DEPARTMENT OF COMPUTER SCIENCE
AND MATHEMATICS

UNIVERSITY OF APPLIED SCIENCES MUNICH

Master's Thesis in Computer Science

On the generalization capabilities of interactive segmentation methods

Alexander Fertig

DEPARTMENT OF COMPUTER SCIENCE
AND MATHEMATICS

UNIVERSITY OF APPLIED SCIENCES MUNICH

Master's Thesis in Computer Science

On the generalization capabilities of interactive segmentation methods

Author:	Alexander Fertig
Supervisor:	Prof. Dr. David Spieler
Advisor:	Advisor
Submission Date:	Submission date

I confirm that this master's thesis in computer science is my own work and I have documented all sources and material used.

Munich, Submission date Alexander Fertig

Acknowledgments

Abstract

1. Introductions
    a) DL in Industry
    b) Application of DL and gathering Labels
2. Basics
    a) ML, Dl, CNN
    b) Semantic Segmentation (and IoU)
    c) Interactive Semantic Segmentation (Methods of comparison)
3. Methods
    a) Extreme Points
    b) IOG
4. Benchmark
    a) Motivation and structure of the Benchmark
    b) Applied Methods
    c) Evaluation (or put Evaluation as own chapter)
5. Conclusion

Contents

1 Introduction

1.1 Section

1.1.1 Subsection

See Table 1.1, Figure 1.1, Figure 2.2, Figure 1.3.

Table 1.1: An example for a simple table.

A	B	C	D
1	2	1	2
2	3	2	3

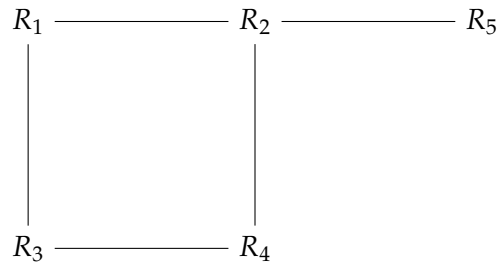


Figure 1.1: An example for a simple drawing.

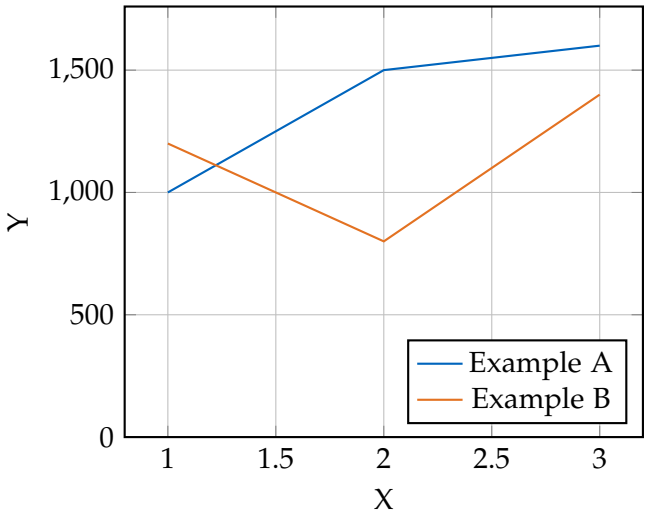


Figure 1.2: An example for a simple plot.

```
SELECT * FROM tbl WHERE tbl.str = "str"
```

Figure 1.3: An example for a source code listing.

2 Basics

2.1 ML, DL, CNNs

The last decade was revolutionary for the sector of information technology. Due to technical advancement, the computational power of processors, especially Graphical Processing Units (GPU) rise significantly. Further, the wider creation and use of data introduced the domain of big data, that allows operators to gain more insights and benefits. Both of these recent advancements benefited another field of study commonly described as Artificial Intelligence (AI). The term AI describes machines that are show characteristics of human intelligence that allow them to handle various tasks. But there are several gradations, that hide behind the powerful bus word AI and are illustrated in this section.

2.1.1 Machine Learning

Machine Learning (ML) is a subsection of AI and is based on the ability of algorithms to analyze, detect and learn patterns in various kinds of data. Applying these learned pattern ML algorithms may reach human level performance or even better, but they are limited specifically to their scope. For other tasks mostly a new ML algorithms needs to be defined [18b]. In contrast AI may be able to solve various tasks and learn independently by showing strong characteristics of human intelligence.

In ML

2.1.2 Deep Learning

2.1.3 Convolutional Neural Networks

Table 2.1: An example for a simple table.

A	B	C	D
1	2	1	2
2	3	2	3

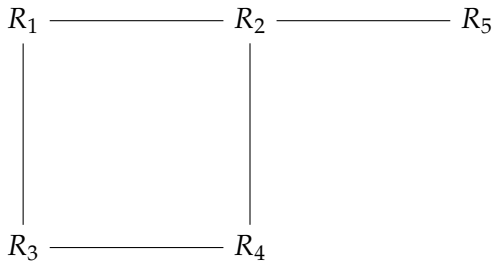


Figure 2.1: An example for a simple drawing.

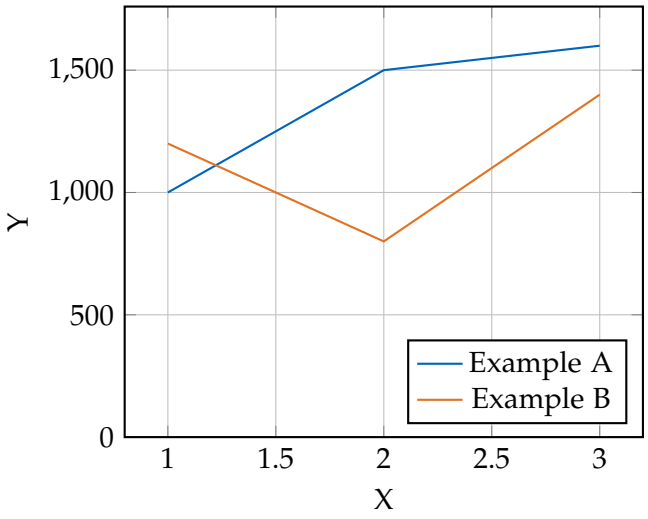


Figure 2.2: An example for a simple plot.

```
SELECT * FROM tbl WHERE tbl.str = "str"
```

Figure 2.3: An example for a source code listing.

2.2 Semantic Image Segmentation

2.2.1 General

Semantic Image Segmentation is an advanced task of modern computer vision. In general, segmentation means to obtain regions or structures from an image and portioning the image into segments. So instead of processing the image, this approach aims to achieve a high level understanding of the image. In the scope of semantic Image Segmentation, the segmentation is performed on pixel level trough the classification of each pixel in an image. Pixels with a common class form a segment.

In order to fulfill this target modern techniques of Deep Learning have proven themselves to be most adequate for this task. Usually deep convolution networks are applied to process the input image and output a segmentation results. As this forms a problem of supervised learning, a dataset with labels on pixel-level is required.

Further, there also exists the Semantic Instance Segmentation, which not only aims to predict one class, but also several instances of one class.

2.2.2 Evaluation Metric

To ensure an objective comparison of several methods a evaluation metric is required, which incorporates the basic idea of semantic segmentation. As this challenge is an classification task on pixel-level, a measure of evaluation is the Overall Pixel (OP) accuracy, which represents the proportion of all correctly labeled pixels in an image. Further, the OP measurement can be refined by calculating the accuracy for each class. This results in the Per-Class (PC) accuracy, which represents the proportion of correctly labeled pixels of one class. The most commonly used evaluation metric is the Jaccard Index, also known as the Intersection over Union (IoU), which is used in the PASCAL VOC challenge [Eve+10] since 2008 [CL13]. The IoU measures the ratio of overlap between GT and prediction (true positives) and of the total area. It is defined as

$$IoU = \frac{\text{true positives}}{\text{true positives} + \text{false negatives} + \text{ false positves}} \tag{2.1}$$

and is calculated for each instance or semantic class. To evaluate all instances or classes of an image or a dataset the IoU is averaged, which results in the *mean Intersection over Union* (mIoU) [21] [Fer19].

An advantage is the inclusion of *false positives* and *false negatives* into the IoU calculation. A limitation of the IoU metric is that the correctness of the segments boundaries is not taken into account, as discussed in [CL13]. For this issue they suggest to combine the IoU with another complementary metric, evaluating the boundary of a segment. Regardless the IoU is a suitable and commonly used metric for semantic segmentation.

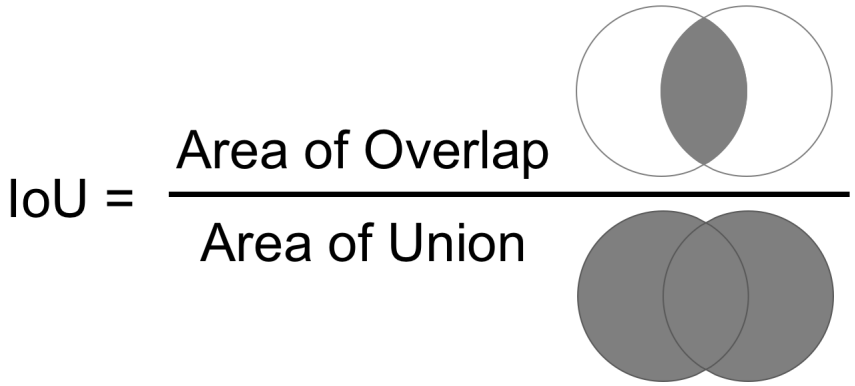


Figure 2.4: Intersection over Union. The *area of overlap* represents the intersection of the GT with the made prediction. The *area of union* represents the amount of the total area of GT and prediction [SM18].

2.2.3 Architecture

The established network architectures for the task of image classification follow a common scheme: An input image with lots of features is processed and continuously downsized to make only one prediction. In contrast, for semantic image segmentation a class is predicted for each pixel of the image. Therefore some kind of enlargement in the architecture is required to enable the model to make that high amount of predictions. In the following important architectures and their characteristics are examined.

Encoder-Decoder-Architecture

The Encoder-Decoder-Architecture as its name anticipates is based on two main parts: the encoder network and the decoder network, visualized in Figure 2.5.

The encoder network is very similar to a CNN. It consists out of convolution and pooling layers, that reduce the size of the feature maps and extract features. The encoder networks of the DeConvNet [NHH15] and the SegNet [BKC17] even mostly include parts of a popular CNN, the VGG-16 [SZ15]. In this context the process of applying the encoder network is also called *downsampling*, due to the size reduction of the feature maps.

The decoder network is the counterpart of the encoder network. It reconstructs the feature maps to their original size, which is also referred to as *upsampling*. To reach this original size often a reversed architecture of the encoder network is used. The elemental components of this reconstruction are the operations *unpooling* and *transpose convolution*.

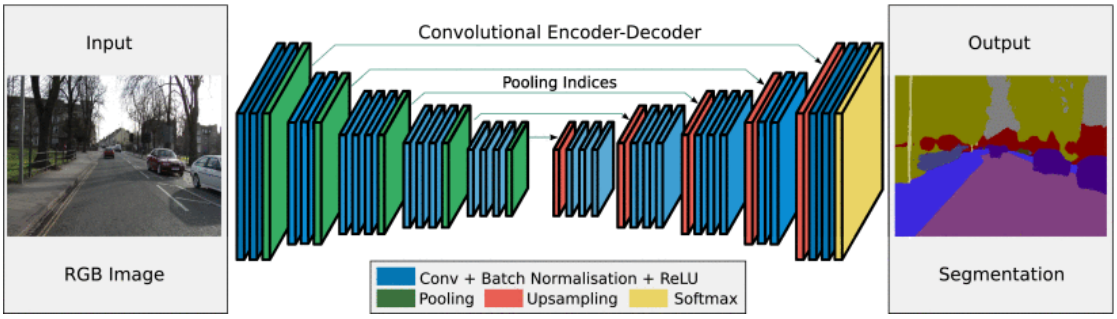


Figure 2.5: Encoder-Decoder-Architecture [BKC17]. On the left the encoder network, which reduces the size of the feature maps while processing. On the right is the decoder network, which reconstructs the feature map to the size of the original input. The yellow layer on the very right is the classification layer, here represented as softmax layer to create the output segmentation.

After the encoder network a generally a softmax classifier is applied, that creates a prediction in the form of a probability map.

Representatives of the encoder-decoder-architecture are among others the U-Net [RFB15], the DeConvNet [NHH15] and the SegNet [BKC17]. [SZ15]

Unpooling [18a] [NHH15]

Transpose Convolution [DV18]

Fully Convolutional Networks

[LSD15]

CRF [Che+16] [KK12]

Atrous Spatial Pyramid Pooling

Lateral connections

[RFB15]

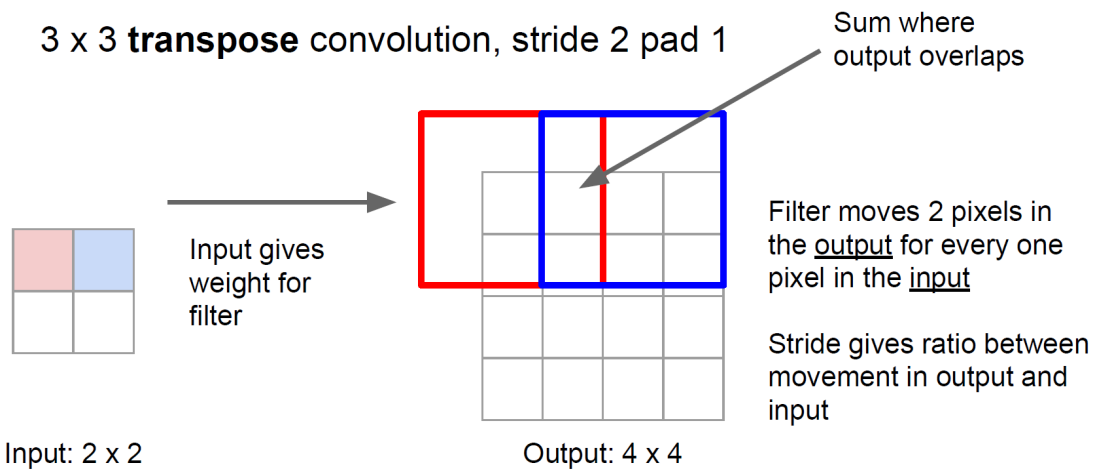


Figure 2.6: tbd

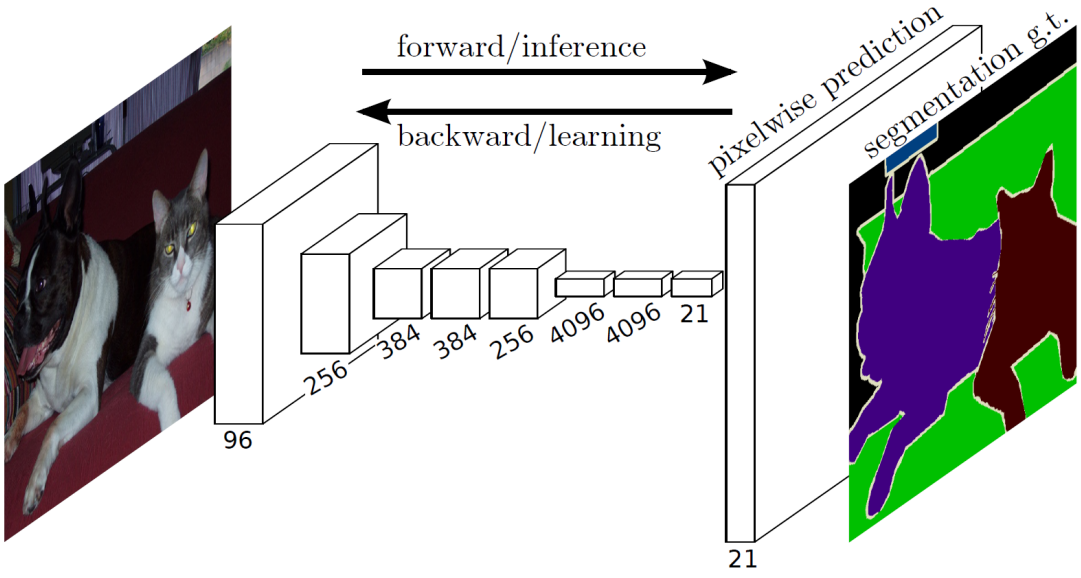


Figure 2.7: tbd

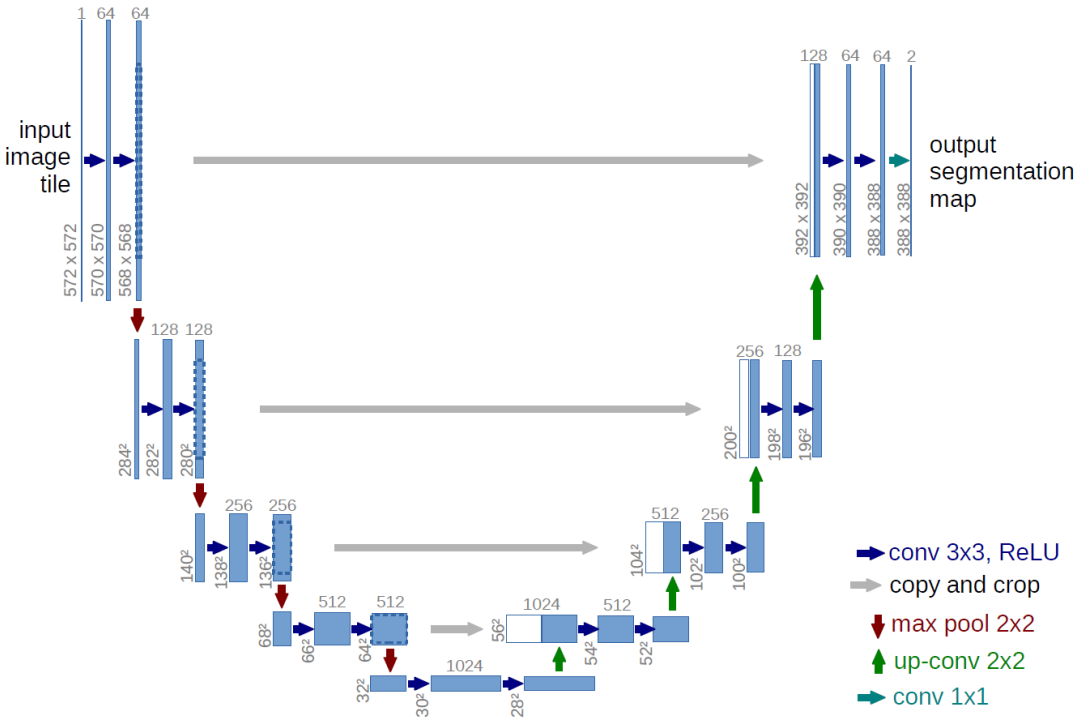


Figure 2.8: tbd

2.2.4 Data

As semantic segmentation is a problem of supervised learning it requires labeled data (GT). For a dataset to be suitable in the field of DL among others the following criteria should be met: Quantity, quality and representation capabilities.

- The quantity of a dataset used for training a DL model is crucial for its success. In general, small datasets, may not cover all vital characteristics to completely map a given objective. It has been shown in [BB01], that the performance of networks can improve significantly using a larger dataset for training. Also, in [HNP09] the effect of larger datasets is examined. It is claimed that, using a larger dataset for training can benefit the networks performance more than modifying the architecture of the network [Gér17] [Fer19]. This highlights the importance of datasets with sufficient quantity to increase the performance of networks.

- The performance of a model highly depends on the quality of the training data. Data, that is inconsistent, incomplete, erroneous or too noise, can lead to significant decrease in performance [GAD17]. Training with poor quality data makes it more difficult for a model to detect and understand the elemental features and patterns, that are required by a model to perform well [Gér17].

- Another elemental characteristic of a dataset is the representation of the problem, that the corresponding network should solve. To enable a model to generalize and perform well, it is essential for the training data to be representative to the problem [Gér17]. The best approach to do so, is to include samples of this specific problem or of samples from the same domain. But instead, often general 'all-use-datasets', like Pascal VOC [Eve+10], COCO [Lin+14] or ImageNet [Den+09], are applied on a specific problem, that is not covered within the samples of these datasets. This may results in a decrease of performance, because the capabilities of DL model are strongly connected with the representation of the data [GBC16].

It can be a challenge to obtain a dataset, that meets these criteria. The creation of new image datasets can be considered very expensive in time and cost. Datasets for semantic segmentation are even more difficult to create due to the high effort required to label images on pixel-level. Especially, datasets that cover uncommon or even restricted domains (e.g. medical or industrial domain) are rare to find. For example, the manufacturing process in a closed industrial environment may contain unique objects or uncommon surroundings, that are hardly ever represented in common datasets.

To facilitate the process of creating a new dataset and label images with pixel-level accuracy, new approaches have been created. An efficient and common way is a program, that simplifies the process of labeling by providing an user interface and

multiple methods to create and save label. These programs are often called *Labeltools* or *Annotation tools* and due to the high demand on labeled training data there are various Labeltools available [20]. To simplify the quite manual labeling process for a human user there are interactive methods (see Chapter 2.3), that support the applicant to create a label. Another approach is to create synthetic datasets like the SYNTHIA dataset [Zol+19] and use them to as training data for semantic segmentation [Che+18].

2.2.5 State-of-the-art

2.2.6 Application

Semantic Segmentation finds application in various tasks and is widely used over different domains. Due to its capability to perform classification on pixel-level it is applied on scene understanding [LSL09] or the evaluation of satellite images [Li+18]. In the field of autonomous driving Semantic Segmentation is used for street scene analysis [Cor+16] [MG15] [Neu+17]. In medicine this method can be used to segment blood cells [Tra+18]. Or it is applied in order to fulfill abstract tasks like the reconstruction of indoor scenes [Dai+17]. This listing of only some applications gives an idea of how versatile and functional Semantic Segmentation is and what can be achieved with it in the future.

2.3 Interactive Semantic Segmentation

While semantic segmentation performs the task of segmenting an image just with the image itself, Interactive Semantic Segmentation takes advantage of additional information interactively provided by an user. The idea of this concept is to enhance the segmentation result by adding a new sort of information, that is already processed by an user. Because of this, the user input has great value for the network and provides high level guidance for the task of segmentation. Depending on the type of interaction, the receipt of the user input may be more or less elaborately, which leads to a weighing of the advantages and disadvantages. On the one side interactively provided user input contains high level information, but on the other side user interactions may be very expensive, especially in the context of the big amount of images in datasets required for deep learning tasks. In the following basic concepts of interactive semantic segmentation are introduced by presenting specific methods.

2.3.1 Subsection

2.3.2 Points from users

A common and well known practice to obtain user input dates back to 1985, when Apple with co-founder Steve Jobs introduced the **Xerox**-computer with the first stage of development of the computer mouse.

Multiple methods use user-clicks on specific characteristics o

3 Methods

3.1 Deep Extreme Cut

The paper "Deep Extreme Cut: From Extreme Points to Object Segmentation" from Manisis et al. published in 2018 introduces another method to perform interactive object segmentation [Man+18].

3.1.1 Method Description

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

3.1.2 Architecture

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

3.1.3 Refinement

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no

sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

3.1.4 Results

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

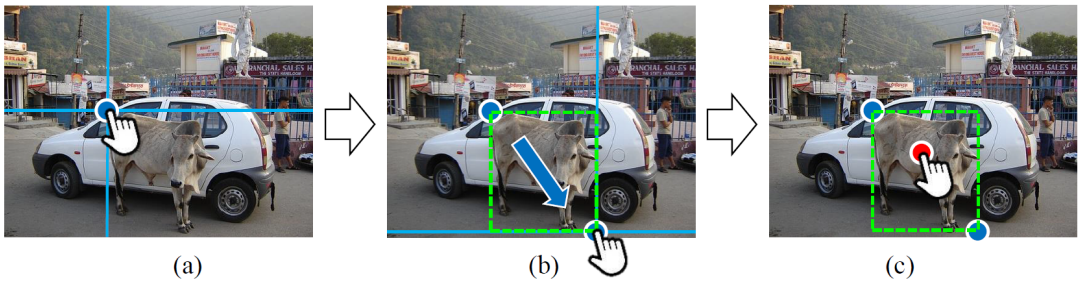
3.2 IOG

The paper "Interactive Object Segmentation with Inside-Outside-Guidance"[Zha+20] published by S. Zhang, Liew, Wei, et al. *et al*. in 2020 provides a state-of-the-art method to perform interactive object segmentation.

3.2.1 Method Description

The execution of this method outputs a binary segmentation for a single object of interest within an image. To segment multiple objects in one image, the method has t be applied for each of them sequentially.

IOG is an interactive segmentation method and hence requires user input. The input is given by a three mouse clicks on the object's foreground and on its background. The procedure is shown in Figure 3.1 and described in the following: first, in order to form an *"almost-tight bounding box"*[Zha+20, p. 12235] two exterior clicks are set at the two diagonal locations corners of the object (top-left and bottom-right or bottom-left and top-right). Based on these two points the other two corner points are derived, which leads to four points on the background. Second, to define the object inside the bounding box a single click around the center of the desired object is made, this click is processed as foreground point. The background points *"provide "outside" guidance (indicating the background regions) while the interior click gives an "inside" guidance (indicating the foreground region), thus giving the name Inside-Outside-Guidance"*[Zha+20, p. 12235].



(a) (b) (c)

Figure 3.1: Procedure of setting the three IOG clicks [Zha+20]. Set the two background clicks (blue) at the diagonal corner locations of the object. Gather a bounding box based on the background clicks. Set a foreground point (red) at the middle of the object.

These three points are preprocessed before they are input to the actual model. To include context from the surrounding region the bounding box is enlarged by p_{box} pixels. In order to focus on the object of interest the enlarged bounding box is cropped and resized to the size of 512×512 px. For background and foreground points, a separate heatmap is created by centering a 2D Gaussian at each point with

$$Gauss = \frac{\exp -4 * \log 2}{\sigma^2} \tag{3.1}$$

The two heatmaps have the size of 512×512 px and are concatenated with the input RGB image to create a 5-channel input for the model.

3.2.2 Architecture

The architecture of the IOG method is based on a *"coarse-to-fine design"*[Zha+20, p. 12237] (see Figure 3.2), containing two main parts: the CoarseNet and the FineNet.

CoarseNet The CoarseNet contains the heavy encoder part, that mainly consists of a classifier often referred to as backbone. In IOG a ResNet-101 [He+16] is used . This ResNet-101 is implemented without the head of fully connected layers. It contains four ResNet blocks and the fourth block outputs 2048 feature maps of the size 32×32 px. After the backbone a PSP-network is applied in order to enrich "the representation with global contextual information"[Zha+20]. The coarse prediction from the PSP-Network [Zha+17] has a spatial dimension of 32×32 px with 512 feature maps. From this onward the layers are enlarged by a four staged upsampling process to obtain the original input size of 512×512 px. During the upsampling process activations from the

residual parts of the ResNet are transferred from the ResNet using so lateral connections and concatenated with the upsampled feature maps. A benefit of this architecture is the fusion of information from different network stages.

FineNet The FineNet is based on a "multi-scale fusion structure"[Zha+20]. The activations from all four stages of the upsampling process from the CoarseNet are further processed along different paths. Depending on the spatial dimension, a number of additional convolution and upsampling operations are applied in order to use *"features at deeper layers for better trade-off between accuracy and efficiency"* [Zha+20, p. 12237]. These different paths are concatenated to create the networks final layer. A sigmoid is applied to this final layer, which results in a probability map as final prediction of the IOG network. The author shows in an ablation study, that the FineNet enhances the networks IoU by 0.8%. The ablation study is performed with a ResNet-50 as backbone and PASCAL-1k [Eve+10] as dataset.
This architecture especially performs well due to its application of lateral connections from different levels in order to recover local detail. The combination of layers with high localization detail with the layers, that contain high detection details, is helpful to prevent a information loss during the down- and upsampling process.

3.2.3 Refinement

If a segmentation results does not meet the user's expectations a refinement can be performed iteratively. This is done by an additional user click, which can be a fore- or background click on the region with the greatest error. In the refinement iteration of the model, this new point is processed in the same way as the initial user click positions to create a heatmap for fore- and background. These two heatmaps are combined into a two-channel input, which is processed in a so called "lightweight-branch". In this branch five convolution operations are applied and the result is concatenated with the ResNet's output of the first iteration. Hence, the ResNet does not require another execution and leads to a fast refinement process. Further, the normal IOG process is executed from the PSP-module. Zhang states that the usage of the lightweight-branch performs better than adding the refinement click into the normal 5-channel input.

In their experiments Zhang compares the IOG method to other state-of-the-art methods on different benchmarks, as shown in Figure They also evaluate the generalization abilities of IOG on unseen classes. Zhang claims that IOG outperforms all other methods.

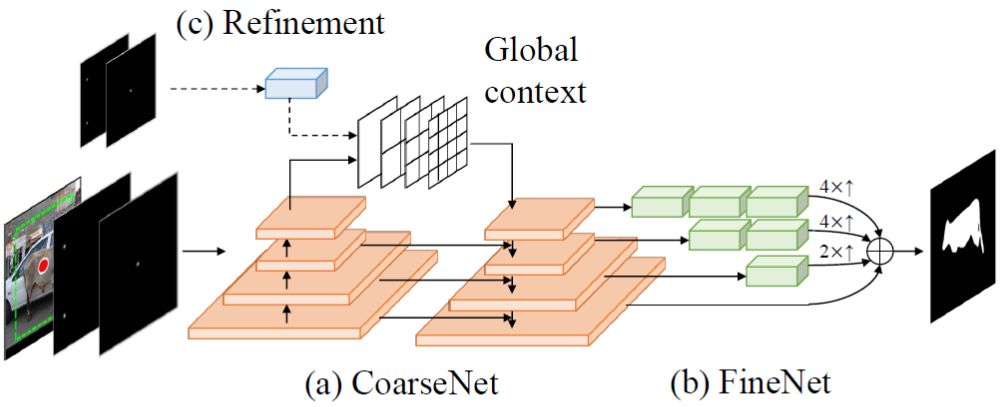


Figure 4. **Network Architecture.** (a)-(b) Our segmentation network adopts a coarse-to-fine structure similar to [14], augmented with a pyramid scene parsing (PSP) module [68] for aggregating global contextual information. (c) We also append a lightweight branch before the PSP module to accept the additional clicks input for interactive refinement.

Figure 3.2: IOG architecture (not final).

List of Figures

List of Tables

Bibliography

[18a] *Detection and Segmentation*. May 1, 2018. URL: http://cs231n.stanford.edu/slides/2018/cs231n_2018_lecture11.pdf.

[18b] *Selbstlernende Maschinen - wie Künstliche Intelligenz entsteht*. July 29, 2018. URL: https://www.hr-inforadio.de/podcast/wissen/selbstlernende-maschinen---wie-kuenstliche-intelligenz-entsteht,podcast-episode-53312.html.

[20] *Image Annotation Tools: Which One to Pick in 2020?* Feb. 11, 2020. URL: https://bohemian.ai/blog/image-annotation-tools-which-one-pick-2020/.

[21] *tf.keras.metrics.MeanIoU*. Apr. 14, 2021. URL: https://www.tensorflow.org/api_docs/python/tf/keras/metrics/MeanIoU.

[BB01] M. Banko and E. Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation." In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France: Association for Computational Linguistics, July 2001, pp. 26–33. DOI: 10.3115/1073012.1073017.

[BKC17] V. Badrinarayanan, A. Kendall, and R. Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pp. 2481–2495. DOI: 10.1109/TPAMI.2016.2644615.

[Che+16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." In: *arXiv:1606.00915* (2016).

[Che+18] Y. Chen, W. Li, X. Chen, and L. V. Gool. "Learning Semantic Segmentation from Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach." In: *CoRR* abs/1812.05040 (2018). arXiv: 1812.05040.

[CL13] G. Csurka and D. Larlus. "What is a good evaluation measure for semantic segmentation?" In: vol. 26. Jan. 2013. DOI: 10.5244/C.27.32.

[Cor+16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. B. and Uwe Franke, S. Roth, and B. Schiele. "The Cityscapes Dataset for Semantic Urban Scene Understanding." In: *CoRR* abs/1604.01685 (2016). arXiv: 1604.01685.

[Dai+17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner. "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes." In: *CoRR* abs/1702.04405 (2017). arXiv: 1702.04405.

[Den+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." In: *CVPR09*. 2009.

[DV18] V. Dumoulin and F. Visin. *A guide to convolution arithmetic for deep learning*. 2018. arXiv: 1603.07285 [stat.ML].

[Eve+10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge." In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338.

[Fer19] A. Fertig. "Semantic Segmentation: State of the Art." unpublished seminar paper. 2019.

[GAD17] V. Gudivada, A. Apon, and J. Ding. "Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations." In: *International Journal on Advances in Software* 10 (July 2017), pp. 1–20.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org [Zugriff am: 02.05.2021]. MIT Press, 2016.

[Gér17] A. Géron. *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly, 2017. ISBN: 978-1-491-96229-9.

[He+16] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Computer Vision and Pattern Recognition (CVPR)*. 2016.

[HNP09] A. Halevy, P. Norvig, and F. Pereira. "The Unreasonable Effectiveness of Data." In: *Intelligent Systems, IEEE* 24 (May 2009), pp. 8–12. DOI: 10.1109/MIS.2009.36.

[KK12] P. Krähenbühl and V. Koltun. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." In: *CoRR* abs/1210.5644 (2012). arXiv: 1210.5644.

[Li+18] W. Li, C. He, J. Fang, and H. Fu. "Semantic Segmentation Based Building Extraction Method Using Multi-source GIS Map Datasets and Satellite Imagery." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, pp. 233–2333. DOI: 10.1109/CVPRW.2018.00043.

[Lin+14] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft COCO: Common Objects in Context." In: *CoRR* abs/1405.0312 (2014).

[LSD15] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.

[LSL09] L.-J. Li, R. Socher, and F.-F. Li. "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework." In: June 2009, pp. 2036–2043. DOI: 10.1109/CVPRW.2009.5206718.

[Man+18] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. "Deep Extreme Cut: From Extreme Points to Object Segmentation." In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[MG15] M. Menze and A. Geiger. "Object scene flow for autonomous vehicles." In: June 2015, pp. 3061–3070. DOI: 10.1109/CVPR.2015.7298925.

[Neu+17] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder. "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes." In: *International Conference on Computer Vision (ICCV)*. 2017.

[NHH15] H. Noh, S. Hong, and B. Han. "Learning Deconvolution Network for Semantic Segmentation." In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1520–1528. DOI: 10.1109/ICCV.2015.178.

[RFB15] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597.

[SM18] R. Shanmugamani and S. Moore. *Deep Learning for Computer Vision: Expert Techniques to Train Advanced Neural Networks Using TensorFlow and Keras*. Packt Publishing, 2018. ISBN: 9781788295628.

[SZ15] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556.

[Tra+18] T. Tran, O.-H. Kwon, K.-R. Kwon, S.-H. Lee, and K.-W. Kang. "Blood Cell Images Segmentation using Deep Learning Semantic Segmentation." In: *2018 IEEE International Conference on Electronics and Communication Engineering (ICECE)*. 2018, pp. 13–16. DOI: 10.1109/ICECOME.2018.8644754.

[Zha+17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. "Pyramid Scene Parsing Network." In: *Computer Vision and Pattern Recognition (CVPR)*. 2017.

[Zha+20] S. Zhang, J. H. Liew, Y. Wei, S. Wei, and Y. Zhao. "Interactive Object Segmentation With Inside-Outside Guidance." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12234–12244.

[Zol+19] J. Zolfaghari Bengar, A. Gonzalez-Garcia, G. Villalonga, B. Raducanu, H. H. Aghdam, M. Mozerov, A. M. Lopez, and J. van de Weijer. "Temporal Coherence for Active Learning in Videos." In: *arXiv preprint arXiv:1908.11757* (2019).