Perception and Learning for Robotics
Exercise 1-Real World Semantic Segmentation

Yujie He
yujihe@ethz.ch

I. Performance in city scenarios

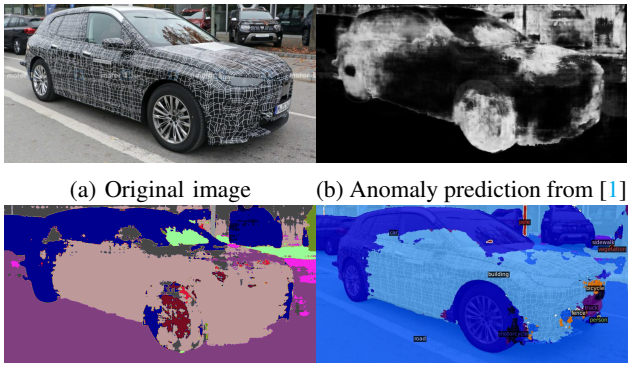
To identify the potentials and limitations of visual semantic segmentation methods, the pixel-wise anomaly detection [1] is employed for qualitative analysis. Inspired by anomaly segmentation methods via image re-synthesis, the proposed method ensembles different uncertainty maps such as softmax entropy, softmax distance, and perceptual difference to perform better.

As shown in Fig. 1, two example images are taken on the Avenue des Champs-Élysées in the direction of the Arc de Triomphe, where the baseline method shows good performance on the top image while showing inferior performance on the bottom one when a challenging anomaly occurs. In terms of the top image, the baseline method generally shows good prediction performance except for the billboard on the right and the motorcycle at the bottom. Although the generated softmax entropy implies high uncertainty and achieves explicit ensemble anomaly prediction in these areas, the predicted semantics are still misclassified as traffic signs and a combination of different classes.

The bottom image shows several objects in the complex street scene pose significant challenges to the baseline method. The first anomaly is the bird instance. In the cityscape dataset, the animal is not included in the training labels, resulting in the misclassification of the bird instance into the combination of multiple classes. It is noted that the baseline predicts a high uncertainty score in the area successfully, which verifies its effectiveness. Another anomaly is the sky and tree area under the Arc. The resynthesis images show degraded performance in this area, which leads to a high uncertainty score but still generates wrong mix of construction and sky classes.

II. Performance in camouflage scenes

Objects with camouflage usually have high similarities between the candidate objects and noisy background, which poses challenges to existing semantic segmentation methods [3]. In this section, an image of a pre-production car with camouflage[1] is employed to evaluate the performance of the given pixel-wise anomaly detection method.



(a) Original image (b) Anomaly prediction from [1]

(c) Semantic map from [1] (d) Semantic map from [2]

Figure 2: Original camouflage image and evaluation results

Fig. 2 shows the original camouflage image and evaluation results of the baseline method. On the one hand, the method gives valid semantic predictions on the road and background cars. However, it shows overconfidence on the side and front face of the car, resulting in incorrect predictions as fence and construction class according to the anomaly prediction result. High uncertainty scores can be visualized in tire rim and background objects, but the method still fails to obtain valid semantic outputs. For example, the tire rim is oversegmented with multiple classes, including bicycle, fence, and car classes.

III. MaskFormer*

In this section, a state-of-the-art (SOTA) segmentation method MaskFormer [2] is introduced for further analysis. Compared to formulating semantic and instance-level segmentation as a per-pixel and mask classification task separately, this method proposed a novel framework based on mask classification. It solves the aforementioned tasks in a unified manner by predicting a set of binary masks, each associated with a single global class label prediction. Further, it also utilizes the novel transformer architecture with attention mechanisms to improve performance compared to convolutional networks. It is noted that the model for the Cityscape dataset is used to this comparision[2].

The final column of Fig. 1 shows the predicted semantic map of MaskFormer. In comparison to the baseline method,

[1]Source: BMW's Electric Crossover iNext Displayed in Camouflage

[2]MaskFormer model: https://github.com/facebookresearch/MaskFormer/blob/main/MODEL_ZOO.md#cityscapes-semantic-segmentation

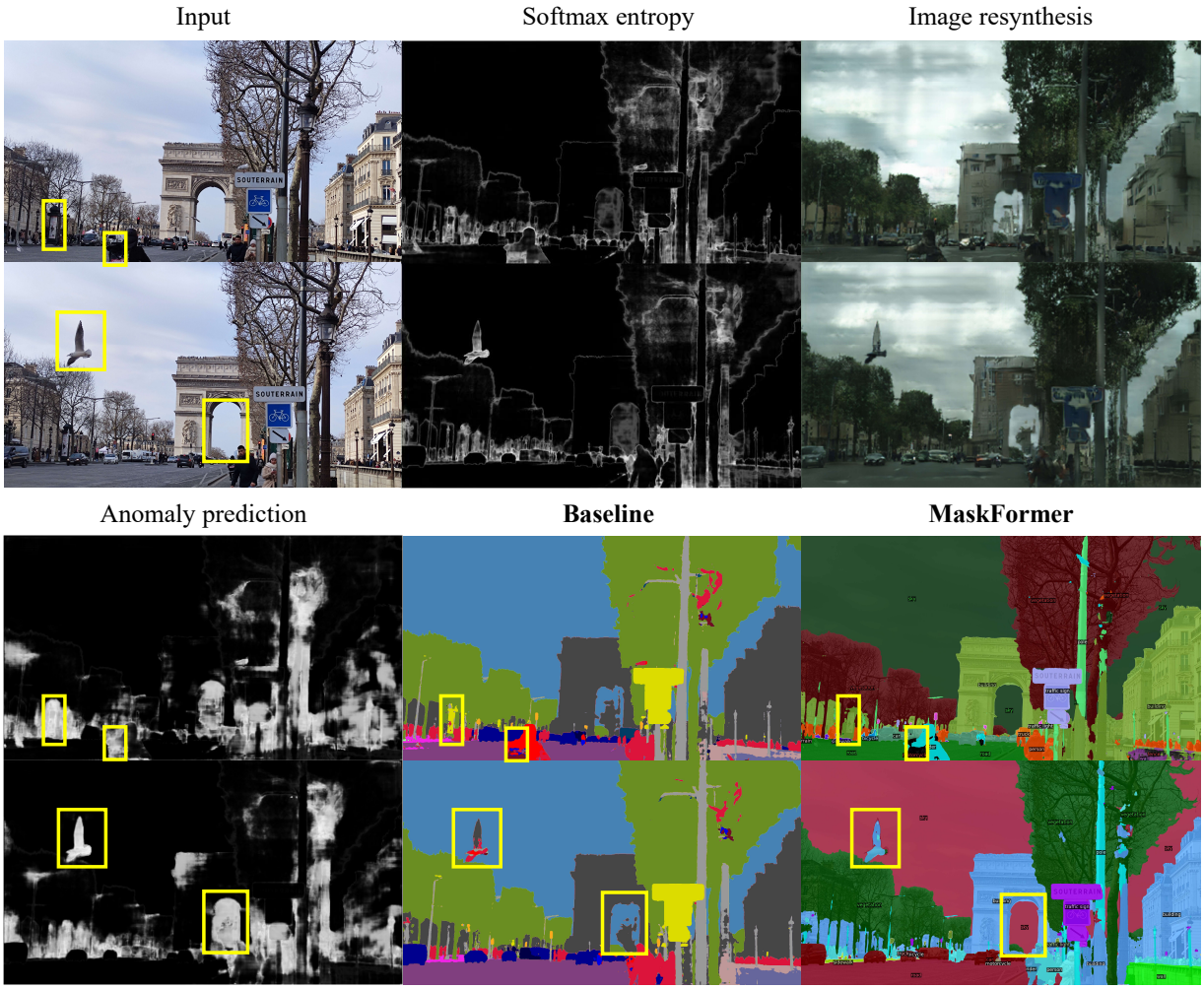


Figure 1: Examples of city scenarios and corresponding uncertainty and semantics prediction results of baseline [1] and MaskFormer [2]. Softmax entropy, image resynthesis, anomaly predictions are generated by the baseline method.

MaskFormer classifies the billboard as construction, and the rider is on the motorcycle with less fragmented areas regarding the top image. Concerning the bottom images, MaskFormer shows better prediction on the area under the Arc with correct sky and vegetation prediction. In terms of the challenging bird instance, MaskFormer achieves correct instance segmentation but with the incorrect label.

In terms of camouflage, the camouflage strategy also deceives the method with SOTA network architecture as shown in Fig. 2d. Compared to the predicted semantic map of the baseline method, MaskFormer predicted large parts of the camouflaged car as the combination of car and construction but showed better performance on the background predictions.

References

[1] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise Anomaly Detection in Complex Driving Scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *NeurIPS*, 2021.

[3] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged Object Segmentation with Distraction Mining," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 8768–8777.