	3×3 Conv + tanh	2/3

curacy. With the restriction of Eq. 6 and 7, the prediction time by SMN was not significantly prolonged. It was only 20% longer than that by a conventional deep learning model. Despite the limitation of the optimization by parameter mutation, the fine-tuning of architecture occupies most of the prediction time. Even though the predic-

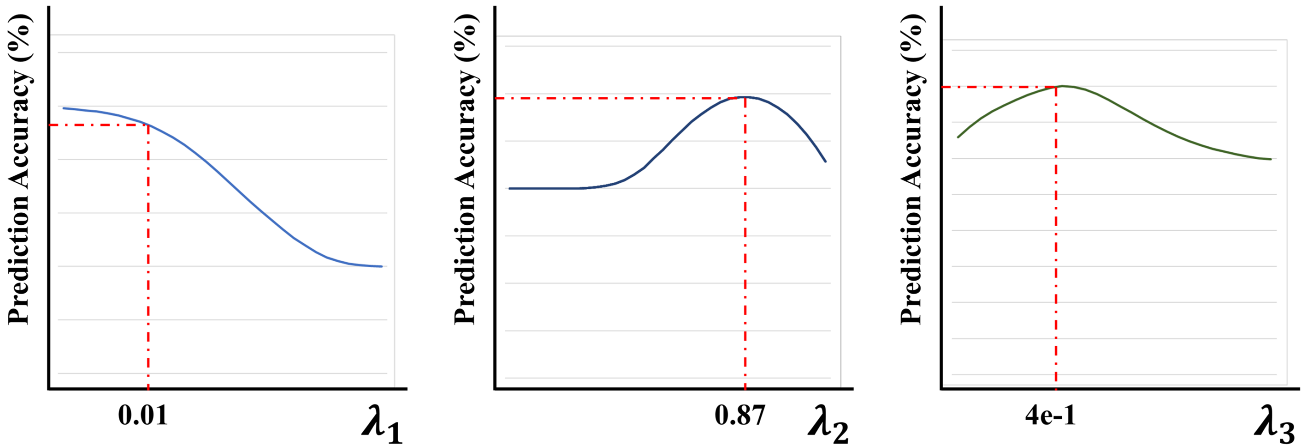


Figure 1: The prediction accuracy along the hyper-parameter values of SMN. The values of hyper-parameters, which generate accuracy within an appropriate range, were adopted to the SMN. The graphs have been ensemble-averaged.

tion time is not important in the field of an aerial image segmentation, the prediction time needs to be reduced. It remained as a future work.

4. Mathematical proof

Since the center of convolution parameters is the same after PF, the summation of convlution filters in the vector space is also the same. Therefore, $\sum p_i = \sum p_i$‘ where p_i are parameters before PF, and p‘$_i$ are parameters after PF. Therefore, PF is invariant to invertible linear transformation and it can be considered as Centered Kernel Alignment (Kornblith 2019). Thus, the similarity index remains after PF. In addition, while a fluctuation vector (f_i) is added to parameters, a feature-map (F) after PF-applied convolution (C) can be approximated as $F \circ C + |f_i F|$. Therefore, the variance of a generated feature-map remains, but the expectation is $\mathbb{E}(F \circ C) + |f_i F| \leqq \mathbb{E}(F \circ C) + \lambda_1 |F|$. That is, after PF, only a small value of λ_1 can guarantee a similar expectation value which leads to a similar normal distribution as well as a similar KL divergence. Furthermore, we investigated the structural similarity between predictions by the original optimized model and the fine-tuned model with PF. As shown in Fig. 2, a small value of λ_1 guarantees a similar prediction.

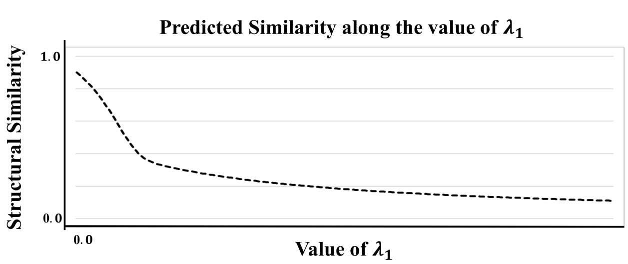


Figure 2: The similarity index between predictions by the original and fine-tuned models.

5. Details of Experiment

We compared the performance of SMN to that of other state-of-the-art models utilized in the domain adaptation of aerial images these days. As illustrated in the manuscript, Fig. 9 illustrates the comparison results of SMN and the others. Fig. 9(a) represents the IoU values of predicted buildings by FDA, DDA, DATA, TreeUNet, and SMN. Here, the illustrated dataset is utilized as a training set, and images in other domains are utilized as a test set. SMN exhibits the highest IoU values of 0.643, 0.635, 0.627, and 0.624 compared to other state-of-the art models in the case of Inria, Mass, WHU, and OUD, respectively. In particular, SMN offers an 8.48% higher IoU value than others. Fig. 9(b) represents the IoU values of predicted buildings by FDA, DDA, DATA, TreeUNet, and SMN. Here, two datasets are used as a training set whereas images in other two domains are used as a test set. SMN shows the highest 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. Furthermore, SMN offers the 11.19% higher IoU value than other models. The corresponding results are illustrated in Table 3 and Table 4. Furthermore, as illustrated in Fig. 9, SMN shows the highest IoU values in every domain, and its mean IoU values are also higher than those of other state-of-the-art networks. In addition, Fig. 3-8 show the predicted segmentation maps of buildings using SMN. It shows that SMN offer the clear boundaries of buildings in all domains of aerial images with high performance.

6. Additional Experiment

To search a novel deep learning network that has a higher performance of domain adaptation and segmentation using SMN, we carried out additional experiments using the state-of-the-art models(the 1st and 2nd ranked models) listed in CSAILVision and the baseline models introduced in the related papers. The BiseNet and DeepUNet are used as baseline models. The experimental results demonstrated that our method outperformed other DA methods, with a significant improvement in terms of IoU. The results demonstrate that our model shows the maximum and minimum improvements of 12.06% and 3.97%, which are huge improvements in domain adaptation and segmentation, compared to other models.

Table 2: Supplemental experiments using state-of-the-art and baseline models. Tree indicates TreeUNet, BL indicates a baseline model, and M.I indicates a maximum improvement (SMN - *min*(others)).

BiseNet	BS	FDA	DATA	Tree	SMN	M.I
I+O	0.552	0.648	0.610	0.656	0.731	12.06%
W+M	0.555	0.625	0.662	0.626	0.682	5.72%
W+O	0.550	0.658	0.617	0.640	0.684	6.74%
M+O	0.555	0.639	0.653	0.646	0.701	6.26%
I+M	0.552	0.620	0.643	0.660	0.686	6.57%
I+W	0.557	0.634	0.619	0.659	0.698	7.96%
I+O	0.539	0.596	0.606	0.597	0.636	3.97%
W+M	0.547	0.613	0.625	0.612	0.694	8.25%
W+O	0.555	0.650	0.645	0.629	0.701	7.16%
M+O	0.545	0.613	0.641	0.611	0.705	9.39%
I+M	0.551	0.620	0.606	0.638	0.663	5.68%
I+W	0.558	0.666	0.620	0.654	0.687	6.68%
DeepUNet	BS	FDA	DATA	Tree	SMN	M.I
I+O	0.562	0.650	0.631	0.627	0.682	5.54%
W+M	0.548	0.612	0.622	0.647	0.687	7.54%
W+O	0.559	0.619	0.641	0.663	0.712	9.28%
M+O	0.552	0.651	0.609	0.661	0.702	9.29%
I+M	0.550	0.656	0.617	0.650	0.702	8.45%
I+W	0.560	0.632	0.622	0.670	0.690	6.79%
HRNetV2	BS	FDA	DATA	Tree	SMN	M.I
I+O	0.591	0.654	0.658	0.662	0.756	10.19%
W+M	0.558	0.640	0.661	0.647	0.723	8.30%
W+O	0.554	0.640	0.621	0.637	0.695	7.36%
M+O	0.558	0.664	0.620	0.640	0.688	6.75%
I+M	0.557	0.669	0.643	0.644	0.688	4.50%
I+W	0.547	0.613	0.610	0.605	0.664	5.91%
EfficientNet	BS	FDA	DATA	Tree	SMN	M.I
I+O	0.596	0.686	0.674	0.705	0.746	7.21%
W+M	0.550	0.615	0.615	0.624	0.681	6.64%
W+O	0.553	0.623	0.661	0.629	0.708	8.49%
M+O	0.547	0.603	0.647	0.637	0.674	7.12%
I+M	0.559	0.643	0.615	0.635	0.679	6.42%
I+W	0.547	0.631	0.618	0.631	0.713	9.47%

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

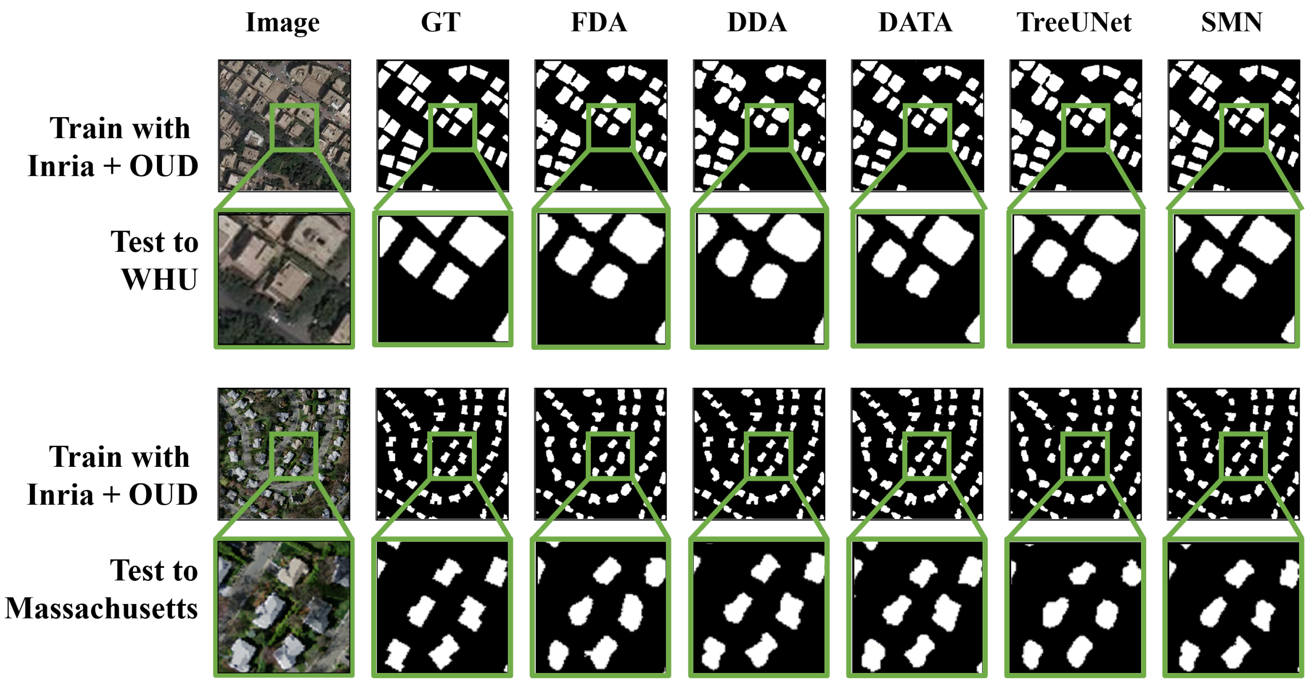


Figure 3: The predicted segmentation-map by deep learning models which are trained using Inria Dataset and OUD. The images in the second row are cropped and resized from the original images of the predicted segmentation-maps.

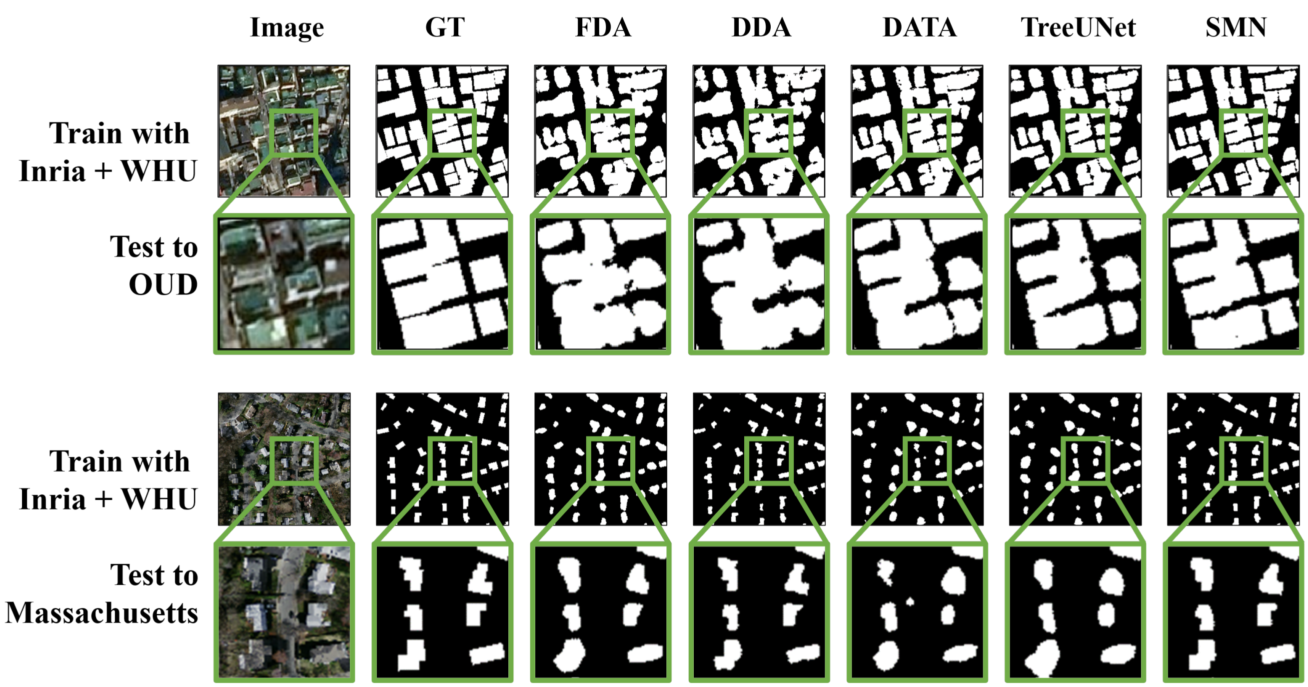


Figure 4: The predicted segmentation-map by deep learning models which are trained using Inria and WHU Building datasets. The images in the second row are cropped and resized from the original images of predicted segmentation-maps.

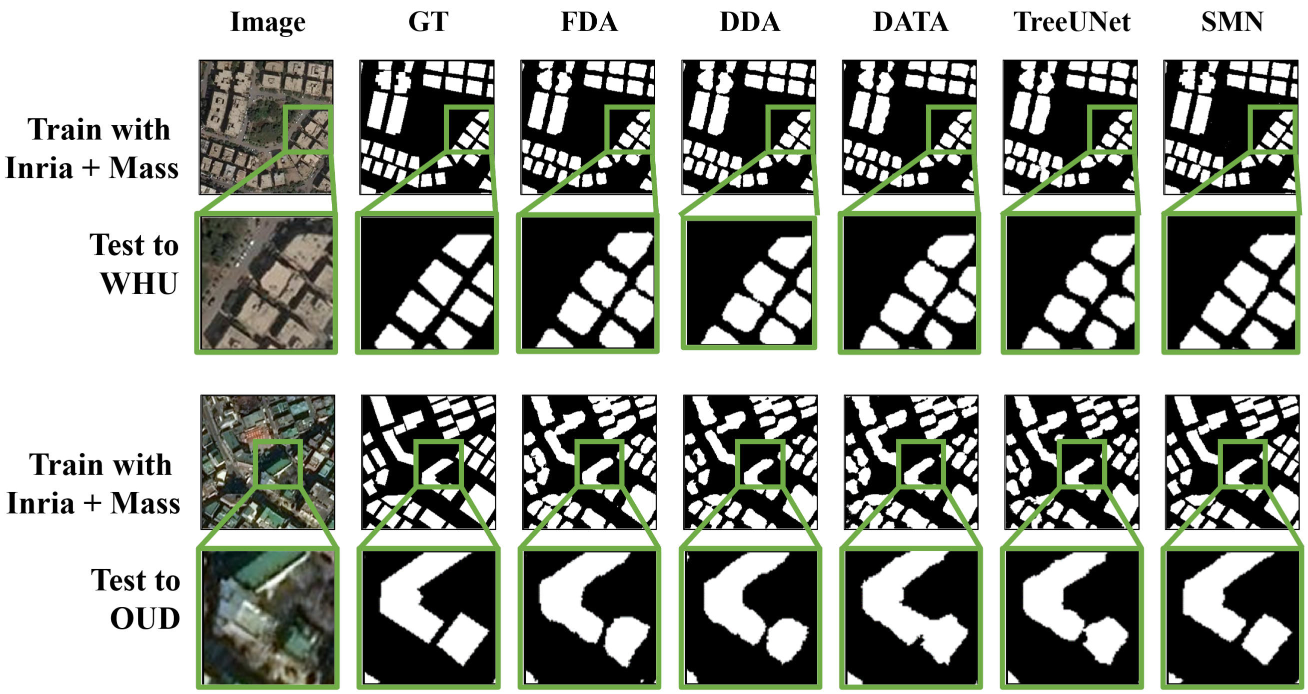


Figure 5: The predicted segmentation-map by deep learning models which are trained using Inria and Massachusetts Building datasets. The images in the second row are cropped and resized from the original images of predicted segmentation-maps.

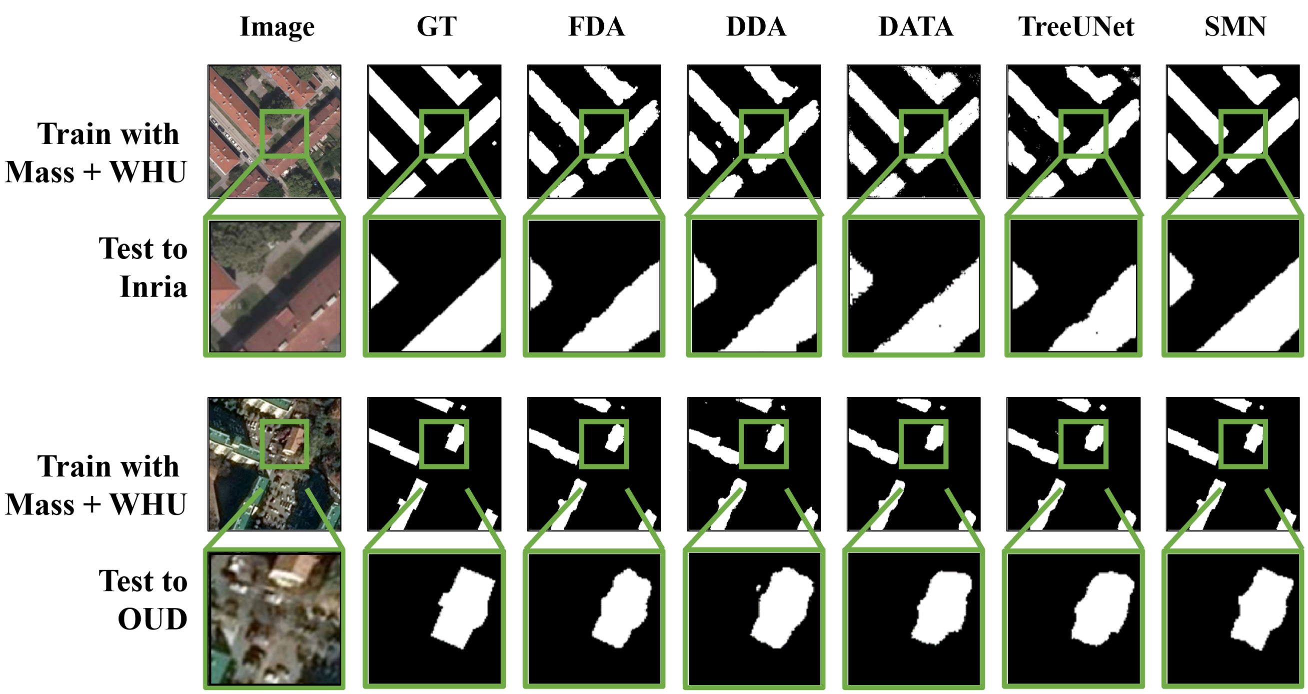


Figure 6: The predicted segmentation-map by deep learning models which are trained using Massachusetts Building and WHU Building datasets. The images in the second row are cropped and resized from the original images of predicted segmentation-maps.

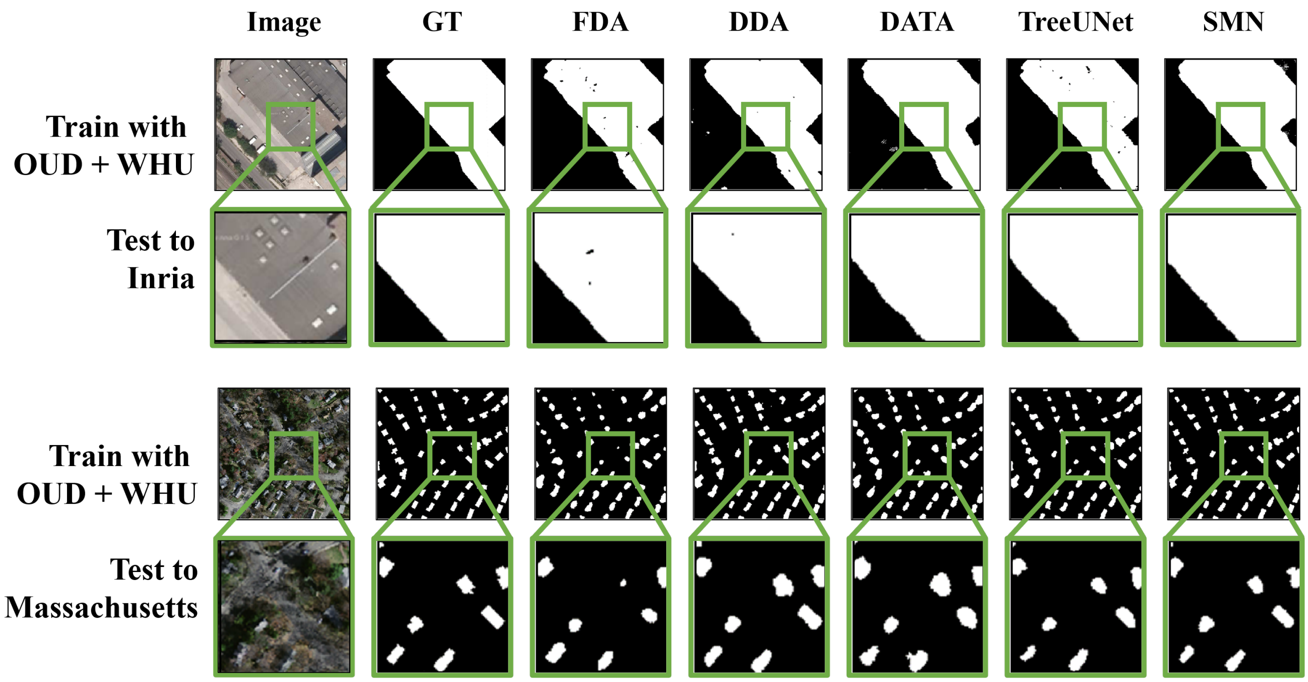


Figure 7: The predicted segmentation-map by deep learning models which are trained using OUD and WHU Building datasets. The images in the second row are cropped and resized from the original images of predicted segmentation-maps.

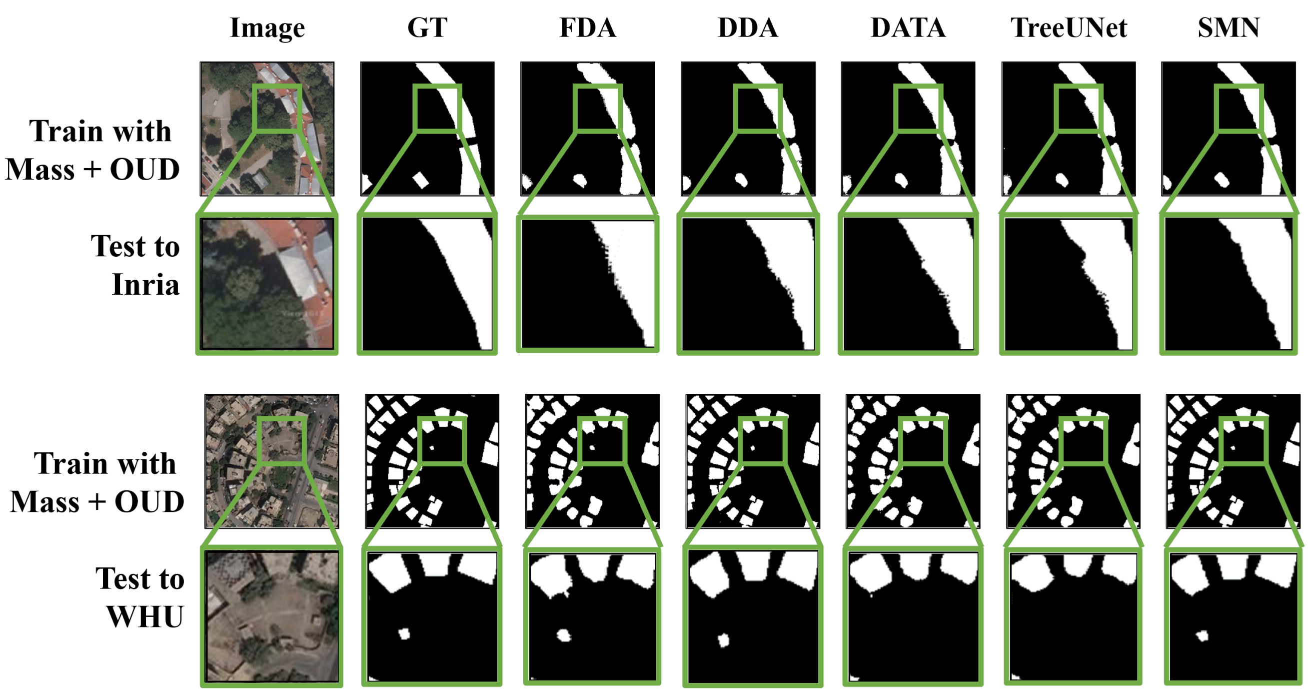


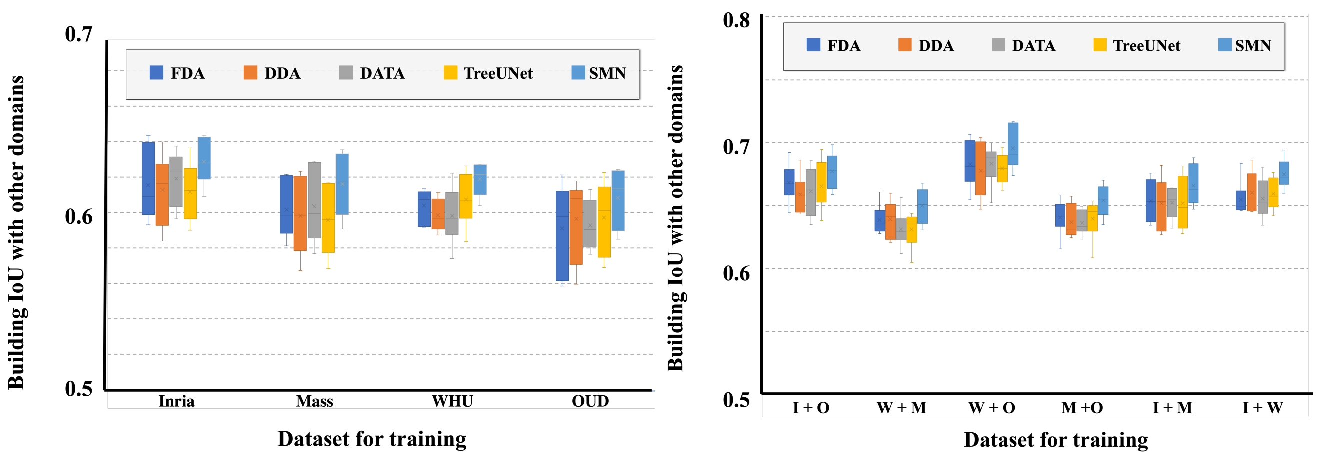
Figure 8: The predicted segmentation-map by deep learning models which are trained using Massachusetts Building and OUD datasets. The images in the second row are cropped and resized from the original images of predicted segmentation-maps.

Table 3: The experimental result. The illustrated domain is utilized as a training set whereas other domains are used as a test set. For example, in the first case of **Inria Building Dataset**, the Inria Building dataset is used for a training set and other three datasets of Massachusetts Building, WHU Building, and Our Urban datasets are utilized as a test set. The highest value is marked as **bold**, and the second one is underlined. The maximum performance difference between SMN and other networks is 8.48%.

IoU	Trial I	Trial II	Trial III	Trial IV	Trial V	Trial VI	Average	MAX	MIN
Inria Building Dataset									
Max. Improvement							0.97%	5.95%	
FDA	0.5933	0.6437	0.6010	0.6384	0.6060	0.6121	0.6158	**0.6437**	0.5933
DDA	0.5842	0.6402	0.6232	0.6118	0.6218	0.5958	0.6128	0.6402	0.5842
DATA	0.6058	0.6377	0.5966	0.6295	0.6259	0.6201	0.6193	0.6377	0.5966
TreeUNet	0.5903	0.6367	0.6212	0.6172	0.5987	0.6076	0.6120	0.6367	0.5903
SMN	0.6092	0.6437	0.6280	0.6426	0.6283	0.6220	**0.6290**	**0.6437**	**0.6092**
Massachusetts Buildings Dataset									
Max. Improvement							1.23%	6.82%	
FDA	0.5908	0.6217	0.6046	0.5814	0.5918	0.6207	0.6018	0.6217	0.5814
DDA	0.5675	0.6060	0.6196	0.5916	0.5824	0.6235	0.5984	0.6235	0.5675
DATA	0.5770	0.6023	0.6283	0.5970	0.5886	0.6293	0.6038	0.6293	0.5770
TreeUNet	0.5685	0.6088	0.6172	0.5839	0.5805	0.6164	0.5959	0.6172	0.5685
SMN	0.5910	0.6293	0.6324	0.6017	0.6066	0.6357	**0.6161**	**0.6357**	**0.5910**
WHU Building Dataset									
Max. Improvement							1.15%	5.33%	
FDA	0.5920	0.5922	0.6068	0.6134	0.6115	0.6082	0.6040	0.6134	0.5920
DDA	0.5999	0.5941	0.5876	0.6116	0.6066	0.5921	0.5986	0.6116	0.5876
DATA	0.5741	0.5943	0.5924	0.6224	0.5993	0.6079	0.5984	0.6224	0.5741
TreeUNet	0.5837	0.6014	0.6087	0.6049	0.6266	0.6199	0.6075	0.6266	0.5837
SMN	0.6040	0.6230	0.6124	0.6268	0.6274	0.6204	**0.6190**	**0.6274**	**0.6040**
Our Urban Dataset									
Max. Improvement							1.13%	6.57%	
FDA	0.5588	0.5978	0.5978	0.6214	0.5628	0.6092	0.5913	0.6214	0.5588
DDA	0.5598	0.6178	0.6063	0.6099	0.5744	0.6109	0.5965	0.6178	0.5598
DATA	0.5820	0.6133	0.5873	0.5940	0.5767	0.6049	0.5930	0.6133	0.5767
TreeUNet	0.5770	0.6228	0.6120	0.6033	0.5692	0.5997	0.5973	0.6228	0.5692
SMN	0.5850	0.6245	0.6132	0.6234	0.5916	0.6138	**0.6086**	**0.6245**	**0.5850**
Total Maximum Improvement (MAX - MIN)							3.77%	8.48%	

Table 4: The experimental result. The illustrated two domains are utilized as a training set wherea other two domains are used as a test set. For example, in the first case of **Inria Building Dataset + WHU Building Dataset**, the Inria Building and WHU Building datasets are used for the training set, and other two datasets of Massachusetts Building and Our Urban datasets are utilized as a test set. The highest value is marked as **bold**, and the second one is underlined. The maximum performance difference between SMN and other networks is 11.19%.

IoU	Trial I	Trial II	Trial III	Trial IV	Trial V	Trial VI	Average	MAX	MIN
Inria Dataset + Our Urban Dataset									
Max. Improvement							1.06%	9.36%	
FDA	0.6446	0.6924	0.6745	0.6640	0.6715	0.6635	<u>0.6684</u>	0.6924	<u>0.6446</u>
DDA	0.6436	0.6863	0.6590	0.6578	0.6632	0.6455	0.6592	0.6863	0.6436
DATA	0.6350	0.6859	0.6654	0.6443	0.6767	0.6607	0.6614	0.6859	0.6350
TreeUNet	0.6381	0.6947	0.6812	0.6634	0.6581	0.6584	0.6657	<u>0.6947</u>	0.6381
SMN	0.6592	0.6985	0.6864	0.6707	0.6847	0.6653	**0.6775**	**0.6985**	**0.6592**
WHU Building Dataset + Massachusetts Building Dataset									
Max. Improvement							0.79%	4.36%	
FDA	0.6303	0.6279	0.6325	0.6414	0.6390	0.6610	0.6387	<u>0.6610</u>	<u>0.6279</u>
DDA	0.6209	0.6436	0.6243	0.6600	0.6397	0.6482	<u>0.6395</u>	0.6600	0.6209
DATA	0.6121	0.6266	0.6298	0.6339	0.6288	0.6568	0.6313	0.6568	0.6121
TreeUNet	0.6049	0.6338	0.6376	0.6438	0.6260	0.6404	0.6311	0.6438	0.6049
SMN	0.6310	0.6514	0.6378	0.6610	0.6504	0.6679	**0.6499**	**0.6679**	**0.6310**
WHU Building Dataset + Our Urban Dataset									
Max. Improvement							1.21%	6.98%	
FDA	0.6549	0.7064	0.7005	0.6832	0.6788	0.6744	0.6830	<u>0.7064</u>	0.6549
DDA	0.6471	0.7040	0.6999	0.6627	0.6623	0.6911	0.6778	0.7040	0.6471
DATA	0.6526	0.6887	0.6999	0.6798	0.6888	0.6907	<u>0.6834</u>	0.6999	0.6526
TreeUNet	0.6621	0.6879	0.6963	0.6709	0.6743	0.6856	0.6795	0.6963	<u>0.6621</u>
SMN	0.6740	0.7157	0.7168	0.6852	0.6900	0.6915	**0.6955**	**0.7168**	**0.6740**
Massachusetts Building Dataset + Our Urban Dataset									
Max. Improvement							1.34%	6.15%	
FDA	0.6159	0.6402	0.6486	0.6420	0.6395	0.6586	<u>0.6408</u>	<u>0.6586</u>	0.6159
DDA	0.6284	0.6579	0.6280	0.6248	0.6330	0.6501	0.6370	0.6579	<u>0.6248</u>
DATA	0.6229	0.6456	0.6325	0.6346	0.6330	0.6497	0.6364	0.6497	0.6229
TreeUNet	0.6088	0.6369	0.6486	0.6432	0.6539	0.6477	0.6398	0.6539	0.6088
SMN	0.6350	0.6633	0.6502	0.6458	0.6607	0.6703	**0.6542**	**0.6703**	**0.6350**
Inria Dataset + Massachusetts Building Dataset									
Max. Improvement							1.22%	6.09%	
FDA	0.6382	0.6408	0.6657	0.6762	0.6345	0.6690	<u>0.6541</u>	0.6762	<u>0.6345</u>
DDA	0.6459	0.6272	0.6613	0.6641	0.6307	0.6820	0.6519	<u>0.6820</u>	0.6272
DATA	0.6444	0.6476	0.6625	0.6637	0.6323	0.6643	0.6525	0.6643	0.6323
TreeUNet	0.6334	0.6279	0.6464	0.6817	0.6507	0.6710	0.6519	0.6817	0.6279
SMN	0.6472	0.6545	0.6706	0.6818	0.6553	0.6882	**0.6663**	**0.6882**	**0.6472**
Inria Dataset + WHU Building Dataset									
Max. Improvement							1.47%	5.94%	
FDA	0.6462	0.6470	0.6469	0.6510	0.6835	0.6548	0.6549	0.6835	<u>0.6462</u>
DDA	0.6460	0.6489	0.6636	0.6453	0.6862	0.6717	<u>0.6603</u>	<u>0.6862</u>	0.6453
DATA	0.6348	0.6530	0.6528	0.6468	0.6661	0.6808	0.6557	0.6808	0.6348
TreeUNet	0.6420	0.6606	0.6518	0.6545	0.6766	0.6706	0.6593	0.6766	0.6420
SMN	0.6602	0.6727	0.6692	0.6714	0.6942	0.6821	**0.6750**	**0.6942**	**0.6602**
Total Maximum Improvement (MAX - MIN)							6.45%	11.19%	



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

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

Figure 9: The segmentation results of SMN compared to other state-of-the-art networks. The illustrated dataset is used as a training set, and other datasets of different domains are utilized as a test set.