Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth- Davy Neven Bert De Brabandere Marc Proesmans Luc Van Gool

**(a) How many citations does the paper have?**

Citation=27

**(b) How does the proposed method rank on the CityScapes dataset?**

Panoptic Semantic Labeling Task: Not applicable

Instance-Level Semantic Labeling Task: Rank =26

**When and where exactly was the method presented for the first time?**

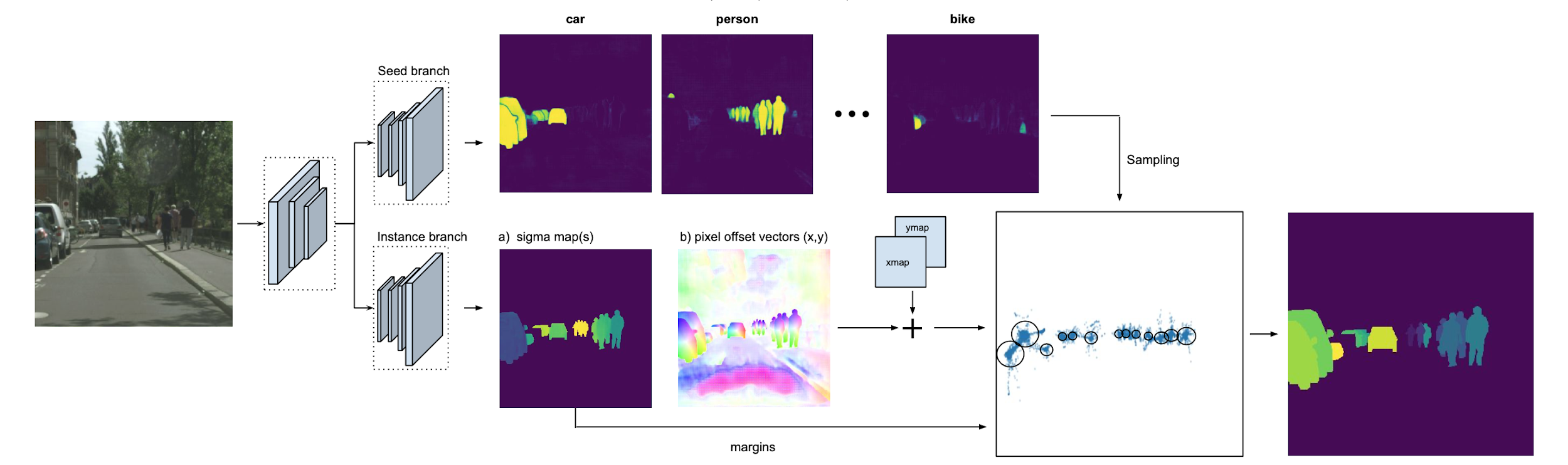
The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

**(c) Does it present a top-down approach (also known as proposal-based) or a bottom-up one (also known as proposal-free)?**

The paper presents a bottom up approach a.k.a. proposal free approach.

The authors explain that even though proposal-based methods have high accuracy, they are slow and generate masks at a fixed and low resolution. Proposal-free methods, on the other hand, can generate masks at high resolution and are often faster, but fail to reach the same accuracy as the proposal-based methods. In this paper the authors offer a combination of the benefits of both with the proposal free approach: accurate, high resolution masks combined with real-time performance.

**(d) Briefly describe and sketch the full architecture of the proposed model.**



The above sketch represents the architecture of the proposed model. The model is divided into 2 branch networks: The first branch i.e. the seed branch, predicts N seed maps(one for each semantic class), the second branch i.e. the instance branch, predicts the sigma and offset values.

Seed branch: The top branch, predicts a seed map for each semantic class. At inference time we need to cluster around the center of each object. The seed map is trained with a regression loss function. Background pixels are regressed to zero and foreground pixels are regressed to the output of the gaussian since it converts the distance between an embedding and the instance center into a closeness score.The pixels in the seed map with the highest value indicate which embeddings lay closest to an object’s center. This would also mean that the borders will have a low value, since they have more difficulty knowing to which center to point.

Instance branch: The bottom branch of the network predicts:   
a) a sigma value for each pixel, which directly translates into a clustering margin for each object. Bigger objects are more blueish, meaning a bigger margin, and smaller objects are more yellowish, meaning a smaller margin.   
The sigma is not fixed but a learnable parameter, the network modifies the sigma to minimise the loss more efficiently. Intuitively this would mean that for a big object it would adapt sigma to make the region around the centroid bigger, so that more instance pixels can point inside this region, and for small objects to choose a smaller region, so that it is easier for background pixels to point outside the region.   
b) Offset vectors for each pixel, pointing at the center of attraction, and displayed using a color-encoding where the color indicates the angle of the vector.

The pixel embeddings (= offset vectors + coordinate vectors) and margins calculated from the predicted sigma are also displayed. The cluster centers are derived from the seed maps.

ERFNet architecture is used as the backbone.It is a dense-prediction encoder-decoder network optimized for real-time semantic segmentation. The model is converted into a 2-branch network, by sharing the encoder part and having 2 separate decoders. The first branch predicts the sigma and offset values, with 3 or 4 output channels depending on sigma (σ vs σxy). The other branch outputs N seed maps, one for each semantic class. The offset values are limited between [-1,1] with a tanh activation function, sigma is made strictly positive by using an exponential activation function, effectively letting the network predict.

**(f) If the paper is about a bottom-up approach, briefly describe how the final instance segmentation is obtained from the prediction of the model: Is the approach using some kind of post-processing or clustering algorithm? If so, which one?**

A post processing approach is used to obtain the instance segmentation. At inference time, a sequential clustering approach is followed for each class-specific seed map. The pixels in the seed map with the highest value indicate which embeddings lay closest to an object’s center. The procedure is to sample the embedding with the highest seed value and use that location as an instance center. At the same location, the sigma value is also taken into consideration. By using this center and accompanying sigma, the pixel embeddings are clustered into the instance. Next, all clustered pixels in the seed map are masked out and the sampling is continued until all seeds are masked. This process is repeated for all classes.