Contribution to Panoptic Segmentation

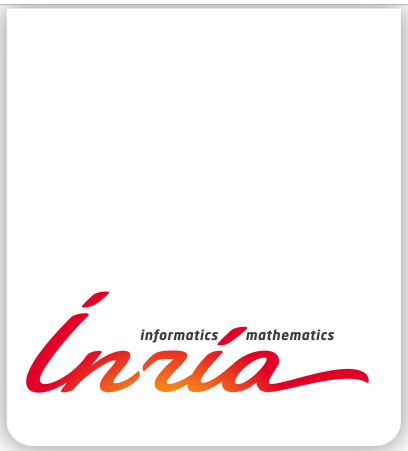
Manuel Alejandro Diaz-Zapata

Contribution to Panoptic Segmentation

Manuel Alejandro Díaz Zapata

TECHNICAL REPORT
N° 0506
September 2019
Project-Team Chroma

ISSN 0249-0803 ISRN INRIA/RT--0506--FR+ENG



Contribution to Panoptic Segmentation

Manuel Alejandro Díaz Zapata*

Project-Team Chroma

Technical Report n° 0506 — September 2019 — 9 pages

Abstract: Full visual scene understanding has always been one of the main goals of machine perception. The ability to describe the components of a scene using only information taken by a digital camera has been the main focus of computer vision tasks such as semantic segmentation and instance segmentation, where by using Deep Learning techniques, a neural network is capable to assign a label to each pixel of an image (semantic segmentation) or define the boundaries of an instance or object with more precision than a bounding box (instance segmentation).
The task of Panoptic Segmentation tries to achieve a full scene description by merging semantic and instance segmentation information and leveraging the strengths of these two tasks. On this report it is shown a possible alternative to solve this merging problem by using Convolutional Neural Networks (CNNs) to refine the boundaries between each class.

Key-words: computer vision, panoptic segmentation, semantic segmentation, instance segmentation, deep learning.

* This work has been accomplished during the internship of Manuel Alejandro Diaz Zapata at Inria-Rhone Alpes under supervision of Ozgur Erkent, Victor Romero-Cano and Christian Laugier at Chroma Project Team. Manuel Alejandro Diaz Zapata was a student of Mechatronic Engineering at Universidad Autónoma de Occidente, Colombia during his internship.

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Contribution au Segmentation Panoptique

Résumé : La compréhension visuelle complète de la scène a toujours été l'un des objectifs principaux de la perception de la machine. La capacité à décrire les composants d'une scène en utilisant uniquement les informations prises par un appareil photo numérique a été le principal objectif des tâches de vision par ordinateur telles que la segmentation sémantique et la segmentation d'instances. En utilisant des techniques d'apprentissage en profondeur, un réseau de neurones est capable d'attribuer une étiquette à chaque pixel d'une image (segmentation sémantique) ou de définir les limites d'une instance ou d'un objet avec plus de précision que le cadre de sélection (segmentation d'instance).

La tâche de segmentation panoptique proposée par Kirillov et. al tente d'obtenir une description complète de la scène en fusionnant les informations de segmentation sémantique et par instance et en exploitant les points forts de ces deux tâches. Ce rapport indique une alternative possible pour résoudre ce problème de fusion en utilisant des réseaux de neurones à convolution (CNN) pour affiner les limites entre chaque classe.

Mots-clés : vision par ordinateur, segmentation panoptique, segmentation sémantique, segmentation d'instance, apprentissage en profondeur.

1 Introduction

One of the two main research themes for the Cooperative and Human-aware Robot Navigation in Dynamic Environments (CHROMA) Project Team, is the perception and situation awareness in human-populated environments. Between CHROMA's main application domains is the task of autonomous vehicle driving, where vision plays an important role by gathering information of its environment using a camera as sensor.

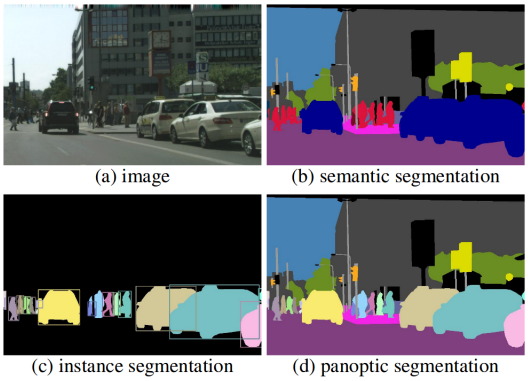


Figure 1: Panoptic segmentation, image taken from [9].

Currently, two of the mainstream vision tasks evaluated in datasets such as the Cityscapes Dataset[3], ADE20k [18], KITTI Dataset [1] and Mapillary Vistas [14], are the tasks of semantic segmentation and instance segmentation. Semantic segmentation's goal is to label an image at the pixel level, where amorphous regions of similar texture or material such as grass, sky or road are given a label depending on the class. Instance segmentation focuses on countable objects such as people, cars or animals by delimiting them in the image using bounding boxes or a segmentation mask.

Kirillov et al [9] said, that there has been a gap on the methods used to detect *stuff* or uncountable objects, and *things* or countable objects, where semantic segmentation has been mainly focused towards *stuff* and instance segmentation towards *things*. This is why on 2018 they proposed the task of Panoptic Segmentation [9], where the information can be merged into a joint task to get a better understanding of the images at the pixel level.

Since Panoptic segmentation combines segmentation and recognition tasks, a new metric called Panoptic Quality (PQ) is needed to measure the performance of the algorithms. Panoptic Quality measurement can be explained by two terms, one is the Segmentation Quality (SQ) and the other one is the Recognition Quality (RQ).

$$PQ = SQ * RQ$$

$$SQ = \frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP|}$$

$$RQ = \frac{|TP|}{|TP| + 0.5|FP| + 0.5|FN|}$$

Where p is the mask prediction given by either the semantic or instance segmentation and g is the ground truth. IoU is the intersection over union, also known as the Jaccard Index, or the overlap between both masks. And TP, FP and FN are respectively the amount of True Positives, False Positives and False Negatives.

The task of Panoptic Segmentation has already been proposed as a challenge for the Cityscapes and COCO datasets.

2 Related work

Since the publication of [9], these following papers focused on working in the panoptic segmentation task.

In the Attention-guided Unified Network for Panoptic Segmentation [10] (AUNet), proposed by Li et al., two modules are proposed, one for the background (stuff) and another for the foreground (things). Both modules share a Feature Pyramid Network as the backbone, which is then divided into a Background branch, Region Proposal Network branch and a Foreground branch. Through these modules they were able to use the instance segmentation information to improve the predictions done by the semantic segmentation module.

Kirillov et al. propose a Panoptic Feature Pyramid Network [8], which takes the backbone of Mask-RCNN [6], a Feature Pyramid Network (FPN) [11], and propose an architecture to use the FPN features to do semantic segmentation. Heuristics are used to merge the outputs given by the semantic and instance segmentation. They also find that FPNs are more efficient for increasing feature resolution for semantic semantic segmentation, compared to dilated networks and symmetric decoders.

UPSNet [17], proposed by Xiong et al., uses the same FPN backbone as Mask-RCNN as the feature extractor for what they call the Semantic Head and the Instance head. They use a deformable convolution based sub-network to do the semantic segmentation, and Mask-RCNN for the Instance head. To merge the output given by both heads they use a series of rules to define the label that will be given to each pixel. A mechanism is also proposed to give the network an added *unknown* class, so it can be used when the confidence score of a pixel is not enough after the merging heuristics.

De Geus et al. on Single Network Panoptic Segmentation for Street Understanding [4], use a ResNet-50 [7] with an output stride of 8 as the backbone feature extractor. These features go to a Instance segmentation branch and to a Semantic segmentation branch, where bounding box information is shared from the semantic branch to improve the masks on the instance branch. Once the predictions from each branch are ready, they are merged by a set of advanced heuristics that combine overlap removal, pixel scoring and removal of stuff classes with a given pixel count.

Liu et al. proposed an End-to-End Network for Panoptic Segmentation [13], which uses a FPN as a backbone, Mask-RCNN for the Instance Segmentation Branch and two 3x3 convolution layers stacked on top of the RPN feature maps from the Mask-RCNN for the semantic segmentation. To merge the outputs, a Spatial-Ranking Module (SRM) is used. This SRM takes the outputs from the instance segmentation branch, groups each instance in a channel per class and uses a Large Kernel convolution [15] to create a Spatial ranking score map, which is used to decide which pixels appear in the foreground.

3 Proposed model

3.1 Convolutional Panoptic Head

The final proposed model consists of three modules: the semantic segmentation module, the instance segmentation module and the panoptic head. Here the semantic segmentation is done by the MobileNetV2 [16], instance segmentation is done by Mask R-CNN [6] and the output of both networks are joint by the Panoptic Head.

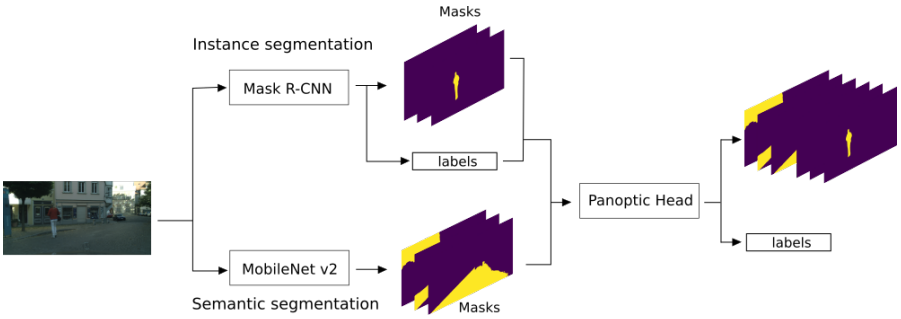


Figure 2: Proposed model

The goal of the Panoptic Head is to refine the Mask R-CNN output by using a set of convolutional neural networks (CNNs) for each class of the instance classes. This module is composed by 8 sets of CNNs, where each follows this process: first, each individual instance is stacked with the masks provided by the semantic segmentation network, resulting in a 20-layer image; then it is fed through a channel-wise convolution, followed by another convolution that reduces its dimensionality from 20 to 1; Finally, the sigmoid function is applied to the resulting one channel image to get all the values scaled between one and zero.

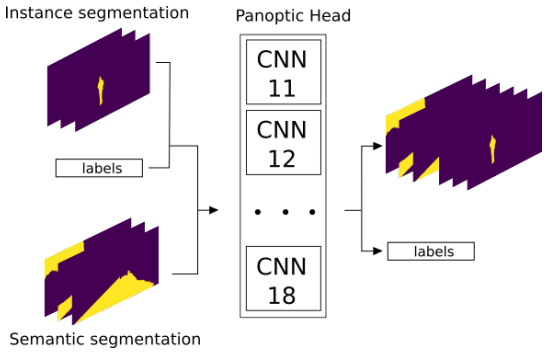


Figure 3: Expanded view of the Panoptic Head

We use one CNN for each class and stack each instance with the semantic segmentation predictions to have a fixed-channel input to the Panoptic Head module, while maintaining flexibility to the possibility of having different number of instances detected on different images. Another advantage is the possibility for each network to take contextual information from the stuff classes to improve the prediction. For example, in a road it is more probable that a detection will be a motorcycle, or that in a sidewalk it will be a bike.

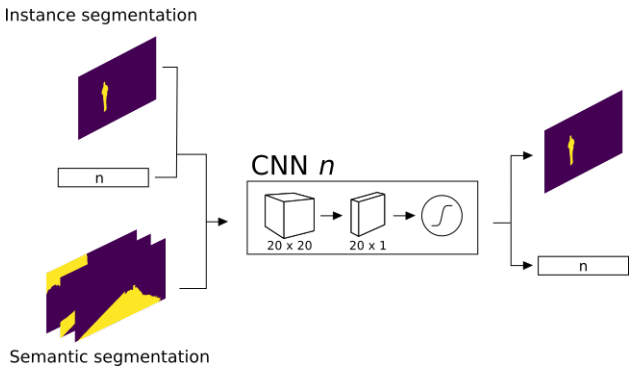


Figure 4: Isolated view of CNN for class n

One of the previous architectures proposed was just one set of CNNs for the merging. This had the drawback of forcing the output to be the fixed for different images, which meant that the input image would have a fixed amount of channels per class, causing a problem where we had to limit the amount of detections allowed per image.

4 Results

4.1 Heuristics vs Proposed Model

As a baseline, we look at the performance of the network by stacking the direct output form Mask R-CNN on top of the predictions of MobileNetV2 for the stuff classes.

On table 1 is the comparison between the mean Intersection over Union (mIoU) of the predictions from the merging using Heuristics against the results from the Convolutional Panoptic Head. This test was done on the first 2000 images of the Cityscapes training dataset. The Train class was omitted due to its low amount of instances and the Rider class was also omitted since this label is not included on the dataset the instance segmentation network used was trained on.

Class	Heuristics (mIoU)	Panoptic head (mIoU)	Detected instances
Person	71.1	**75.7**	6730
Car	84.6	**85.4**	13153
Truck	90.0	14.4	140
Bus	79.3	**87.3**	154
Motorcycle	62.2	25.8	204
Bicycle	66.4	2.19	788

Table 1: Mean Intersection over Union for Heuristics and the proposed model

Some of the classes have some improvement on their mIoU after going through the Convolutional Panoptic Head. These classes happen to be the ones with the higher number of instances, which suggests that letting the heads train for more epochs (i.e. seeing more training instances) could improve the performance for the other classes.

It is important to emphasise that since the proposed architecture uses pretrained models, the results can be limited by the by the performance of the instance and semantic segmentation networks. This problem could be solved by doing an end-to-end training.

4.2 Panoptic Quality comparison

On table 1 is the comparison between our proposed network and some of the current Panoptic Segmentation networks on the Cityscapes Validation set. OANet [13] is ommited since its results were reported on the COCO Dataset [12]. Table 2 also shows the difference in the inference time between some of the networks.

Network	PQ	SQ	RQ
Human performance [9]	69.7	84.2	82.1
AUNet [10]	59.0	n.a	n.a
Panoptic FPN [8]	58.1	n.a	n.a
UPSNet-101-M-COCO [17]	61.8	81.3	74.8
De Geus et al. [4]	45.9	74.8	58.4
Ours	37.7	54.0	69.8

Table 2: Panoptic Quality metrics on the Cityscapes Validation set

Network	Mean Prediction time
UPSNet [17]	236 ms
De Geus et al. [4]	590 ms
Ours (512×1024)	411 ms

Table 3: Mean Prediction time on the Cityscapes Validation set

5 Future Work

Tests to measure the Panoptic Quality on the KITTI dataset could be done. For this, the available instance segmentation data can be used since it includes masks for both stuff and things. Currently, there are not any scores available for the task of Panoptic Segmentation.

Since our proposed architecture uses two networks with different backbones, we propose to use a common backbone in order to cut inference time as some of the presented panoptic networks do. Sharing the backbone can also facilitate the end to end training of the network.

Using faster methods of instance segmentation like YOLACT [2], would be another option to speed up the prediction times. But in the case of YOLACT, its mask mAP is much lower than Mask R-CNN's.

Following the work done by Erkent et al. [5] on semantic grid estimation, panoptic segmentation can be projected onto an occupancy grid and be used for perception and tracking tasks.

6 Acknowledgments

This work was supported by the European Union project Cyber-Physical Systems for the EU (CPS4EU).

Experiments presented in this report were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

7 Repositories

The code developed in the internship is available on this Inria repository.

Guides to use the Grid5000's services can be found at the Chroma team repository.

References

[1] Hassan Alhaija, Siva Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision (IJCV)*, 2018.

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: real-time instance segmentation. *CoRR*, abs/1904.02689, 2019.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Daan de Geus, Panagiotis Meletis, and Gijs Dubbelman. Single network panoptic segmentation for street scene understanding. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[5] Özgür Erkent, Christian Wolf, and Christian Laugier. Semantic Grid Estimation with Occupancy Grids and Semantic Segmentation Networks. In *ICARCV 2018 - 15th International Conference on Control, Automation, Robotics and Vision*, pages 1–6, Singapore, Singapore, November 2018.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[8] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *CoRR*, abs/1901.02446, 2019.

[9] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *CoRR*, abs/1801.00868, 2018.

[10] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. *CoRR*, abs/1812.03904, 2018.

[11] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[12] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[13] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. *CoRR*, abs/1903.05027, 2019.

[14] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017.

[15] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters - improve semantic segmentation by global convolutional network. *CoRR*, abs/1703.02719, 2017.

[16] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018.

[17] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. *CoRR*, abs/1901.03784, 2019.

[18] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Contents

1 Introduction **3**

2 Related work **4**

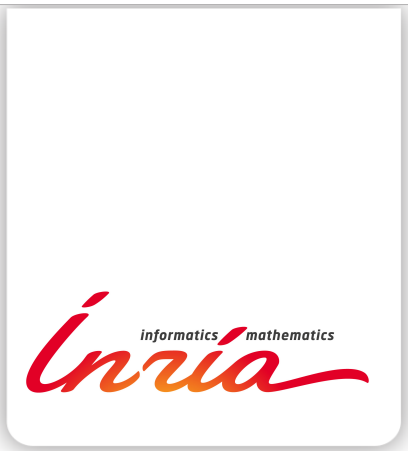
3 Proposed model **5**
3.1 Convolutional Panoptic Head 5

4 Results **6**
4.1 Heuristics vs Proposed Model 6
4.2 Panoptic Quality comparison 7

5 Future Work **7**

6 Acknowledgments **7**

7 Repositories **8**



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803