International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Impact Factor (2024): 7.741

Enhancing Semantic Segmentation with CLIP: Leveraging Cross-Modal Understanding for Image Analysis

Ziyi Han

[1] Institution: North China Electric Power University, School of Control and Computer Engineering, Beijing, CN 102206
Email: *825064077[at]qq.com*

Abstract: ***Image semantic segmentation, although not a new concept, has found significant application in various domains. For instance, it is widely used in autonomous driving for scene understanding and obstacle detection, in medical imaging for organ segmentation and anomaly detection, and in satellite imagery for land cover classification and urban planning. Despite numerous research efforts to improve image semantic segmentation, challenges such as fine-grained object delineation, handling complex scenes with multiple overlapping objects, and achieving robustness to diverse environmental conditions persist. To address these challenges, we propose leveraging the CLIP (Contrastive Language-Image Pretraining) framework for image semantic segmentation. CLIP, a recent breakthrough in computer vision and natural language processing, learns visual representations by jointly training on large-scale image-text pairs. By fine-tuning CLIP on image semantic segmentation tasks, we aim to leverage its ability to understand the semantic context of images and improve the accuracy and generalization of segmentation models. Through this approach, we anticipate overcoming some of the limitations of traditional segmentation methods and achieving more robust and effective semantic segmentation results across various applications.***

Keywords: Semantic Segmentation CLIP Transformer

1. Introduction

When you submit your paper print it in two-column format, including figures and tables. In addition, designate one author as the “corresponding author”. This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only. Semantic segmentation, a crucial task in image processing, entails assigning each pixel in an image to predefined semantic categories, thereby enabling a detailed understanding and analysis of the image content. This technology finds widespread applications across various fields, including scene understanding in autonomous driving, lesion localization in medical imaging, and crop monitoring in agriculture.

Traditionally, image segmentation relied on handcrafted features and algorithms like edge detection and region growing. However, these methods[0][2]often struggled to accurately capture complex image semantics and were limited in handling diverse visual data. With the advent of deep learning and other advanced technologies, semantic segmentation has witnessed significant improvements in performance and accuracy. Deep learning approaches, particularly convolutional neural networks (CNNs), have revolutionized semantic segmentation by automatically learning hierarchical features from data. This has led to more robust and accurate segmentation results compared to traditional methods. Consequently, semantic segmentation has become increasingly vital in practical applications across a wide range of domains. Among them, [0]introducing a graph-based approach revolutionized traditional segmentation by leveraging graph theory to partition images based on structural and feature similarities. Additionally, [2] innovatively combining superpixel concepts with k-means clustering reduced computational complexity while maintaining accuracy. Furthermore, segmentation accuracy and efficiency were improved through multi-scale feature extraction using atrous convolution. Moreover, extending CNN architectures to pixel-level tasks achieved state-of-the-art performance.

However, despite these advancements, current CNNs in semantic segmentation have a drawback. They often struggle to understand the global context of the scene. While CNNs excel at extracting rich feature information from local regions, they frequently lack effective modeling of the overall context. This limitation can result in instances of missegmentation or undersegmentation, particularly in complex scenes where objects may overlap or occlude each other. With the continuous development of Transformer models in the field of computer vision, an increasing number of researchers aspire to apply the advantages of pre-trained large-scale Transformer models to semantic segmentation tasks. This trend is gradually augmenting the conventional methods of semantic segmentation, previously heavily reliant on CNNs.

Additionally, within the current pre-trained paradigm, leveraging high-quality semantic information from images is also a challenge that this paper aims to address. To address the aforementioned challenges, this paper proposes a hierarchical semantic segmentation network based on CLIP (Contrastive Language-Image Pre-training).

The proposed hierarchical semantic segmentation network leverages CLIP for enhanced semantic understanding. By integrating language and image information, CLIP provides a powerful framework for semantic segmentation. The network architecture consists of multiple layers, each responsible for different levels of semantic abstraction. At the lowest level, the network captures basic visual features, while higher levels focus on capturing more abstract semantic concepts. This hierarchical approach allows the network to effectively model

the complex relationships between different semantic elements in an image, leading to improved segmentation accuracy.

Furthermore, the network utilizes pre-trained CLIP embeddings to incorporate high-quality semantic information into the segmentation process. By leveraging the rich semantic representations learned by CLIP during pre-training, the network can better understand the underlying semantics of the input image, leading to more accurate and reliable segmentation results. Additionally, the network is trained using a contrastive loss function, which encourages the model to learn discriminative representations for different semantic classes. This helps improve the model's ability to differentiate between different objects and background regions in the image, further enhancing segmentation performance.

In conclusion, the proposed hierarchical semantic segmentation network based on CLIP offers a promising solution for addressing the challenges faced by current CNN-based segmentation methods. By leveraging the power of pre-trained language-image representations, the network achieves state-of-the-art performance in semantic segmentation tasks, providing more reliable and efficient solutions for various application scenarios.

The contributions of this paper are as follows:
Enhanced Segmentation Accuracy with CLIP Integration: The integration of CLIP into our semantic segmentation framework has significantly improved segmentation accuracy and correctness. CLIP's cross-modal understanding provides additional semantic cues and contextual information, leading to more precise delineation of objects and scenes in images.

Improved Feature Representation through Triple Feature Fusion: The introduction of triple feature extraction and fusion has led to superior feature representation. By integrating multiple scales of features extracted from the input image using ImageNet pre-trained backbones, along with textual features encoded by CLIP, our approach achieves a more comprehensive representation of the visual and semantic content. This triple feature fusion results in a richer and more discriminative feature representation, facilitating more accurate and robust semantic segmentation.

2. Related work

2.1 Semantic segmentation

Semantic segmentation was historically approached as a pixel classification task using CNNs[6][7][8][9]. Recent advancements[10][11] have demonstrated the efficacy of transformer-based techniques in semantic segmentation, inspired by their success in language and vision domains [2, 37]. MaskFormer [11], among these approaches, reframed semantic segmentation as a mask classification challenge, building upon earlier methodologies [3,14,16], by employing a transformer decoder with object queries. Similarly, we also reinterpret semantic segmentation as a mask classification problem.

2.2 CLIP Transformer

CLIP (Contrastive Language-Image Pre-training) was first introduced by Radford et al. in 2021. It is a cross-modal deep learning model developed by OpenAI. CLIP stands out for its ability to understand both images and text simultaneously and is effectively pre-trained on a large-scale dataset of image-text pairs. It learns rich semantic representations by understanding the contrastive relationships between images and text.

In the field of semantic segmentation, the application of CLIP brings several benefits. Firstly, CLIP provides additional semantic cues and contextual information for semantic segmentation models, thereby improving the accuracy and robustness of segmentation results. Secondly, by combining the semantic information from both images and text, CLIP can supplement information in images that may be difficult to obtain from images alone, such as object names, attributes, and relationships, enriching the visual understanding capability of semantic segmentation models. Additionally, CLIP's pre-trained representations demonstrate strong generalization abilities, aiding semantic segmentation models in achieving better performance across various scenarios and datasets.

In this study, we apply CLIP to semantic segmentation tasks with the aim of leveraging its cross-modal understanding capabilities to enhance semantic segmentation models. Specifically, we propose a method that integrates region information from CLIP to learn semantic understanding of images. This approach not only harnesses the pre-training of CLIP on a large-scale dataset of image-text pairs but also provides richer and more accurate semantic representations for semantic segmentation models. By combining CLIP's pre-trained representations with semantic segmentation tasks, we aim to achieve more accurate and efficient semantic segmentation, offering new avenues and methodologies for further research and applications in the field of image understanding.

3. Methods

In this section, we present our model. The overall model diagram is as follows: Initially, we introduce the feature extraction module, followed by an explanation of the semantic fusion module. Finally, the loss function is applied.

3.1 Features encoder

The input to the feature extractor is an image along with its corresponding textual annotation. We utilize popular ImageNet pre-trained backbones [17, 30, 31] to extract multi-scale feature representations from the input image. Denoted as $\boldsymbol{r}$, the feature extractor is instrumental in this process. Commonly acknowledged is the ability of encoder's shallow features to extract surface-level details like texture and edges in images. Conversely, deeper layers delve into more profound semantic representations. This paper adopts a hierarchical structure, aiming to harness diverse expressions from multiple dimensions and enrich visual understanding. By leveraging this approach, we seek to capture a comprehensive view of visual content, bridging the gap

between low-level features and high-level semantics for enhanced image comprehension and analysis.

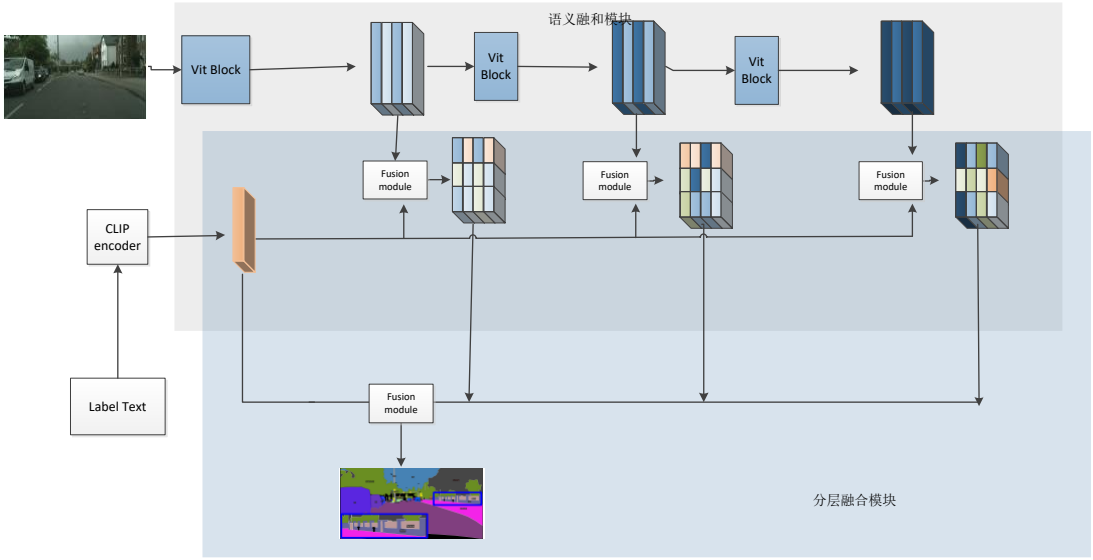

Figure 1: Model architecture diagram

Images represented in different dimensions : r_1 , r_2 , r_3 . Similarly, we utilize the pre-trained CLIP (Contrastive Language-Image Pre-training) model, trained on a dataset of 300 million image-text pairs, to encode textual information. The utilization of the pre-trained CLIP model underscores its pivotal role in contemporary artificial intelligence, as it stands at the forefront of bridging the crucial gap between images and text. At the core of CLIP's significance lies its remarkable ability to understand the nuanced associations between images and text. This is facilitated by its extensive pretraining on a vast corpus of image and text data. Unlike conventional models, which merely map images and text into a shared embedding space, CLIP employs contrastive learning to capture the underlying semantic similarities and differences between them. Trained on such an extensive dataset, CLIP transcends traditional limitations, enabling it to capture the subtleties of language and visual content with remarkable accuracy. Its comprehensive understanding of the interplay between images and text empowers CLIP to excel across a wide range of tasks, including image classification, image-text retrieval, and zero-shot learning. The text represented output from CLIP can be represented as T.

3.2 Semantic fusion

Integrating semantics with images enhances semantic segmentation capabilities. Semantic information provides additional context and understanding for image segmentation, allowing segmentation models to better comprehend image content. By incorporating semantic knowledge, we can precisely locate and segment different objects and regions in the image, reducing instances of missegmentation and omission. This holistic approach improves the quality and accuracy of segmentation results, providing more support and guidance for image understanding tasks, especially in scenarios with complex scenes and dense objects.

Due to the inconsistency in dimensionality between the output of the text encoder and that of the image, we first expand the features from the text encoder to match the dimensions of the image. The formula representation is as follows:

$$T_{r1\_expand} = unsqueeze(FC_{r1}(T, \theta_1)) \quad (1)$$

Here, FC_{r1} and θ_1 are learnable parameters. T_{r1_expand} refers to features expanded to match the dimensionality of r1 features, Similarly, through the formulas provided below, we can obtain the expressions for the expanded features of the other two dimensions.

$$T_{r2\_expand} = unsqueeze(FC_{r2}(T, \theta_2)) \quad (2)$$

$$T_{r3\_expand} = unsqueeze(FC_{r3}(T, \theta_3)) \quad (3)$$

Here, FC_{r2} , FC_{r2} , θ_2 , θ_3 carries a similar meaning, both referring to learnable parameters.

Then, we concatenate the features obtained with the same dimensionality. The formula representation is as follows:

$$T_{r1\_expand} = concat[T_{r1\_expand}, FC_1(r_1, \delta_1)] \quad (4)$$

$$r_{new\_1} = MultiHead(T_{r1\_expand}, r_1) \quad (5)$$

Here, FC_1 and δ_1 are learnable parameters, $concat[\square]$ refers to concatenation operation, $MultiHead$ represents a multi-head self-attention function, aimed at injecting semantics into the feature representation of the image, to obtain a semantically enriched feature representation. r_{new_1} refers to the result after the fusion operation. Similarly, we can obtain the results for the other two dimensions, as follows:

$$T_{r2\_expand} = concat[T_{r2\_expand}, FC(r_2, \delta_2)] \quad (6)$$

$$r_{new\_2} = MultiHead(T_{r2\_expand}, r_2) \quad (7)$$

$$T_{r3\_expand} = concat[T_{r3\_expand}, FC(r_3, \delta_3)] \quad (8)$$

$$r_{new\_3} = MultiHead(T_{r3\_expand}, r_3) \quad (9)$$

Thus, we have obtained the image representation with injected semantics across three dimensions.

Finally, we merge the obtained image features from the three dimensions with the original text. This fusion process aims to establish semantic correlations between images and text, facilitating deeper semantic understanding and expression. By combining image features with textual information, we leverage the complementarity between images and text, enabling a more comprehensive depiction and explanation of the involved scenes or objects. Such integrated representation not only enhances the understanding of image content but also provides richer information for advanced visual reasoning and applications. The formula representation is as follows:

$$r_{3-1} = FC_{3-1}(concat[r_{new\_3}, r_{new\_1}], \omega_{3-1}) \quad (10)$$

$$r_{3-2} = FC_{3-2}(concat[r_{new\_3}, r_{new\_2}], \omega_{3-2}) \quad (11)$$

$$r_{1-2} = FC_{1-2}(concat[r_{new\_1}, r_{new\_2}], \omega_{1-2}) \quad (12)$$

$$r_{fusion} = FC_{fusion}(concat[r_{new\_1}, r_{new\_2}, r_{new\_3}], \theta_{fusion}) \quad (13)$$

Here, $FC_{3-1}, FC_{3-2}, FC_{1-2}, FC_{fusion}$ are learnable parameters. Similarly, $\omega_{3-1}, \omega_{3-2}, \omega_{1-2}, \theta_{fusion}$ also are. Through the same three concatenation methods, different representations of the image were obtained, and then these representations were aggregated to obtain a comprehensive representation of the image in three dimensions. The significance of this process lies in the ability to capture various aspects of the image's feature information by using different concatenation methods, thus obtaining a more comprehensive and diversified representation of the image. Aggregating these representations across different dimensions allows for a comprehensive consideration of the relationships between various features, providing a more accurate and rich representation of the image. This integrated representation better reflects the semantics and essence of the image.

3.3 Loss

Following [10], we use a binary cross-entropy as our loss function to learn the model parameters. This loss function is commonly employed in binary classification tasks and measures the discrepancy between the predicted and actual binary labels. It is particularly suitable for scenarios where each data instance belongs to one of two classes. By minimizing the binary cross-entropy loss, our model learns to accurately classify instances into their respective classes, thereby improving its overall performance. The formula representation is as follows:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[yilog(p_i) + (1-y_i)log(1-p_i)] \quad (14)$$

4. Experience

In this section, our goal is to assess the effectiveness of the "Semantic Segmentation of Images Using CLIP" method proposed in this paper from two primary perspectives: the semantic fusion attained by integrating CLIP and the performance comparison between using and not using CLIP, as well as the influence of employing a hierarchical strategy on the results. All experiments were conducted using the computing resources of an NVIDIA GeForce GTX 8090 GPU.

4.1 Datasets

The CamVid dataset consists of 367 training images, 101 validation images, and 233 testing images, all with a resolution of 480×360 pixels.

The Cityscapes [12] consists of a total 19 (11“stuff” and 8 “thing”) classes with 2,975 training, 500 validation and 1,525 test images.

4.2 Evaluation Metrics

In this paper, we employ the mean Intersection over Union (mIoU) as the metric for evaluating the network accuracy. mIoU is a commonly used image segmentation metric that considers the accuracy of the model's segmentation results for each class and then computes the average across all classes to provide a comprehensive accuracy assessment.

4.3 Experimental Results

Comparing our model with other segmentation networks on the Cityscapes validation set, this paper showcases the superior segmentation accuracy. The comparative results, illustrated in Table 1, reveal that our model achieves the highest average segmentation accuracy on the CamVid dataset.

The comparative results, illustrated in Table 2, reveal that our model achieves the highest average segmentation accuracy on the Cityscapes dataset. Notably, it achieves the highest segmentation accuracy for classes such as building, road, and sidewalk.

Table 1: Comparison of IoU and mIoU (%) for each network in CamVid dataset.

网络名称	*sky*	*fence*	*Pole*	*tree*	*Side-walk*
ENet	91.0	21.7	25.6	67.9	75.0
ERFNet	91.7	36.4	35.9	72.9	79.8
ESNet	91.3	39.2	36,6	71.8	81.8
DSANet	91.5	45.8	34.5	76.4	80.9
Our	91.6	45.6	35.7	75.4	81.9

Table 2: Comparison of IoU and mIoU (%) for each network in Cityscapes dataset.

网络名称	*road*	*wall*	*fence*	*Traffic light*	*Traffic sign*
ERNet	97.9	45.3	50.4	62.9	68.4
CGNet	95.5	40.0	43.0	59.8	63.9
ESNet	98.1	48.3	36,6	62.5	72.3
DSANet	96.8	45.8	50.8	64.0	71.7
Our	97.6	48.5	50.7	65.4	70.9

Illustrative semantic segmentation results of our modal and other networks on the Cityscapes dataset are presented in Figure 1. In Figure 1, it can be seen that our modal exhibits lower misclassification rates and higher segmentation completeness compared to other networks.

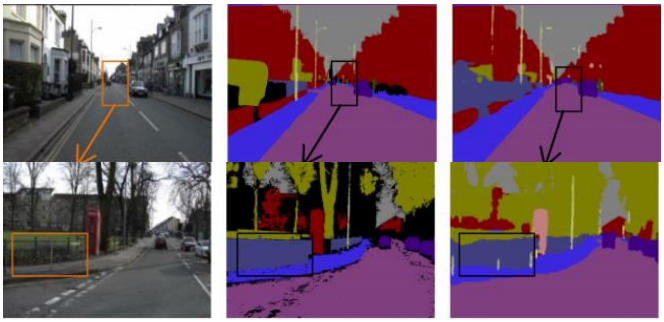

Figure 1: Comparison of semantic segmentation results with other networks on the CamVid dataset

5. Conclusion

Semantic segmentation of images has become indispensable across various domains, including autonomous driving, medical imaging, and satellite imagery analysis. Despite its widespread use, challenges persist in accurately delineating fine-grained objects, handling complex scenes with overlapping objects, and ensuring robustness in diverse environmental conditions.

To tackle these challenges, we propose leveraging the CLIP (Contrastive Language-Image Pretraining) framework for image semantic segmentation. CLIP, a recent breakthrough in computer vision and natural language processing, learns visual representations by training on large-scale image-text pairs. By fine-tuning CLIP for semantic segmentation tasks, we aim to harness its capability to understand the semantic context of images, thereby enhancing the accuracy and generalization of segmentation models.

This innovative approach holds promise for overcoming the limitations of traditional segmentation methods. By integrating CLIP's multimodal understanding of images and text, we anticipate achieving more robust and effective semantic segmentation results. Ultimately, this advancement has the potential to revolutionize various applications, from enhancing safety in autonomous vehicles to improving medical diagnosis and urban planning

Through the proposed integration of CLIP into semantic segmentation tasks, we envision a future where segmentation models can better understand the intricate relationships between visual and textual information, leading to more accurate and reliable segmentation results across diverse real-world scenarios.

References

[1] J. P. Serra, "Image Analysis and Mathematical Morphology," Academic Press, Inc., Orlando, FL, USA, 1982.
[2] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.
[3] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).
[4] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).
[5] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).
[6] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation.
[7] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. In BMVC, 2019.
[8] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In ECCV, pages 108–124, 2016.
[9] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In CVPR, pages 4424– 4433, 2020.
[10] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In CVPR, pages 18145– 18154, 2022. 1, 2, 6.
[11] Muchen Li and Leonid Sigal. Referring transformer: A one step approach to multi-task visual grounding. In NeurIPS, 2021. 2, 6, 7.
[12] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In ICCV, pages 4683– 4693, 2019.
[13] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. MAttNet: Modular attention network for referring expression comprehension. In CVPR, pages 1307–1315, 2018.
[14] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In ICML, pages 6105–6114, 2019.
[15] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In ACMMM, pages 1274–1282, 2020.
[16] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In CVPR, pages 8530–8539, 2020.
[17] Justin Lazarow, Weijian Xu, and Zhuowen Tu. Instance segmentation with mask-supervised polygonal boundary transformers. In CVPR, pages 4372–4381, 2022.
[18] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916, 2022
[19]] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In ECCV, pages 59–75, 2020.
[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence ´ Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, pages 740–755, 2014.