SEMANTIC PHOTOGRAMMETRY – BOOSTING IMAGE-BASED 3D RECONSTRUCTION WITH SEMANTIC LABELING

E.-K. Stathopoulou [1], F. Remondino [1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: http://3dom.fbk.eu
Email: <estathopoulou><remondino>@fbk.eu

Commission II

KEY WORDS: image-based 3D reconstruction, label transfer, semantic photogrammetry, dense image matching.

ABSTRACT:

Automatic semantic segmentation of images is becoming a very prominent research field with many promising and reliable solutions already available. Labelled images as input for the photogrammetric pipeline have enormous potential to improve the 3D reconstruction results. To support this argument, in this work we discuss the contribution of image semantic labelling towards image-based 3D reconstruction in photogrammetry. We experiment semantic information in various steps starting from feature matching to dense 3D reconstruction. Labelling in 2D is considered as an easier task in terms of data availability and algorithm maturity. However, since semantic labelling of all the images involved in the reconstruction may be a costly, laborious and time consuming task, we propose to use a deep learning architecture to automatically generate semantically segmented images. To this end, we have trained a Convolutional Neural Network (CNN) on historic building façade images that will be further enriched in the future. The first results of this study are promising, with an improved performance on the quality of the 3D reconstruction and the possibility to transfer the labelling results from 2D to 3D.

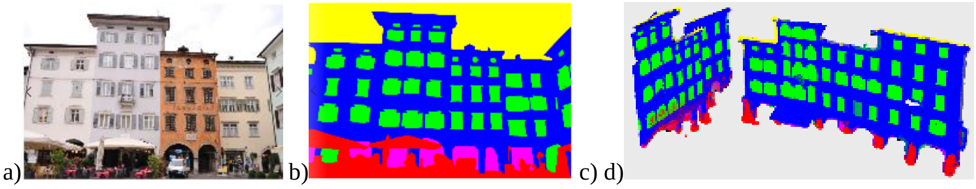


Figure 1: Boosting photogrammetric 3D reconstruction with semantic information: an input image (a), its corresponding automatic labelling in 4 classes (b) and the semantic 3D data (c) image network for 3D reconstruction (d). Semantic aids also the various photogrammetric steps, from tie point extraction to dense image matching.

1. INTRODUCTION

The image-based 3D reconstruction pipeline, based on photogrammetry and computer vision principles, has become in the last years a powerful approach for the generation of detailed and precise 3D data. It is getting commonly preferred over costly laser scanning methods in many fields, including heritage documentation. Indeed, photogrammetry can generally reassure adequate automation, high quality results and ease of use, even for non-expert users along with cost efficiency. Recent advances were achieved in all core components of the photogrammetric pipeline enhanced even more the aforementioned arguments: from the image pre-processing (Verhoeven et al., 2015) to keypoints extraction (Hartmann et al., 2015), bundle adjustment (Schoenberger and Frahm, 2016) as well as dense points clouds generation (Remondino et al., 2014).

At the same time, the latest decade has witnessed a massive usage of machine and deep learning techniques in the fields of computer vision and signal processing among the fields of data science. The availability of an extensive amount of image data along with the increasing computational power of PC but, most importantly, the use of GPU processing, has promoted the so-called deep learning techniques (Goodfellow et al., 2016) letting the data scientists to achieve unexpected good results. Machine and deep learning methods are getting more and more popular also in the photogrammetric community. Indeed, they are used for image classification, scene semantic segmentation as a pre-requisite for several tasks in robotics (e.g. object modelling and recognition, autonomous grasping and manipulation, object tracking, etc.), autonomous driving, indoor and urban modelling, etc. (Finman et al., 2014; Martinovic et al. in 2015; Marmaris et al., 2016; Tateno et al., 2017; Tsai et al., 2018). They are also used to segment 3D heritage data (Grilli et al., 2018) or remove outliers in point clouds (Stucker et al., 2018).

1.1 Aim of the paper

Since 2D semantic segmentation algorithms are robust enough and can achieve high level performance scores, the work proposes to support photogrammetric 3D reconstructions with state-of-the-art semantic image segmentation methods (Fig. 1). We propose to use semantic image labelling in various steps of the photogrammetric pipeline and also to apply label transfer from 2D to 3D. We apply our method to heritage datasets developing our own training set. In more detail, the paper suggests semantic segmentation of images in order to (Fig. 2):

- Constrain the search area of image features when extracting tie points among images in order to minimize mismatched correspondences - only features under the same label should be matched.
- Use semantic labelling to mask out areas that should not be considered during the generation of the dense point cloud (e.g. the sky).
- Reinforce dense image matching using pixel labels and reduce noisy 3D points.

- Label transferring from the image to the 3D data.

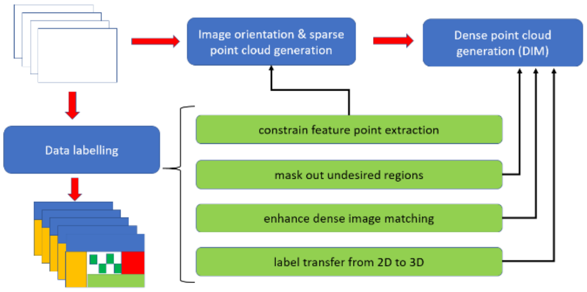


Image orientation & sparse point cloud generation
Dense point cloud generation (DIM)
Data labelling
constrain feature point extraction
mask out undesired regions
enhance dense image matching
label transfer from 2D to 3D

Figure 2: The proposed semantic photogrammetric pipeline where semantic segmentation using deep learning methods are implemented to support 3D results.

To reach these goals, we use a Convolution Neural Network (CNN) trained on our own historic façades dataset (Fig. 3) to facilitate automatic semantic labelling of similar image sets. The images depict various architectural scenes, namely facades of historical buildings of several towns in Italy. The data was carefully selected in order to cover a large variety of historic buildings.

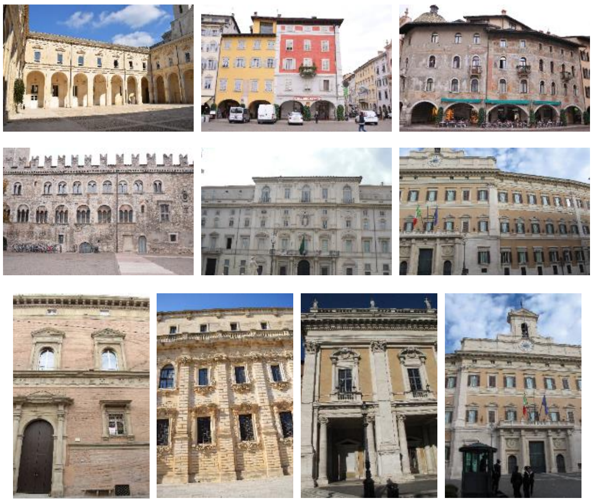


Figure 3: Example of the historic building façades used in our tests to train a CNN for automatic semantic segmentation.

In the next Sections, the related work on semantic segmentation methods is discussed (Section 2), followed by the description of our training method (Section 3). Section 4 presents our experiments towards boosting image-based 3D reconstruction by exploiting semantic information within the processing pipeline. Section 5 draws some final conclusions and our vision for future works.

2. RELATED WORK

The recent technological advances in computation power led to a popularization of deep learning methods and their broad use in image processing and scene understanding. Image classification refers to the assignment of a category label to an image, commonly based on its most salient objects. On the other hand, image segmentation refers to the assignment of a predicted label for each single image pixel in a semantic meaningful way.

State-of-the-art approaches in image classification and segmentation as well as object detection benefit from the recent advances in Convolutional Neural Networks (CNNs) which tend to outperform other methods in efficiency and accuracy (Simonyan et al., 2014; Long et al., 2015; Girshick et al., 2015; He et al., 2017). The concept of CNNs was originally introduced in the 1980s (Fukoshima et al., 1982), but gained popularity because of the recently achieved results published to the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) challenge (Deng et al., 2009; Russakovsky et al., 2015), one of the pioneer efforts to collect an extensive dataset harvested from the web and train an image classification deep CNN. The most famous CNN architectures probably are AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan et al., 2014), ResNet (He et al., 2016), Inception (Szegedy et al., 2015). As an extension of the standard CNN, the so-called Fully Convolutional Network (FCN) (Chen et al., 2015) are adapted in order to deal with arbitrary sized images and are commonly used for semantic segmentation problems (Long et al., 2015). To train these networks, a large amount of dense pixel annotations must be collected and used as training data. Like all machine learning methods, training with a poor amount of annotated data leads to models that may not generalize well to all test images and cases. Semantic annotation of objects and classes is performed on single images or video sequences. An extensive bibliography on semantic segmentation of sequences exist, especially on urban street scenes towards complete scene understanding mainly for autonomous driving purposes. To this end, several benchmark datasets are introduced for training and testing like CamVid (Brostow et al., 2009), Rue-Monge2014 (Riemenschneider et al., 2014) or CityScapes (Cordts et al., 2016). However, to the best of our knowledge, no dataset is available to specifically tackle heritage and architectural scenarios, including historical façades.

3. DATA LABELLING AND NETWORK TRAINING

3.1 Image annotation and labelling

Labelling ground truth data (either simple boxes or more complex shapes) is something that still remains unsolved in supervised learning, due to the complexity of the task and the fact that it is time consuming. Up to now, it is an indispensable and laborious stage that cannot be automated but necessary to train the deep learning model. Moreover, the labeller should pay large attention to avoid gross errors that would affect the training quality. The labouring time can be significantly reduced by using specially designed interfaces for annotation (e.g. RectLabel, Labelbox, LabelMe) whereas research is going on towards efficient object annotation (Papadopoulos et al., 2017). Recently, the high demand for data labelling gave rise to specialized services, such as Edgecase, HumanInThe Loop, Raidon, etc. Crowdsourcing services are also used for creating large annotated datasets (e.g. MTurk, Clickworker) providing relatively fast results under low cost although they could be risky in terms of low quality and inconsistency. Weak and semi-supervised approaches were also proposed to reduce the heavy labelling cost of collecting segmentation ground truths (Khoreva et al., 2017), while synthetic semantically annotated data have been also introduced for urban scenes (Ros et al., 2016).

For our experiments, ground truth data was created by manually labelling the available images (Fig. 4). Up to now, 5 classes were used to semantically segment the scenes, namely “building”, “sky”, “obstacle”, “window” and “door”. The classes were chosen as such to boost the second objective of this study, i.e. the photogrammetric 3D reconstruction. Thus, classes that are not contributing positively to the 3D reconstruction could be masked out, or, in other words, the reconstruction will be implemented for just the areas of interest among our scene.

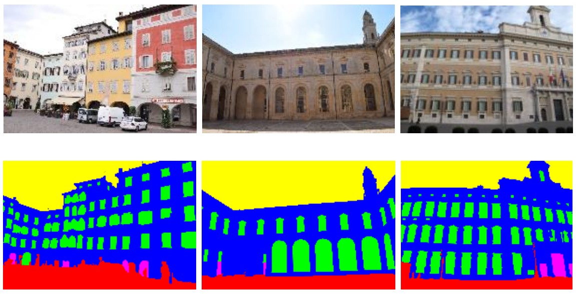


Figure 4: Examples of our training dataset (above) and the corresponding manually labelled images (below). The classes correspond to the following colors: "sky"=yellow, "building" = blue, "window"=green and "obstacle"=red.

3.2 Deep learning model and training

Various implementations of deep learning models exist and are available as open-source libraries and frameworks (Tensorflow, Caffe, PyCharm etc.). There exist also several tools particularly designed to tackle specific challenges e.g. semantic segmentation. Our pipeline is built upon the Semantic Segmentation Suite tool based on Tensorflow, known for its good performance on the CamVid dataset (Brostow et al., 2009). Among all the state-of-the-art models, we used a fully convolutional FC-DenseNet (Jégou et al., 2017) with 56 layers on a GPU Tensorflow implementation. The network was trained over 300 epochs using a learning rate of 0.00001 on our ground truth dataset. The learning rate value was adjusted regarding the needs of the project and selected as the best performing rate within several attempts (visual inspection of the tested images – Fig.5).

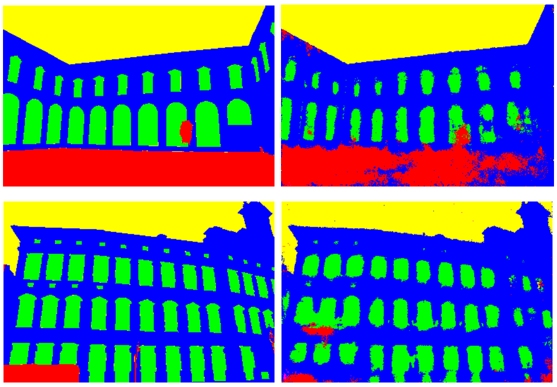


Figure 5: Ground truth labels (left) and testing (right) results.

During the testing phase, we evaluated the accuracy of the tested images with respect to their ground truth (Table 1). Considering that our training set is not extensive and still expanding and that we are considering a large diversity of architectural scenes (and thus objects that belong to the same class), the achieved values are considered satisfying.

Measure	*Average value*
Accuracy	0.81
Precision	0.87
Recall	0.81
F1 score	0.83

Table 1: Statistics of the testing phase.

The resulting pixel annotations are used as input in the 3D reconstruction pipeline providing the additional information needed to undertake the aforementioned subtasks (Section 1.1).

4. BOOSTING PHOTOGRAMMETRY WITH SEMANTIC INFORMATION - EXPERIMENTS AND DISCUSSION

The ultimate goal of the work is to support the photogrammetric pipeline with semantic information, till the generation of semantically enriched 3D point clouds (Fig. 2). Thus, 2D images with available labelled data are used as input in our 3D reconstruction pipeline. The labelled data could be ground truth, if available, or test data resulting from our trained architecture (Section 3.2).

4.1 Constrained tie point extraction

Given a labelled set of images, each putative tie point on the master image is to be matched only within a subset of points on the search image that have the same labelling information. With such a "constrained matching", certain false correspondences can be avoided and the final list of tie points is more reliable. In our pipeline, we used the ORB (Rublee et al., 2011) feature detector and descriptor coupled with brute force matching (OpenCV library) guided by the labelling information. The matches were then filtered by comparing the closest match to the second closest based on the ratio test criterion (Lowe, 2004), the so-called "good matches". Rejecting all matches in which the distance ratio is greater than 0.65, we perform the experiments on the image pairs while constraining the search areas based on the labelling information. Examples results of the constrained matching are shown in Figure 6 for label "window" (Figure 6a-b) and "building (Figure 6c).

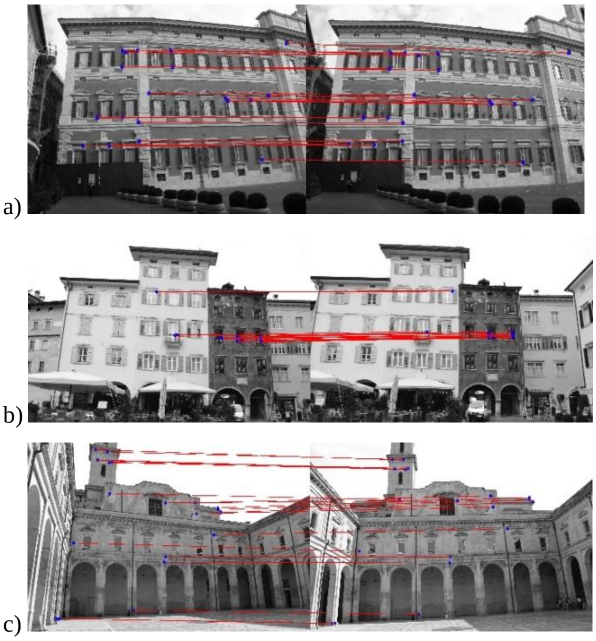


Figure 6: Constrained feature matching for tie points extraction based on the feature's label. For a clearer visualisation, only few features are drawn on grey scale images.

The number of good matches is compared to the respective number of good matches while no labelling constrain is applied.

According to these experiments, constrained matching resulted to 10% -20% less good matches.

4.2 Automatic masking for dense image matching

Using the deep learning network results, undesired parts of the scene (e.g. sky, occlusions, etc.) can be automatically excluded from the dense image matching (DIM) procedure. This procedure is normally called "masking", it is generally manually performed and time consuming. The masking of undesired areas is performed to remove areas that would degrade the dense point clouds adding unwanted or noisy points.

Figure 7 shows an example where the sky in background and in between the columns would produce outliers close to the temple borders. Semantically classified images allow to automatically mask the unwanted areas ("sky" class), robustly eliminating large part of the outliers and the noise close to the edges (white/blue points in Figure 8a). The number of removed points was around 5% of the total reconstructed 3D points. Using this approach a large part of the post-processing/editing of the dense point cloud (manual or automatic filtering) can be avoided.

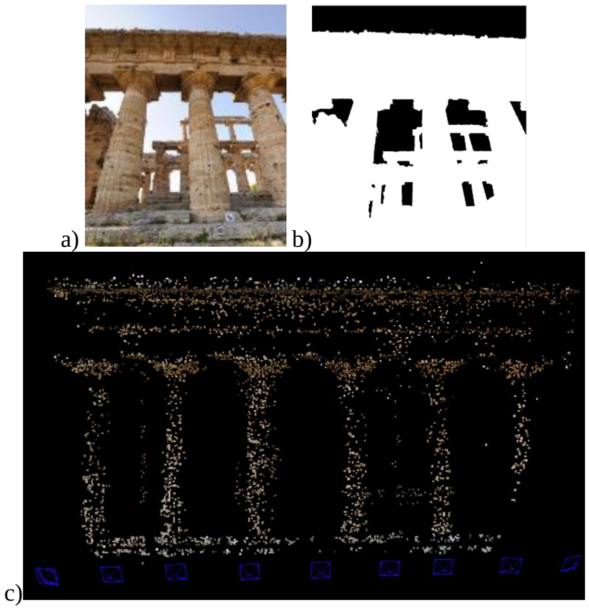
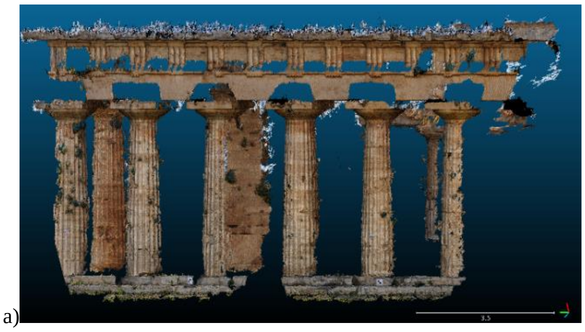


Figure 7: An image of the considered dataset (a), its corresponding automatically generated semantic mask in binary format (b) and the orientation results (c).



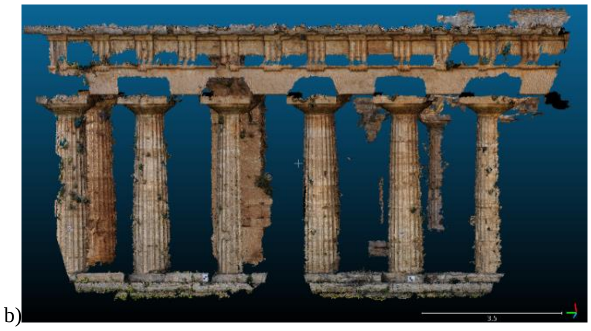


Figure 8: Dense 3D reconstruction of part of the temple without semantic masking (a) versus dense 3D reconstruction with semantic masking and exclusion of the sky category (b).

4.3 Label transfer from 2D images to 3D data

The semantic enrichment of 3D point clouds can be performed either on the sparse or on the dense point cloud.

The sparse point cloud generated during the bundle adjustment procedure can be expressed, for every camera, using the projection matrix P which connects the 3D with the 2D space. Thus, giving the correspondences between an image point x and its projection to the object space X, a semantic label can be assigned to each 3D point based on the label of its back-projection to the 2D image (Fig. 9). All images that contribute to a 3D point are considered for label contribution. If the assigned labels to each back-projected point do not match, the most weighted label wins. In case of inconsistent labels with the same weight, the pixel is considered as unlabeled and is subject to further investigation (white points in Figure 9c).

The quality of the sparse point cloud is fundamental for the success of this procedure. Outliers and noisy points are not projected correctly on the images and, consequently, they are assigned to erroneous or inconsistent labels, thus producing an undesired result with randomly assigned labels. In our experiments, points with an ambiguous label were below the 2% of the total number of points, so this issue is considered to be of minor importance.

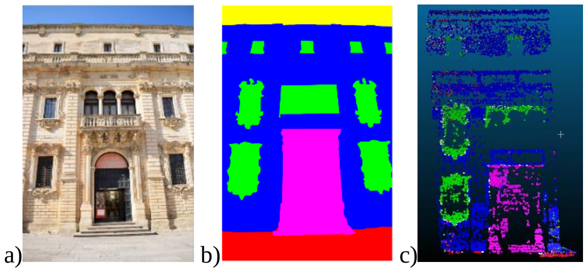


Figure 9: Image (a), its semantic segmentation (b) and the labelled sparse point cloud after label transfer (c).

In a similar fashion, the projection of each pixel label to the dense cloud will result to a labelled dense cloud. If the information about image contribution on each 3D point is available, the winner label can be projected onto the dense point cloud given the image orientation. However, this procedure is not straight forward since most of the MVS algorithms may apply some filtering or interpolation between the dense 3D points and thus, the back-projection may lead to wrong traces giving misleading label information.

The labelling can also be performed on mesh 3D models, whereas each pixel label is projected to the mesh nodes and interpolated to the faces.

5. CONCLUSIONS AND FUTURE WORK

The paper presented some initial results of using semantic labels to boost the photogrammetric processing of terrestrial datasets. Semantically segmented images are automatically generated using deep learning methods (CNN). Then labels are used as constraints in the photogrammetric processing or transferred to the generated 3D information. The initial results are promising and the developed methods will be integrated in our web-based pipeline (Tefera et al., 2018) to extend its functionalities and potentially boost the reconstruction results. Similarly to feature matching constraint, dense image matching can be constrained based on the semantic information provided for each pixel.

The training dataset will be extended in order to have a larger input of our CNN method and optimize its accuracy. More classes will be annotated as there is no such a training dataset in the heritage community. The dataset, as well as our pre-trained network, will be available to the public in the near future to facilitate research towards this direction.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Matteo Rapana’ for the manual labelling of the employed images.

REFERENCES

Brostow, G.J., Fauqueur, J., Cipolla, R., 2009. Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, Vol. 30(2), pp.88-97.

Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R., 2008. Segmentation and recognition using structure from motion point clouds. Proc. ECCV, pp. 44-57.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2015: Deeplab: Semantic image segmentation with deep convolutional nets and fully connected CRFs. Proc. ICLR

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. Proc. CVPR, pp. 3213-3223.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.

Finman, R., Whelan, T., Kaess, M., Leonard, J. J., 2014. Efficient incremental map segmentation in dense RGB-D maps. Proc. Int. Conf. on Robotics and Automation (ICRA)

Fukushima, K. and Miyake, S., 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and cooperation in neural nets, pp. 267-285, Springer, Berlin, Heidelberg.

Girshick, R., 2015. Fast R-CNN. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., 2016. Deep learning (Vol. 1). Cambridge: MIT press.

Grilli, E., Dininno, D., Petrucci, G., and Remondino, F., 2018. From 2D to 3D supervised segmentation and classification for Cultural Heritage Applications. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2, pp. 399-406

Hartmann, W., Havlena, M., Schindler, K., 2015. Recent developments in large-scale tie-point matching. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 155, pp. 47-62.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. CVPR, pp. 770-778.

He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017, October. Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 2980-2988). IEEE.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation. Proc. CVPR, pp. 1175-1183.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B., 2017. Simple does it: Weakly supervised instance and semantic segmentation. Proc. CVPR

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Proc. NIPS, pp. 1097-1105.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proc. CVPR, pp. 3431-3440.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, Vol. 60(2), pp.91-110.

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 Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 3, p.473.

Martinovic, A., Knopp, J., Riemenschneider, H., Van Gool, L., 2015. 3D all the way: Semantic segmentation of urban scenes from start to end in 3D. Proc. CVPR, pp. 4456-4465.

Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V., 2017. Extreme clicking for efficient object annotation. Proc. ICCV, pp. 4940-4949.

Remondino, F., Spera, M.G., Nocerino, E., Menna, F., Nex, F., 2014. State of the art in high density image matching. The Photogrammetric Record, Vol. 29, pp. 144-166

Riemenschneider, H., Bódis-Szomorú, A., Weissenberg, J. and Van Gool, L., 2014. Learning where to classify in multi-view semantic segmentation. Proc. ECCV, pp. 516-532.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D. and Lopez, A.M., 2016. The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. Proc. CVPR, pp. 3234-3243.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. Proc. ICCV, pp. 2564-2571.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. Int. Journal of Computer Vision, Vol. 115(3), pp.211-252.

Schoenberger, J.-L., Frahm, J.-M., 2016. Structure-from-motion revisited. Proc. CVPR

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Proc. CVPR, pp. 1-9.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Stucker, C., Richard, A., Wegner, J. D., Schindler, K., 2018. Supervised outlier detection in large-scale MVS point clouds for 3D city modeling applications. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. IV-2, pp. 263-270.

Tateno, K., Tombari, F., Navab, N., 2017. Large scale and long standing simultaneous reconstruction and segmentation. Journal CVIU.

Tefera, Y., Poiesi, F., Morabito, D., Remondino, F., Nocerino, E., Chippendale, P., 2018. 3DNOW: Image-based 3D reconstruction and modeling via web. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2, pp. 1097-1103

Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M., 2018. Learning to Adapt Structured Output Space for Semantic Segmentation. Proc. CVPR

Verhoeven, G., Karel, W., Stuhec, S., Doneus, M., Trinks, I., Pfeifer, N., 2015. Mind your grey tones - examining the influence of decolourization methods on interest point extraction and matching for architectural image-based modelling. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. 40(5/W4), pp. 307-314.

Web references

https://rectlabel.com/
https://www.labelbox.com/
http://labelme.csail.mit.edu/Release3.0/
https://www.edgecase.ai/
https://humansintheloop.org/
https://raidon.io/
https://www.mturk.com/
https://www.clickworker.com/about-us/
https://www.tensorflow.org/
http://caffe.berkeleyvision.org/
https://pytorch.org
https://opencv.org/

https://doi.org/10.5194/isprs-archives-XLII-2-W9-685-2019 |