AN INVESTIGATION OF THE INFLUENCES OF TRAINING DATA PROPERTIES ON AERIAL IMAGE SEGMENTATION

Güren Tan Dinga

Lab for Geoinformatics and Geovisualization, HafenCity University Hamburg, Germany - gueren.dinga@hcu-hamburg.de

Commission II,WG II/6

KEY WORDS: Remote Sensing, Segmentation, Deep Learning, Machine Learning, Training Data, UNET

ABSTRACT:

While numerous studies are being conducted to improve neural network performance for image segmentation, studies on the impact of training data in terms of data quality, bias and labeling noise are comparatively scarce. When opening state of the art algorithms to a large and varying dataset, they do not achieve the same results as under optimal and controlled conditions due to a mismatch of the data used for training and the data that is to be predicted. This paper presents an approach to show the influences of diverging image properties such as scale, contrast, brightness and saturation between the training data of a model and data that is to be predicted. For this purpose, a U-Net is trained to segment buildings in aerial images. It was found that while changes in brightness have a strong effect on precision, recall and F-score, a change in saturation does not have too much or even positive effect on segmentation. In general, however, it can be said that any differences between training and prediction data have a negative effect on segmentation results.

1. INTRODUCTION

The increasing overlap between remote sensing, computer vision and computer science led to rapid advancements regarding the analysis of geodetic data. The new possibilities, also due to increasing computer performance, opened up new approaches to geodetic data processing (Maxwell et al., 2018). At the same time, breakthroughs in machine learning led to rapid progress and new developments for remote sensing applications (Sagan et al., 2020). While the initial focus of industry and academia was on object detection using bounding boxes (e.g. with YOLO (Redmon et al., 2016)), further developments in the field of neural networks enabled the segmentation of raster data using Deep Learning (e.g. with U-Net (Ronneberger et al., 2015)). This in turn resulted in different approaches for the segmentation of satellite and aerial imagery (Kim et al., 2019), which made it possible to detect objects instead of bounding boxes.

Many of the aerial segmentation studies conducted since then have been concerned with the neural networks themselves, showing advances in the design of architectures that ultimately lead to better detection and segmentation results on benchmark datasets (Rahman M.A., 2016). Compared to the volume of research on neural networks and (fine tuning) their architectures, relatively little research is being done on the used data. Mostly, research is done in the context of data on topics such as class imbalance (Japkowicz and Stephen, 2002) or dataset size (Soekhoe et al., 2016), but rarely on the actual properties of the data such as exposure, color shift, or contrast values as they occur in the real world and in real-world applications during prediction. Especially in the field of aerial image segmentation, studies on image properties are pending.

This paper presents an approach to show the influences of scale, brightness, contrast and saturation as well as the combination of these factors on the segmentation results when predicting buildings from aerial imagery unknown to the neural network. The goal is to investigate whether these properties affect the segmentation results and if so, to what extent. For this purpose, said combinations are contrasted and compared with the segmentation results of data that has not been altered.

2. RELATED WORK

2.1 Direct dataset evaluation

As mentioned, much research in the field of aerial image segmentation regarding datasets amounts to a) the size of datasets (Sun et al., 2017) and b) the class imbalance of the training data (Li et al., 2021). While, to our knowledge, no publications deal with the direct impact of image properties mentioned in the introduction, in adjacent areas where training sets are as important as in image segmentation such as audio and signal processing, the impact of data quality and quantity is discussed when introducing new datasets (Manilow et al., 2019).

2.2 Data augmentation

Data augmentation, in this case image augmentation, is the process of automatically creating additional training data by making minor (or depending on the use case moderate) changes to it. Image augmentation algorithms can include flips, rotations, color space augmentations, geometric transformations and many more (Shorten and Khoshgoftaar, 2019). Some publications in the field of satellite and aerial segmentation use augmentations and describe those that haven been used like rotations and chromatic distortions (Li et al., 2018, Khryashchev et al., 2019). At the same time, information about how classification or segmentation results would look like without augmentations is often missing since it is not the main focus of given publications. When augmented and non-augmented segmentation and classification results are compared, they usually aren't compared granularly so the effect of each augmentation method is not always clear (Wu et al., 2019). It must also be mentioned that augmentations sometimes are only introduced into the training data to prevent the neural network from overfitting which means that in that case the network learns to perfectly

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2022-245-2022

model the training data which would lead to bad performance when introducing new data.

3. METHODOLOGY

3.1 Segmentation approach and data

In order to perform a study regarding the effects of different image properties for segmentation, images must first be segmented. After successes in the field of biomedical image processing, the U-Net found great appeal in the field of image segmentation and became a popular choice for segmentation tasks in various fields, including aerial and satellite image segmentation (Zeng et al., 2019, Tang et al., 2019, Darapaneni et al., 2020). Accordingly, for this experiment, a U-Net was used to segment aerial images. The training data[1] has three channels (RGB), a resolution of 10000 x 10000 pixels with a ground sampling distance of 10 cm. The data was chosen to include both developed urban areas as well as rural areas. So that the images do not have to be reduced in resolution for training, they are divided into smaller tiles of 512 x 512 pixels. The same applies to the validation and test data. The training labels were automatically generated from outline polygons of buildings of the corresponding areas so labeling noise can not be ruled out entirely. However, the labels were examined for the proportion of building pixels with respect to background pixels, so the effects of class imbalance can be mitigated to some degree.

The training was carried out with around 20.000 image tiles. The best resulting model was used to predict the segmentation maps of unprocessed images. The segmentation maps of the unaltered images are later used as reference for examining the influences of the alteration of image properties on segmentation results. Figure 1 shows an overview of the training and prediction process.

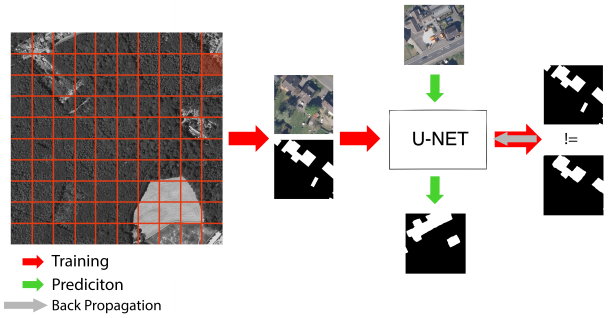


Figure 1. After splitting up the initial true-orthophoto into tiles, the training images and labels (black and white binary images) are being used to train the U-Net. While red arrows show the path to train the U-Net, the green arrows show the path to predict (to the network) unknown images. The gray arrow (back propagation) is supposed to depict that there is a training process included.

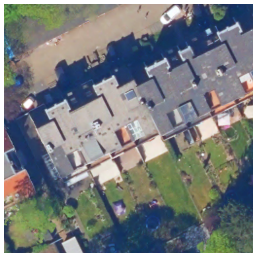
3.2 Image Alteration

While designing the concept, a decision had to be made about which changes in image properties to consider. To look for clues about discrepancies in datasets, publicly available data were examined. A look at the aerial image data publicly available in Germany[2] revealed that taking into account variations in contrast, saturation and exposure/brightness would prove useful. A shift in color could not be observed. Accordingly, the decision was made in favor of the former properties. It should be mentioned that these properties are not specific to Germany as a brief research has shown similar occurrences in other datasets (e.g. (Mnih, 2013)[3], USGS[4]). The image data were modified in both directions: this means that for contrast, images with raised contrast and lowered contrast were used. The same applies to saturation and exposure. Furthermore, as a first attempt to investigate the effect of scale changes, data enlarged by 10 % and by 50 % were used. Figure 2 shows an exemplary section of the alterations.

(a) Original Image

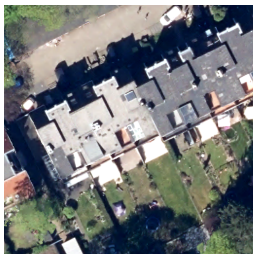
(b) 50 % Crop

(c) Low Saturation



(d) High Saturation

(e) Low Contrast



(f) High contrast

(g) Low Brightness

(h) High Brightness

[1] https://www.opengeodata.nrw.de/produkte/geobasis/lusat/dop/

[2] https://www.govdata.de/

[3] https://www.cs.toronto.edu/ vmnih/data/

[4] https://earthexplorer.usgs.gov/

(i) High Contrast, Low Saturation

(j) High Brightness, Low Saturation

Figure 2. The images are altered to compare the segmentation results in the next step. Due to space limitations, not all possbile combinations are shown.

3.3 Evaluation

To ensure comparability between the predicted images, the same metrics were chosen for all datasets where image alterations have been made. The fact that absolute values are ultimately important for segmentation results cannot be denied. For this experiment, however, the comparison of different variations is important. Accordingly, the results are additionally presented as percentages compared to the segmentation results of the original data. Furthermore, precision (equation 1), recall (equation 2) and the F-score (equation 3) are highlighted (Powers, 2007).

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (1)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (2)$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

With *True Positive* being an outcome where the model correctly predicts the positive class (model predicts a building pixel as building) and *True Negative* correctly predicts the negative class (model predicts a background pixel as background). A *False Positive* being an outcome where the model predicts the positive class incorrectly (model predicts a background pixel as building) and a *False Negative* being an outcome, where the model incorrectly predicts the negative class (model predicts a building pixel as background).

4. RESULTS AND DISCUSSION

The segmentation maps and results generated with the original and unaltered data serve as a reference for the segmentation maps and results generated with the altered data. Table 1 shows the results of the evaluation of the unaltered image data whereas tables 2, 3 and 4 show the recall, precision and F-score of the evaluation of the of the altered data respectively.

	Unaltered Data
Precision	0.716
Recall	0.630
F-score	0.656

Table 1. Recall, Precision and F-score of the segmentation results when using unaltered data for the predictions

Table 2 shows the percentage difference of the recall values in comparison to the recall values of the original data. By observing the recall we can look at the proportion of actual positives that have been identified correctly. According to our results, predicting imagery with lower brightness than our training data, especially in combination with lowering the contrast or saturation, will statistically lead to much worse segmentation results (-30 %). The numbers show that the recall suffers in particular from the fact that brightness is reduced. With an increased brightness, recall values are up to 11 % better than the recall values of the unaltered image data. We assume that especially when brightness *and* contrast are reduced, no sufficiently distinctive features remain that can be picked up during training for better segmentation results.

RECALL						
	lb	hb	lc	hc	ls	hs
lb	-16 %	-	-	-	-	-
hb	-	+10 %	-	-	-	-
lc	-30 %	+5 %	+6 %	-	-	-
hc	-10 %	+1 %	-	+8 %	-	-
ls	-26 %	+11 %	-2 %	+3 %	+8 %	-
hs	-6 %	-1 %	0 %	-1 %	-	+4 %

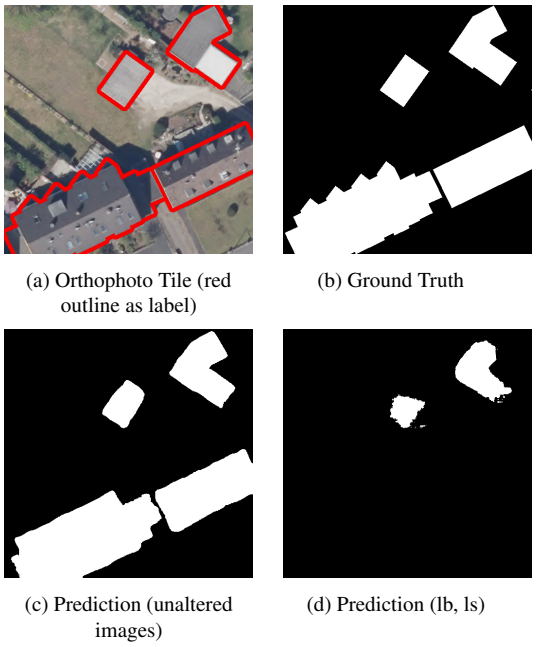
Table 2. Recall percentage values when using altered images in comparison to the recall values when using original, unaltered images

where lb, hb = low & high brightness
lc, hc = low & high contrast
ls, hs = low & high saturation

Figure 3 shows an exemplary segmentation map of the image data of unaltered images alongside the segmentation map of the images with low brightness and low contrast as well as the ground truth. Looking at the prediction maps, the values look plausible.

In table 3 we can see the percentage difference of the precision values in comparison to the original data. The precision shows the proportion of actual positives that have been identified correctly. The numbers show that increasing the brightness in any case leads to massive drops in precision. Increasing the brightness and simultaneously lowering the contrast leads to a percentage change in precision of almost -80 % which leads to the assumption that these alterations represent an unfavorable combination for recognizing strong and distinctive features for segmentation purposes. Solely images with lower contrast or lower saturation in comparison to the unaltered image data prior to the prediction have shown variations under 10 % compared to the predictions generated with unaltered images.

Table 4 shows the percentage difference of the F-score of the segmentations of the altered images in comparison to the F-score of the segmentation of the unaltered images. The F-score lets us combine the precision and recall of our model



(a) Orthophoto Tile (red outline as label)

(b) Ground Truth

(c) Prediction (unaltered images)

(d) Prediction (lb, ls)

Figure 3. Comparison of the ground truth and segmentation results when predicting the segmentation maps of unaltered images (b) and pictures with lower contrast and lower saturation (c)

PRECISION						
	lb	hb	lc	hc	ls	hs
lb	-19 %	-	-	-	-	-
hb	-	-72 %	-	-	-	-
lc	-21 %	-77 %	-9 %	-	-	-
hc	-27 %	-65 %	-	-30 %	-	-
ls	-33 %	-70 %	-12 %	-31 %	-2 %	-
hs	-12 %	-60 %	-21 %	-39 %	-	-15 %

Table 3. Precision precentage values when using altered images in comparison to precision values when using original, unaltered images

and is defined as the harmonic mean of precision and recall. As already seen with precision, the F-score for segmenting the altered image data is lower overall. Especially the images with an increased brightness show F-score deviations of consistently more than 50 %. As with precision, a combination of increased brightness and decreased contrast strongly influences the F-score. At the same time, it can be seen that a reduction in saturation and contrast only leads to a decrease of 3 % and 9 %, respectively.

F-SCORE						
	lb	hb	lc	hc	ls	hs
lb	-23 %	-	-	-	-	-
hb	-	-61 %	-	-	-	-
lc	-28 %	-68 %	-9 %	-	-	-
hc	-26 %	-55 %	-	-23 %	-	-
ls	-35 %	-58 %	-13 %	-25 %	-3 %	-
hs	-16 %	-51 %	-19 %	-32 %	-	-13 %

Table 4. F-score percentage values when using altered images in comparison to F-score values when using original, unaltered images

In addition to contrast, brightness and saturation, two crop factors were examined: Images cropped by 50 % and 90 %, respectively, and resized to the original size were predicted using the model trained on the unaltered data. Table 5 shows the results, i.e. the deviations between the recall, precision, and F-score, compared to the results of the unaltered data. The numbers show that due to the changing ratio between foreground and background pixels, strong losses in precision, recall and F-score occur when the image data is enlarged by 50 %. We assume that the shifted ratio between foreground and background pixels lead to this drop. At the same time, we can see that a magnification of 10 % (90 % crop) leads to better results compared to the 50 % crop. The reason for this is the same, since in this case the ratio between background and foreground pixels is not changed dramatically.

50 % CROP			90 % CROP
Precision	-74 %		-29 %
Recall	- 46 %		+4 %
F-score	-65 %		-16 %

Table 5. Precision, recall and F-score for the predictions conducted on images 50 % and 90 % of their initial resolution

The entire experiment, beginning with the training process, was conducted five times to check for reproducibility. The deviations between the values for Accuracy, Precision, Recall and F-score were less than 11 % on average in between experiments.

5. CONCLUSION

This paper showed an approach to determine the influence of image properties on segmentation results. For this purpose, a U-Net was trained with image data that has not been altered and predictions were made on unknown images with and without altered image properties such as contrast, saturation and brightness. The quality parameters generated from the predictions of unaltered images, namely recall, precision and the F-score, were used to compare these parameters with segmentations of images with altered properties. It was found that increased brightness in particular had a strong negative effect on precision and F-score. At the same time it could be observed that for this specific approach desaturated images either do not show a large effect or, in the case of the recall, show a beneficial effect. In general, however, it could be observed that a discrepancy between the training data and the data to be predicted has a negative effect on the quality of the segmentation.

6. FUTURE WORK

The results shown in this paper are the results of testing a single dataset with a single U-Net implementation. In order to make more sustainable statements about the effects of changes in the image properties of training data or the divergence of training and prediction data, further investigations have to be performed on different implementations and datasets. In our opinion, TernausNetV2 (Iglovikov et al., 2018) and other implementations for segmentation of aerial image data like SegNet (Badrinarayanan et al., 2017) represent further interesting networks for similar tests. Another approach would be to derive properties from the difference of the training data and data to be predicted so changes can be applied to the latter. Overall, the question of the possibilities to use models trained on data with slight variations to the prediction data remains unanswered.

REFERENCES

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image

Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 2481-2495.

Darapaneni, N., Jagannathan, A., Natarajan, V., Swaminathan, G. V., Subramanian, S., Paduri, A. R., 2020. Semantic segmentation of solar pv panels and wind turbines in satellite images using u-net. *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 7–12.

Iglovikov, V., Seferbekov, S., Buslaev, A., Shvets, A., 2018. Ternausnetv2: Fully convolutional network for instance segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 228–2284.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6, 429–449. https://doi.org/10.3233/IDA-2002-6504.

Khryashchev, V., Larionov, R., Ostrovskaya, A., Semenov, A., 2019. Modification of u-net neural network in the task of multichannel satellite images segmentation. *2019 IEEE East-West Design Test Symposium (EWDTS)*, 1–4.

Kim, J. H., Lee, H., Hong, S. J., Kim, S., Park, J., Hwang, J. Y., Choi, J. P., 2019. Objects Segmentation From High-Resolution Aerial Images Using U-Net With Pyramid Pooling Layers. *IEEE Geoscience and Remote Sensing Letters*, 16(1), 115-119.

Li, L., Liang, J., Weng, M., Zhu, H., 2018. A Multiple-Feature Reuse Network to Extract Buildings from Remote Sensing Imagery. *Remote Sensing*, 10(9). https://doi.org/10.3390/rs10091350.

Li, Z., Kamnitsas, K., Glocker, B., 2021. Analyzing Overfitting Under Class Imbalance in Neural Networks for Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(3), 1065-1077.

Manilow, E., Wichern, G., Seetharaman, P., Le Roux, J., 2019. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 45–49.

Maxwell, A. E., Warner, T. A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39(9), 2784-2817. https://doi.org/10.1080/01431161.2018.1433343.

Mnih, V., 2013. Machine Learning for Aerial Image Labeling. PhD thesis, University of Toronto.

Powers, D., 2007. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness Correlation. *Journal of Machine Learning Technologies*. https://csem.flinders.edu.au/research/techreps/SIE07001.pdf.

Rahman M.A., W. Y., 2016. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. *Advances in Visual Computing*, 10072. https://doi.org/10.1007/978-3-319-50835-$1_2$2.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-$4_2$8.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., Maalouf, S., Adams, C., 2020. Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews*, 205, 103187.

Shorten, C., Khoshgoftaar, T. M., 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0.

Soekhoe, D., van der Putten, P., Plaat, A., 2016. On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. *Advances in Intelligent Data Analysis XV*, 9897. https://doi.org/10.1007/978-3-319-46349-0_5.

Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, 843–852.

Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., Coppola, G., 2019. Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging. *Computer Methods and Programs in Biomedicine*, 178, 289-301. https://www.sciencedirect.com/science/article/pii/S0169260719306145.

Wu, M., Zhang, C., Liu, J., Zhou, L., Li, X., 2019. Towards Accurate High Resolution Satellite Image Semantic Segmentation. *IEEE Access*, 7, 55609-55619. https://doi.org/10.1109/ACCESS.2019.2913442.

Zeng, Z., Xie, W., Zhang, Y., Lu, Y., 2019. RIC-Unet: An Improved Neural Network Based on Unet for Nuclei Segmentation in Histology Images. *IEEE Access*, 7, 21420-21428.

APPENDIX

50 % CROP			90 % CROP
Precision	0.189		0.511
Recall	0.340		0.656
F-score	0.233		5.554

Table 6. Precision, recall and F-score values when using images cropped by 50 % and 90 %

RECALL						
	lb	hb	lc	hc	ls	hs
lb	0.528	-	-	-	-	-
hb	-	0.693	-	-	-	-
lc	0.442	0.662	0.666	-	-	-
hc	0.567	0.635	-	0.678	-	-
ls	0.466	0.696	0.6193	0.646	0.682	-
hs	0.589	0.626	0.630	0.622	-	0.652

Table 7. Recall values in when using altered images

PRECISION						
	lb	hb	lc	hc	ls	hs
lb	0.512	-	-	-	-	-
hb	-	0.176	-	-	-	-
lc	0.500	0.145	0.572	-	-	-
hc	0.460	0.219	-	0.442	-	-
ls	0.422	0.190	0.556	0.435	0.619	-
hs	0.553	0.254	0.496	0.385	-	0.534

Table 8. Precision values in when using altered images

F-SCORE						
	lb	hb	lc	hc	ls	hs
lb	0.508	-	-	-	-	-
hb	-	0.254	-	-	-	-
lc	0.469	0.211	0.594	-	-	-
hc	0.483	0.298	-	0.507	-	-
ls	0.425	0.272	0.569	0.494	0.571	-
hs	0.553	0.324	0.534	0.449	-	0.636

Table 9. F-score values in when using altered images

(a) Orthophoto Tile with label as red outline

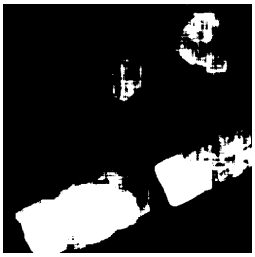
(b) Ground Truth

(c) Unaltered (Original)

(d) High Contrast, High Brightness

(e) 90 % crop

(f) Low Brightness



(g) Low Contrast

(h) Low Saturation

(i) Low Contrast, Low Saturation

(j) Low Contrast, Low Brightness

Figure 4. Exemplary depiction of segmentation maps when using altered images alongside ground truth and orthophoto