A Multi-Stage Deep Learning Framework for Shadow Detection in Aerial Orthophotos

Yohei Kobayashi [1], Mitsuteru Sakamoto [1], Sho Nakamura [1], Toshiaki Satoh [1]

[1] PASCO Meguro Sakura Bldg, 1-7-1 Shimomeguro, Meguro-ku, Tokyo 153-0064, Japan
(yiohhs2526@pasco.co.jp, moittu9191@pasco.co.jp, sahrou7863@pasco.co.jp, tuoost7017@pasco.co.jp)

Keywords: shadow detection, orthophoto, aerial imagery, multi-resolution, deep learning, semantic segmentation

Abstract

Shadow correction is an important preprocessing step not only for visual enhancement but also for improving object recognition performance in remote sensing imagery. Although many datasets and deep learning models have been proposed for shadow detection and removal, most of them focus on natural images. In contrast, high-resolution aerial orthophotos contain large continuous shadows caused by tall buildings, especially in urban areas, and existing models often fail to handle such large-scale structures effectively.
In this study, we construct a new shadow annotation dataset specifically designed for aerial orthophotos with spatial resolutions of 20 cm/pixel and 5 cm/pixel. Furthermore, we propose a three-stage multi-resolution segmentation framework that progressively refines shadow predictions from low to high resolution. Predictions from lower-resolution stages are used as auxiliary information to guide higher-resolution prediction. Experimental results demonstrate that the proposed approach improves fuzzy Intersection over Union (IoU) by approximately 0.05 compared with a previously published shadow detection model, and also outperforms a single-stage baseline, particularly for large continuous shadow regions. The framework is also applicable to other large-scale segmentation tasks requiring extensive receptive fields.

1. Introduction

Shadow correction in images is an important task not only for improving visual appearance but also for enhancing object recognition performance. A simple method for shadow correction is to apply brightness balance adjustments such as gamma correction. However, such global adjustments applied uniformly to the entire image often lead to poor results—for example, when calibrated based on shadow regions, the non-shadow areas may become excessively bright.
To address this issue, it is necessary to segment the image into shadow and non-shadow regions and apply corrections separately. However, if the segmentation is performed using simple brightness thresholds, dark objects under illumination may be incorrectly classified as shadows, leading to problems during correction. For this reason, research has explored the use of learning-based models, such as deep learning, for shadow detection and region segmentation (Hu et al., 2018; Wang et al., 2019).
Recent advances in deep learning have significantly improved image segmentation performance. Convolutional neural network (CNN)–based approaches such as Fully Convolutional Networks (FCN) and U-Net have become widely used for pixel-wise prediction tasks (Long et al., 2015; Ronneberger et al., 2015). These methods have also been applied to remote sensing imagery, where semantic segmentation of aerial and satellite images has become an important research topic (Marmanis et al., 2016; Audebert et al., 2017; Zhu et al., 2017).
Several datasets for shadow correction and detection have been publicly released, and deep learning models for shadow detection and removal have been developed using these datasets (Hu et al., 2018; Wang et al., 2019). Examples include methods based on encoder–decoder architectures such as U-Net, as well as more recent approaches using Vision Transformers and related architectures (Dosovitskiy et al., 2021; Liu et al., 2021).
However, many of these datasets primarily consist of natural images, which naturally limits their applicability to similar types of imagery. In contrast, there are only a limited number of datasets targeting aerial orthophotos, such as AISD (Wang et al., 2020), and these often have restrictions such as limited image sizes. Aerial orthophotos are significantly larger than natural images, and when tall buildings are present, large continuous shadows are likely to occur. As a result, conventional shadow detection models often struggle to detect shadows accurately in such imagery, leading to an increasing number of cases where related correction processes do not perform well.

Area	Number of Images	Total pixel count	Shadow ratio	Resolution
AISD(as a reference)	514	148M	21.49%	20 cm/pix
Urayasu	1	87M	17.1%	20 cm/pix
Tokyo	2	175M	65.7%	5 cm/pix
Kobe	5	106M	40.7%	5 cm/pix
Kamakura	1	70M	25.5%	5 cm/pix

Table 1. Comparison of shadow annotation dataset across different areas. The table summarizes the number of images, total pixel count, shadow ratio, and spatial resolution for each region.

2. Purpose and Method

2.1 Overall Strategy

In this study, to address the problem of shadow detection in orthophotos, we construct a new dataset and propose a multi-stage detection model that is effective for large-area shadow detection.
A key limitation of existing datasets is that many training images are divided into relatively small images of less than 1,000 pixels. Moreover, in many cases, shadows are artificially added through image processing or by blocking light sources in

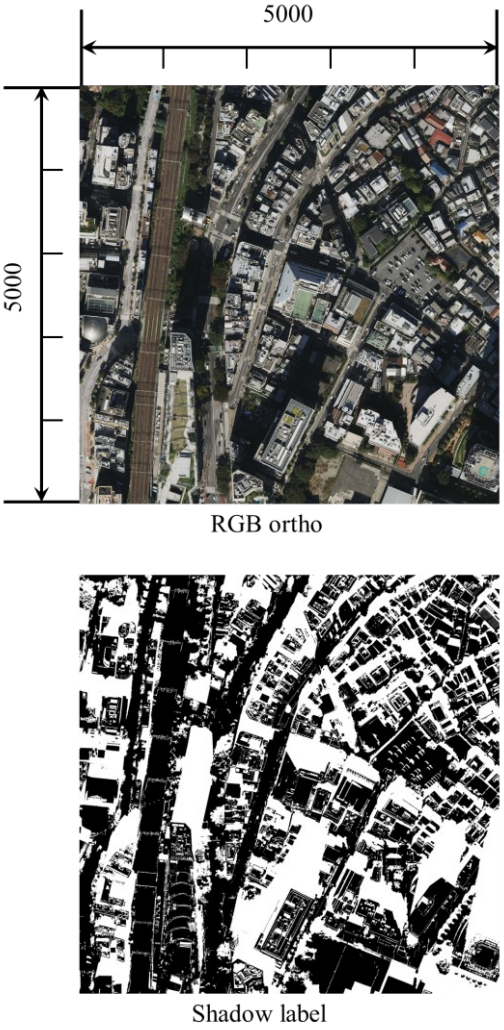

Figure 1. An example of the newly generated shadow label, which contains numerous large shadow regions.

real space, resulting in relatively simple problem structures (Hu et al., 2018; Wang et al., 2019).

In contrast, the problem becomes more complex when factors such as the objects casting the shadows or the appearance of shaded areas (e.g., the back sides of objects) are included in the image. As a result, the domain gap between such datasets and real orthophotos—where these elements are mixed—is too significant to ignore.

To overcome this issue, we first create shadow labels for aerial orthophotos collected from multiple regions. At the same time, we develop a deep learning model capable of handling large shadow regions.

As illustrated in Figure 1, shadows in urban orthophotos are often extremely large, sometimes exceeding several thousand pixels. Therefore, when images are cropped to commonly used deep learning input sizes such as 512 × 512 pixels, it is not uncommon for the entire cropped image to be covered by shadow, making accurate shadow recognition difficult.

Accordingly, we propose a model that enables efficient learning and prediction of shadow labels for such large-scale shadow regions.

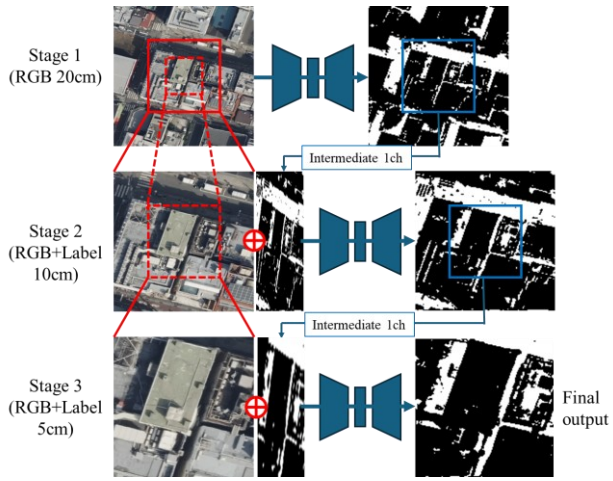

Figure 2. The newly proposed 3-stage inference model improves prediction accuracy by concatenating low-resolution inference results with the RGB image as input.

2.2 Datasets

In this study, we newly constructed a shadow-label dataset specifically designed for orthophotos. The dataset includes images with multiple spatial resolutions, namely 20 cm/pixel and 5 cm/pixel. The image sizes range from approximately 1,000 to 10,000 pixels per side.

Shadow annotations were created for three major regions, primarily covering urban and built-up areas. The number of images, total number of pixels, and proportion of shadow pixels are summarized in Table 1.

Shadow labels were generated as binary values: 0 indicates non-shadow, and 1 indicates shadow. The criterion for determining the presence of shadow was based on direct sunlight. If a surface was directly illuminated by sunlight, it was labeled as non-shadow. If sunlight was obstructed by any object, the area was labeled as shadow. Therefore, surfaces such as slanted roofs, which may visually appear to be in shadow but are in fact directly illuminated by sunlight, were labeled as non-shadow. Examples of the labels are shown in Figure 1. Extremely fine shadows, such as those cast by power lines or utility poles, were not included as shadow labels.

2.3 Model

In this study, shadow detection is formulated as a pixel-wise prediction task, similar to general object segmentation, where each pixel is classified as belonging to a shadow or non-shadow region.

Various image segmentation models have been proposed for such tasks. CNN-based segmentation methods such as FCN, U-Net, and DeepLab have demonstrated strong performance in pixel-level prediction (Long et al., 2015; Ronneberger et al., 2015; Chen et al., 2018). These models have also been widely used in remote sensing applications (Marmanis et al., 2016; Audebert et al., 2017).

However, their input size is fundamentally limited. For example, CNN-based models typically accept input sizes of around 512 × 512 pixels, or at most 1024 × 1024 pixels. Moreover, due to the inherent properties of CNNs, even if the input window size is increased, the effective receptive field of deeper layers remains

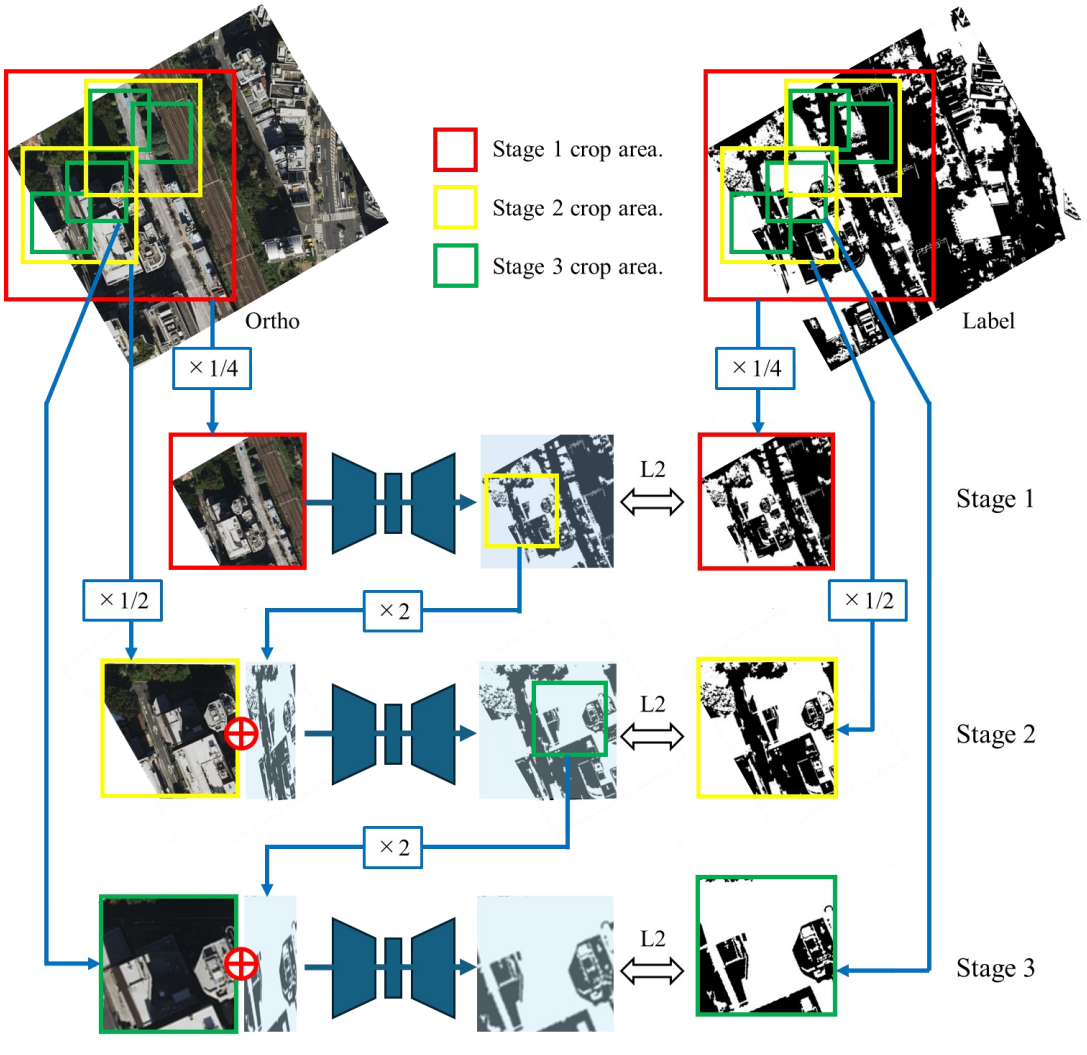

Figure 3. Training procedure of the model with augmentation. After applying geometric transformations, the input image is cropped in three stages. From the region extracted in the first stage, two regions for the second stage are randomly cropped. Subsequently, from each second-stage region, two regions for the third stage are randomly cropped in the same manner.

limited. This makes it difficult to accurately recognize large-scale objects.

Several approaches have been proposed to address this issue, such as dilated convolutions and modified pooling strategies (Yu and Koltun, 2016; Chen et al., 2018). However, these methods also have practical limitations.

On the other hand, Transformer-based segmentation models can analyze global relationships within the entire input region regardless of spatial distance (Dosovitskiy et al., 2021; Liu et al., 2021). Nevertheless, increasing the input size results in a dramatic increase in memory consumption, making such approaches impractical for very large images.

To address these challenges, we propose a multi-stage deep learning model that leverages multiple resolution levels. As illustrated in Figure 2, the model consists of three stages trained from lower to higher resolutions. The prediction results obtained at lower resolutions are used as auxiliary information to assist higher-resolution predictions.

In this study, we employ a ConvNeXt-based architecture (Liu et al., 2022) as the backbone network. In the first stage, the target image is downsampled to one-quarter of its original resolution, and training and inference are performed without auxiliary information. In the second stage, the target image at half resolution is used together with auxiliary information obtained by upsampling the first-stage prediction by a factor of two. Training and inference are then conducted using both inputs. In the final stage, training and inference are performed at the original image resolution, again incorporating auxiliary information from the previous stage.

By employing this multi-resolution approach, the proposed method enables the detection of large-scale shadows while simultaneously capturing fine-grained details.

3. Experiment

3.1 Training stage

As shown in Table 1, the training images are generally larger than 2048 × 2048 pixels, which is the minimum required input size for the model. Therefore, cropping for training may appear relatively straightforward; however, in practice, it becomes somewhat complex due to data augmentation operations such as rotation and scaling.

In this study, we set the model’s window size to 512 × 512 pixels and performed data loading and preprocessing

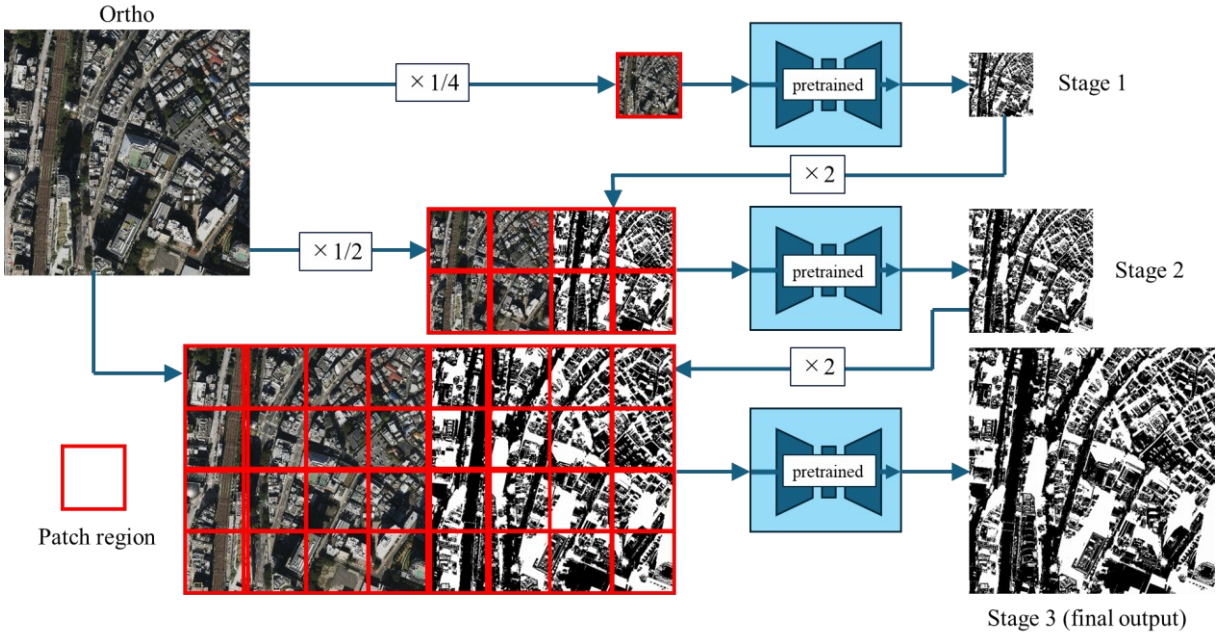

Figure 4. Processing pipeline during inference. The input image is resized to one-quarter, one-half, and full resolution, and inference is performed sequentially from the lowest resolution using a tiling strategy. At each stage, predictions are generated for all tiles, and these results are merged to produce the overall inference output for that stage. In the second and third stages, the prediction result from the previous stage is incorporated as auxiliary information for prediction.

accordingly as shown in Figure 3. First, to enable training at one-quarter resolution, we applied augmentations such as rotation and scaling to the original-resolution training images, and then cropped regions of 2048 × 2048 pixels. At this stage, any empty areas outside the original image were padded with white.

For the first-stage model, these cropped images were downsampled to one-quarter resolution (512 × 512) and used as input. The corresponding labels were also resized to the same resolution and used for supervised learning.

After completing the training of the first-stage model, we proceeded to the second-stage model. In this stage, the shadow probability map predicted by the trained first-stage model was enlarged by a factor of two and used as auxiliary information. The second-stage model used the original cropped 2048 × 2048 images downsampled to 1024 × 1024. Since the model input size is 512 × 512, two random crops were extracted from each 1024 × 1024 image for training. The corresponding regions were also cropped from the ground-truth labels and used for supervision.

The same cropping procedure was applied to the third-stage model. Specifically, for the two regions cropped in the second stage, two additional random sub-crops were extracted from each region, resulting in a total of four cropped regions for training. The corresponding ground-truth labels were also extracted from the full-resolution 2048 × 2048 ground-truth images and used for supervision.

Because the sub-crops were selected randomly within a limited range, some cropped areas could overlap; however, no special processing was performed to exclude such overlaps.

3.2 Inference stage

The multi-stage design allows the model to first capture large-scale contextual information at low resolution and then progressively refine shadow boundaries at higher resolutions. During inference, as shown in Figure 4, the target image is first downsampled to one-quarter of its original resolution. It is then divided into overlapping 512 × 512 tiles, and a per-pixel shadow probability map is predicted for each tile. These tile-wise probability maps are subsequently merged to generate the overall shadow probability map at one-quarter resolution. This probability map is then upsampled by a factor of two and used as auxiliary information for the next stage.

In the second stage, using this auxiliary probability map, the original target image is downsampled to half resolution. The RGB channels are combined with the auxiliary probability map to form a four-channel input (RGB + probability). As in the first stage, the image is divided into overlapping 512 × 512 tiles, and shadow probability maps are predicted for each tile.

In the final third stage, the shadow probability map predicted in the previous stage is again upsampled by a factor of two and used as full-resolution auxiliary information. The image is then tiled into 512 × 512 patches in the same manner, and probability predictions are generated and merged to produce the final shadow probability map. The probability maps were directly compared with the soft ground-truth labels without binarization.

4. Result

The proposed model was trained at each stage by randomly cropping patches of a specified size from the training data corresponding to Table 1, followed by data augmentation. The batch size was set to 4, and training was conducted for 50,000 iterations at each stage.

In the first stage, where the resolution becomes one-quarter of the original, the training images were downsampled to an equivalent resolution of 20 cm. Images originally acquired at 20 cm resolution were used without modification. Patches of size 512 × 512 were randomly extracted and used for training.

In the second stage, where the resolution becomes one-half of the original, the training images were rescaled to an equivalent

Area	Total pixel count	Shadow ratio	Resolution
Tokyo1	25M	45.2%	5 cm/pix
Tokyo2	16M	74.3%	5 cm/pix
Sendai	48M	59.9%	10 cm/pix

Table 2. Specifications of test datasets across different areas.

Area	Model	Precision	Recall	F1	IoU
Tokyo1	FDRNet	0.735	0.828	0.779	0.638
	BDRAR	0.888	0.879	0.883	0.791
	1-Stage	**0.943**	0.828	0.882	0.789
	2-Stage	0.870	0.866	0.868	0.767
	3-Stage	0.881	**0.897**	**0.889**	**0.800**
Tokyo2	FDRNet	0.908	0.870	0.889	0.799
	BDRAR	0.957	0.890	0.922	0.855
	1-Stage	**0.981**	0.707	0.822	0.698
	2-Stage	0.948	0.922	0.935	0.877
	3-Stage	0.945	**0.947**	**0.946**	**0.893**
Sendai	FDRNet	0.768	0.449	0.557	0.395
	BDRAR	**0.924**	0.626	0.747	0.596
	1-Stage	0.892	**0.946**	**0.918**	**0.849**
	2-Stage	0.922	0.906	0.914	0.841
	3-Stage	0.907	0.923	0.915	0.843

Table 3. Comparison of accuracy metrics between the previously published model and the proposed one-stage, two-stage, and three-stage models. The weights of the previously published model were trained on a different publicly available dataset, not on the dataset used in this study. Soft labels were used for the segmentation labels.

resolution of 10 cm as described in the previous section. Images originally at 20 cm resolution were upsampled by a factor of two. From images cropped at twice the patch size, two patches were randomly sampled to form a batch. Specifically, two 1024 × 1024 patches were extracted per batch, resulting in an effective batch size of 4.

In the third stage, the training images with a resolution of 5 cm were used at their original resolution, while the 20 cm resolution data were excluded. A region of size 2048 × 2048 was first cropped, and four patches extracted from this region were used to construct a batch of size 4.

After completing the training of all stages, model performance was evaluated using the test regions listed in Table 2. The comparison included the one-stage, two-stage, and three-stage models proposed in this study, as well as an existing model with publicly available pretrained weights. For the existing model, inference was performed using a sliding-window approach with a window size of 512 × 512, consistent with the inference procedure used for the one-stage model. It should be noted that the pretrained weights of the existing model were trained on publicly available datasets, such as the SBU dataset (Vicente et al., 2016), rather than on the dataset used in this study. SAM-based models were excluded from the comparison because they were not compatible with this dataset and failed to detect shadows reliably.

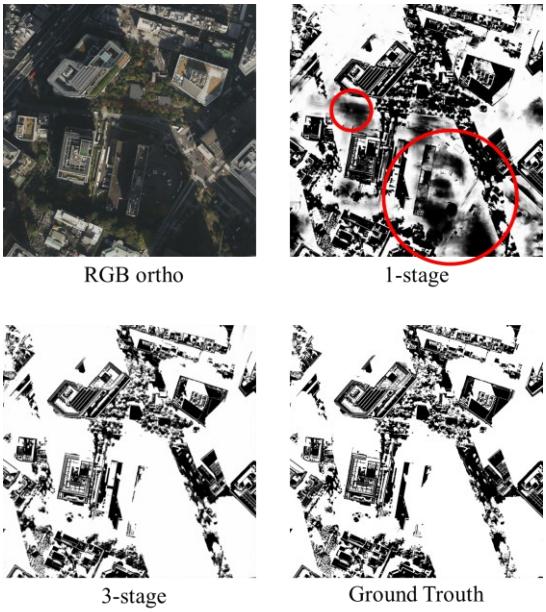

Figure 5. Example of shadow detection results. Compared with the three-stage model, the single-stage model exhibits missed detections inside large shadow regions (highlighted in the red circle).

Table 3 summarizes the evaluation results for the test regions described above. Unlike conventional binary segmentation evaluation using labels of 0 and 1, the metrics in this study were computed using soft labels represented as continuous values. Overall, the proposed three-stage model achieved the best performance across most evaluation metrics. However, in some regions the performance advantage was limited, and in a few cases the accuracy was lower than that of other models.

An example of the prediction results is shown in Fig. 5. The one-stage model using only 5 cm resolution failed to detect the interior regions of large shadows, whereas the proposed three-stage model showed significantly fewer missed detections. In contrast, little difference was observed between the models in the detection of small shadows.

5. Discussion

One advantage of the proposed model is its ability to more accurately segment spatially continuous regions, such as shadows that extend over large areas. Examples of such targets include not only shadows but also large buildings, connected river systems, road networks, land-cover classes such as vegetation, as well as large-scale surfaces such as oceans and ice sheets.

Although ConvNeXt was used in this study, the proposed framework is not limited to this architecture. Various image transformation models, such as Vision Transformers (Dosovitskiy et al., 2021) and hierarchical transformers such as Swin Transformer (Liu et al., 2021), could also be applied. However, applying this method to detection-based models, such as those commonly used in remote sensing object detection (Cheng and Han, 2016; Cheng and Han, 2017), would require additional modifications.

Since the proposed model employs a three-stage training and inference mechanism, it is expected to require more execution time than a single-stage approach. During training, computation

time increases proportionally with the number of stages. However, during inference at full resolution, the auxiliary information is generated from lower-resolution predictions (with areas reduced to 1/4 or 1/16 of the original). As a result, the overall inference time is approximately 1.3 times that of a single-stage model.

On the other hand, the effectiveness of the proposed method depends on the size of the input images. If the input image size is limited—for example, smaller than 2048 × 2048, which is four times the model input size—then the benefit of performing lower-resolution inference becomes relatively small. In particular, for datasets such as AISD or ISTD (Wang et al., 2018), where image resolutions are below 1000 × 1000, the advantage of this approach is likely to be limited.

Furthermore, if a model uses a small input size and the entire region is covered by shadow, resolution-independent approaches—such as recognizing shadows based on global brightness characteristics—may already be effective. In such cases, the advantage of the proposed multi-stage method may diminish.

6. Conclusions

In this study, we newly annotated shadow labels for large shadow regions in orthoimages, focusing on relatively wide and spatially continuous areas.

By employing a three-stage inference model that handles three different resolutions, we achieved both improved recognition of large-scale shadow regions and accurate detection of fine shadow details.

The proposed method is not limited to shadow segmentation. It is also expected to be effective for tasks that require a large receptive field, such as building segmentation and land-cover classification.

References

Audebert, J., Le Saux, B., Lefèvre, S., 2017. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. ISPRS J. Photogramm. Remote Sens., 134, 134–147. doi.org/10.1016/j.isprsjprs.2017.11.002.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell., 40(4), 834–848. doi.org/10.1109/TPAMI.2017.2699184.

Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. ISPRS J. Photogramm. Remote Sens., 117, 11–28. doi.org/10.1016/j.isprsjprs.2016.03.014.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE, 105(10), 1865–1883. doi.org/10.1109/JPROC.2017.2675998.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., 2021. An image is worth 16×16 words: Transformers for image recognition at scale. Int. Conf. Learn. Represent. (ICLR).

Hu, X., Fu, C., Zhu, L., Heng, P.-A., 2018. Direction-aware spatial context features for shadow detection. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 7454–7462. doi.org/10.1109/CVPR.2018.00778.

Liu, Z., Mao, H., Wu, C.-Y., et al., 2022. ConvNeXt: A ConvNet for the 2020s. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 11976–11986. doi.org/10.1109/CVPR52688.2022.01167.

Liu, Z., Lin, Y., Cao, Y., et al., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 10012–10022. doi.org/10.1109/ICCV48922.2021.00986.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 3431–3440. doi.org/10.1109/CVPR.2015.7298965.

Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of CNNs. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., III-3, 473–480. doi.org/10.5194/isprs-annals-III-3-473-2016.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci., 9351, 234–241. doi.org/10.1007/978-3-319-24574-4_28.

Wang, L., Ding, Y., Liu, C., et al., 2020. AISD: Aerial image shadow dataset for shadow detection. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 326–329. doi.org/10.1109/IGARSS39084.2020.9324327.

Wang, J., Li, X., Yang, J., 2019. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 1788–1797. doi.org/10.1109/CVPR.2019.00189.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. Int. Conf. Learn. Represent. (ICLR).

Vicente, T. F. Y., Hou, L., Yu, C. P., Hoai, M., Samaras, D., 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. European Conference on Computer Vision (ECCV).

Wang, J., Li, X., Yang, J., 2018. Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag., 5(4), 8–36. doi.org/10.1109/MGRS.2017.2762307.