Semantic Image Segmentation: Two Decades of Research
Other titles in Foundations and Trends® in Computer Graphics and Vision

Video Summarization Overview
Mayu Otani, Yale Song and Yang Wang
ISBN: 978-1-63828-078-1

A Comprehensive Review of Modern Object Segmentation Approaches
Yuanbo Wang, Unaiza Ahsan, Hanyan Li and Matthew Hagen

Deep Learning for Image/Video Restoration and Super-resolution
A. Murat Tekalp

Deep Learning for Multimedia Forensics
Irene Amerini, Aris Anagnostopoulos, Luca Maiano and Lorenzo Ricciardi Celsi
ISBN: 978-1-68083-854-1

Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art
Joel Janai, Fatma Güney, Aseem Behl and Andreas Geiger

Discrete Graphical Models - An Optimization Perspective
Bogdan Savchynskyy
ISBN: 978-1-68083-638-7
Semantic Image Segmentation: Two Decades of Research

Gabriela Csurka  
Naver Labs Europe  
Gabriela.Csurka@naverlabs.com

Riccardo Volpi  
Naver Labs Europe  
Riccardo.Volpi@naverlabs.com

Boris Chidlovskii  
Naver Labs Europe  
Boris.Chidlovskii@naverlabs.com
Editorial Scope

Topics

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2022, Volume 14, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.
# Contents

## Preface

### 1. Semantic Image Segmentation (SiS)

1.1 Historical SiS Methods ............................... 9
1.2 Deep Learning-based SiS .............................. 14
1.3 Beyond Classical SiS ................................. 30

### 2. Domain Adaptation for SiS (DASiS)

2.1 Brief Introduction into UDA ......................... 46
2.2 Adapting SiS between Domains ...................... 49
2.3 Complementary Techniques .......................... 58
2.4 Beyond Classical DASiS .............................. 68

### 3. Datasets and Benchmarks

3.1 SiS Datasets and Benchmarks ....................... 84
3.2 DASiS Benchmarks .................................. 94

### 4. Related Segmentation Tasks

4.1 Instance Segmentation (InstS) ...................... 100
4.2 Panoptic Segmentation (PanS) ..................... 103
4.3 Medical Image Segmentation ....................... 105
## 5 Summary and Perspectives

5.1 Monograph Summary .................................................. 107
5.2 SiS with Additional Modalities ................................. 108
5.3 Perspectives in SIS .................................................... 109

### Abbreviations

### References
ABSTRACT

Semantic image segmentation (SiS) plays a fundamental role in a broad variety of computer vision applications, providing key information for the global understanding of an image. This survey is an effort to summarize two decades of research in the field of SiS, where we propose a literature review of solutions starting from early historical methods followed by an overview of more recent deep learning methods including the latest trend of using transformers. We complement the review by discussing particular cases of the weak supervision and side machine learning techniques that can be used to improve the semantic segmentation such as curriculum, incremental or self-supervised learning.

State-of-the-art SiS models rely on a large amount of annotated samples, which are more expensive to obtain than labels for tasks such as image classification. Since unlabeled data is instead significantly cheaper to obtain, it is not surprising that Unsupervised Domain Adaptation (UDA) reached a broad success within the semantic segmentation community. Therefore, a second core contribution of this monograph is to summarize five years of a rapidly growing
field, Domain Adaptation for Semantic Image Segmentation (DASiS) which embraces the importance of semantic segmentation itself and a critical need of adapting segmentation models to new environments. In addition to providing a comprehensive survey on DASiS techniques, we unveil also newer trends such as multi-domain learning, domain generalization, domain incremental learning, test-time adaptation and source-free domain adaptation. Finally, we conclude this survey by describing datasets and benchmarks most widely used in SiS and DASiS and briefly discuss related tasks such as instance and panoptic image segmentation, as well as applications such as medical image segmentation.

We hope that this monograph will provide researchers across academia and industry with a comprehensive reference guide and will help them in fostering new research directions in the field.
Semantic image segmentation (SiS) plays a fundamental role towards a general understanding of the image content and context. In concrete terms, the goal is to label image pixels with the corresponding semantic classes and to provide boundaries of the class objects, easing the understanding of object appearances and the spatial relationships between them. Therefore, it represents an important task towards the design of artificial intelligent systems. Indeed, systems such as intelligent robots or autonomous cars should have the ability to coherently understand visual scenes, in order to perceive and reason about the environment holistically.

Hence, semantic scene understanding is a key element of advanced driving assistance systems (ADAS) and autonomous driving (AD) (Teichmann et al., 2018; Hofmarcher et al., 2019) as well as robot navigation (Zurbrügg et al., 2022). The information derived from visual signals is generally combined with other sensors such as radar and/or LiDAR to increase the robustness of the artificial agent’s perception of the world (Yurtsever et al., 2020). Semantic segmentation fuels applications in the fields of robotic control and task learning (Fang et al., 2018; Hong et al., 2018b), medical image analysis (see Section 4.3), augmented reality (DeChicchis, 2020; Turkmen, 2019), satellite imaging (Ma et al., 2019) and many others.
The growth of interest in these topics has also been caused by recent advances in deep learning, which allowed a significant performance boost in many computer vision tasks – including semantic image segmentation. Understanding a scene at the semantic level has long been a major topic in computer vision, but only recent progress in the field has allowed machine learning systems to be robust enough to be integrated into real-world applications.

The success of deep learning methods typically depends on the availability of large amounts of annotated training data, but manual annotation of images with pixel-wise semantic labels is an extremely tedious and time consuming process. As the major bottleneck in SiS is the high cost of manual annotation, many methods rely on graphics platforms and game engines to generate synthetic data and use them to train segmentation models. The main advantage of such synthetic rendering pipelines is that they can produce a virtually unlimited amount of labeled data. Due to the constantly increasing photo-realism of the rendered datasets, the models trained on them yield good performance when tested on real data. Furthermore, they allow to easily diversify data generation, simulating various environments and weather/seasonal conditions, making such data generation pipeline suitable to support the design and training of SiS models for the real world.

While modern SiS models trained on such simulated images can already perform relatively well on real images, their performance can be further improved by domain adaptation (DA) – and even with unsupervised domain adaptation (UDA) not requiring any target labels. This is due to the fact that DA allows to bridge the gap caused by the domain shift between the synthetic and real images. For the aforementioned reasons, sim-to-real adaptation represents one of the leading benchmarks to assess the effectiveness of domain adaptation for semantic image segmentation (DASiS).

The aim of our monograph is to overview the research field of SiS. On the one hand, we propose a literature review of semantic image segmentation solutions designed in the last two decades – including early historical methods and more recent deep learning ones, also covering the recent trend of using transformers with attention mechanism. On the other hand, we devote a large part of the monograph to survey methods
designed ad hoc for DASiS. While our work shares some similarities with some of the previous surveys on this topic, it covers a broader set of DASiS approaches and departs from these previous attempts pursuing different directions that are detailed below.

Amongst the existing works surveying SiS methods, we can mention Thoma (2016) who gives a brief overview of some of the early semantic segmentation and low-level segmentation methods. Li et al. (2018a) and Zhou et al. (2018) discuss some of the early deep learning-based solutions for SiS. A more complete survey on deep SiS models has been proposed by Minaee et al. (2020), while Zhang et al. (2020a) focus on reviewing semi- and weakly supervised semantic segmentation models. We cover most of these methods in Section 1, where we provide a larger spectrum of the traditional SiS methods in Section 1.1. Then, in Section 1.2, we organize the deep SiS methods according to their most important characteristics, such as the type of encoder/decoder, attention or pooling layers, solutions to reinforcing local and global consistency. In contrast to the previous surveys, this section also includes the latest SiS models that use attention mechanisms and transformers as encoder and/or decoder. One of the core contributions of this section is Table 2.1, which presents a broad set of deep models proposed in the literature, and summarized according to the above mentioned characteristics. Finally, in Section 1.3 we review not only semi- and weakly supervised SiS solutions, but also new trends whose goal is improving semantic segmentation, such as curriculum learning, incremental learning and self-supervised learning.

In Section 2, we present and categorize a large number of approaches devised to tackle the DASiS task. Note that previous DA surveys (Gopalan et al., 2015; Csurka, 2017; Kouw and Loog, 2021; Zhang and Gao, 2019; Venkateswara and Panchanathan, 2020; Singh et al., 2020; Csurka, 2020; Wang and Deng, 2018; Wilson and Cook, 2020) address generic domain adaptation approaches that mainly cover image classification and mention only a few adaptation methods for SiS. Similarly, in recent surveys on domain generalization (Wang et al., 2020b; Zhou et al., 2020a), online learning (Hoi et al., 2018) and robot perception (Garg et al., 2020), several DA solutions are mentioned, but yet DASiS received only marginal attention here. The most complete survey – and therefore most similar to the content of our Section 2 –
is by Toldo et al. (2020a), which also aimed at reviewing the recent trends and advances developed for DASiS. Nevertheless, we argue that our survey extends and enriches it in multiple ways. First, our survey is more recent in such a quickly evolving field as DASiS, so we address an important set of recent works appeared after their survey. Second, while we organize the DASiS methods according to how domain alignment is achieved similarly to (Toldo et al., 2020a) – namely on image, feature or output level – we complement it with different ways of grouping DASiS approaches, namely based on their most important characteristics, such as the backbone used for the segmentation network, the type and levels of domain alignments, any complementary techniques used and finally the particularity of each method compared to the others. We report our schema in Table 2.1, which represents one of the core contributions of this monograph. Third, we survey a large set of complementary techniques in Section 2.3 that can help boost the adaptation performance, such as self-training, co-training, self-ensembling and model distillation.

Finally, in Section 2.4 we propose a detailed categorization of some of the related DA tasks – such as multi-source, multi-target domain adaptation, domain generalization, source-free adaptation, domain incremental learning, etc. – and survey solutions proposed in the literature to address them. None of the previous surveys has such a comprehensive survey on these related DA tasks, especially what concerns semantic image segmentation.

To complement the above two sections, which represent the core contributions of our monograph, we further provide in Section 3 a list of the datasets and benchmarks typically used to evaluate SiS and DASiS methods, covering the main metrics and discuss different SiS and DASiS evaluation protocols. Furthermore, in Section 4 we propose a short overview of the literature for three tasks strongly related to SiS, namely instance segmentation in Section 4.1, panoptic segmentation in Section 4.2 and medical image segmentation in Section 4.3.

We hope that our monograph, with its comprehensive survey of the main trends in the field of semantic image segmentation, will provide researchers both across academia and in the industry with a solid basis and background to help them develop new methods and foster new research directions.
References


References

References


Borse, S., Y. Wang, Y. Zhang, and F. Porikli. (2021). “InverseForm: A Loss Function for Structured Boundary-Aware Segmentation”. In: CVPR.


Full text available at: http://dx.doi.org/10.1561/0600000095


Byeon, W., T. M. Breuel, F. Raue, and M. Liwicki. (2015). “Scene Labeling with LSTM Recurrent Neural Networks”. In: CVPR.


Chen, Y.-C., Y.-Y. Lin, M.-H. Yang, and J.-B. Huang. (2019c). “CrDoCo: Pixel-level Domain Transfer with Cross-Domain Consistency”. In: CVPR.


References


References


Full text available at: http://dx.doi.org/10.1561/0600000095
References


References


Full text available at: http://dx.doi.org/10.1561/0600000095


Jaritz, M., R. de Charette, E. Wirbel, X. Perrotton, and F. Nashashibi. (2018). “Sparse and Dense Data with CNNs: Depth Completion and Semantic Segmentation”. In: 3DV.


References


Krähenbühl, P. and V. Koltun. (2011). “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials”. In: NeurIPS.

Kulharia, V., S. Chandra, A. Agrawal, P. Torr, and A. Tyagi. (2020). “Box2Seg: Attention Weighted Loss and Discriminative Feature Learning for Weakly Supervised Segmentation”. In: ECCV.


References


References


References


Ma, X., J. Gao, and C. Xu. (2021). “Active Universal Domain Adaptation”. In: ICCV.

Mansour, Y., M. Mohri, and A. Rostamizadeh. (2009). “Domain Adaptation with Multiple Sources”. In: NeurIPS.


References


Full text available at: http://dx.doi.org/10.1561/06000000095


References


Panareda Busto, P. and J. Gall. (2017). “Open Set Domain Adaptation”. In: ICCV.


References


Pinheiro, P. H. O. and R. Collobert. (2014). “Recurrent Convolutional Neural Networks for Scene Parsing”. In: ICML.


Ranftl, R., A. Bochkovskiy, and V. Koltun. (2021). “Vision Transformers for Dense Prediction”. In: ICCV.


Roth, P. M., S. Sternig, H. Grabner, and H. Bischof. (2009). “Classifier grids for robust adaptive object detection”. In: CVPR.


Schneider, S., E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge. (2020). “Improving robustness against common corruptions by covariate shift adaptation”. In: NeurIPS.


References


Tighe, J., M. Niethammer, and S. Lazebnik. (2014). “Scene Parsing with Object Instances and Occlusion Ordering”. In: CVPR.
References


References


Xu, W., Y. Xu, T. Chang, and Z. Tu. (2021). “Co-Scale Conv-Attentional Image Transformers”. In: ICCV.


Full text available at: http://dx.doi.org/10.1561/0600000095


Full text available at: http://dx.doi.org/10.1561/0600000095


References


