EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF SCIENCE
COMPUTER SCIENCE DEPARTMENT
COGNITIVE SYSTEMS · PROF. DR. A. ZELL

FourierNet: Compact Mask Representation for Instance Segmentation Using Differentiable Shape Decoders

Hamd ul Moqeet Riaz*, Nuri Benbarka* and Andreas Zell

Instance segmentation is one of the techniques used for scene understanding. It categorizes each pixel of an image by a specific class and at the same time distinguishes different instance occurrences, and It is usually solved with deep learning. Each instance needs a mask and it can be represented in one of three ways: *binary grid* [1], *polygon contour* [2], or *shape encoding* [3-4].

In our work, we want to find a representation that uses the least number of parameters to represent the mask without sacrificing performance.

- *Binary grid* is the most widely used representation because of its simplicity, but it needs a large amount of memory.
- In *Polygon contour* representation, the network has to predict points which are used to construct a set of closed polygons which make the mask. This has a lower memory footprint, but it adds to the complexity of the system.
- In *Shape encoding*, the network has to predict a compressed representation of the other two representations then perform a numerical transformation to get the final mask.

In our work, we introduce **FourierNet** which predicts the Fourier coefficients of a polygon contour in polar coordinates. This reduces the number of parameters to represent the mask without sacrificing quality. In addition, the Fourier transform operation is differentiable, thus the network can be trained end-to-end.

Results

- The experiments were conducted on COCO dataset.
- The FourierNet in this experiment has a Resnet-50 backbone and 60 contour points.

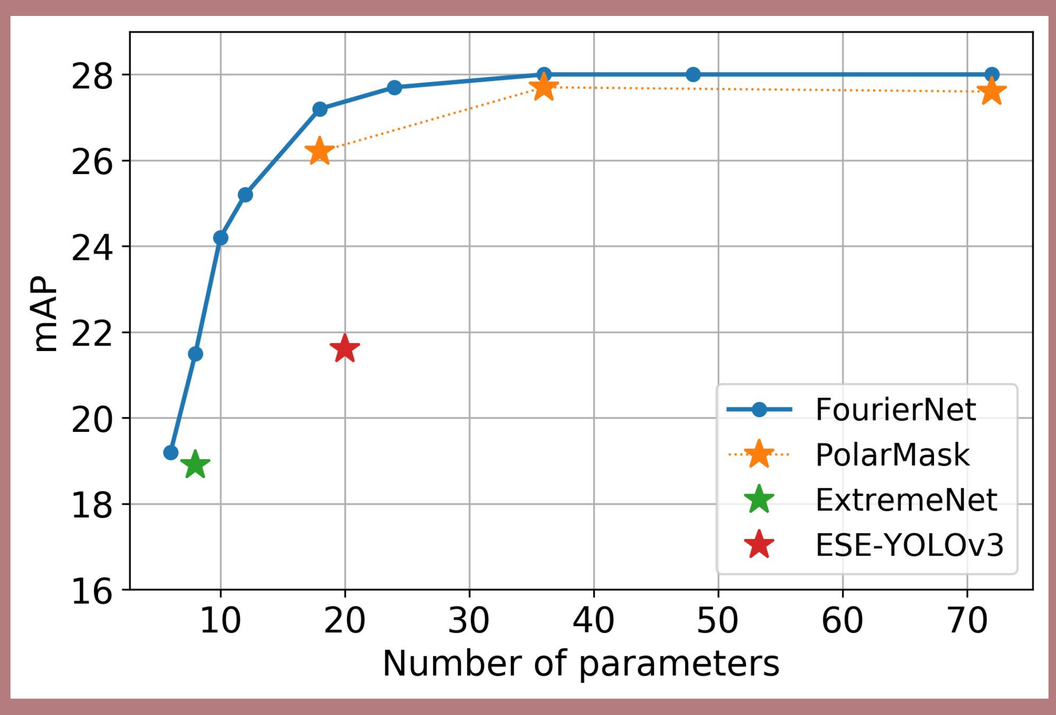
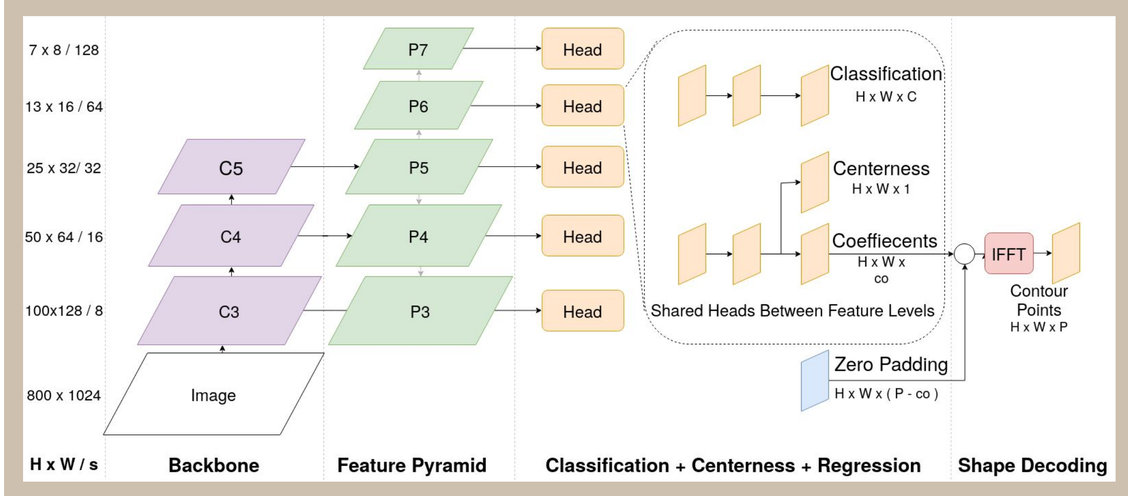


Fig. 1: FourierNet output with different number of coefficients; 2 (top left), 5 (top right), 10 (bottom left) and 20 (bottom right). The top right object has 18 rays extended from the feature point responsible for the detection (the actual mask generated in this image has 90 contour points).

Network architecture



The **FourierNet** architecture. FourierNet has 5 heads at various spatial resolutions. More uniquely, each head predicts coefficients of a Fourier series which is converted into contour points using an *Inverse Fast Fourier Transform*.

References

[1]	He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2961-2969).
[2]	Xie E, Sun P, Song X, Wang W, Liu X, Liang D, Shen C, Luo P. Polarmask: Single shot instance segmentation with polar representation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 12193-12202).
[3]	Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019 (pp. 850-859).
[4]	Xu W, Wang H, Qi F, Lu C. Explicit shape encoding for real-time instance segmentation. InProceedings of the IEEE International Conference on Computer Vision 2019 (pp. 5168-5177).

DAAD
Deutscher Akademischer Austauschdienst
German Academic Exchange Service

University of Tübingen · Sand 1 · 72076 Tübingen
Phone: +49-7071-29-70441
hamd.riaz@uni-tuebingen.de
www.cogsys.cs.uni-tuebingen.de