CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	Auto-labelling of pedestrian road crossing from a monocular camera
Author's name:	Jonáš Koditek
Type of thesis :	bachelor
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Cybernetics
Thesis reviewer:	Victor Besnier
Reviewer's department:	Valeo

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **ordinarily challenging**

How demanding was the assigned project?

The assigned project was clear despite being challenging, I think the project is do-able for a bachelor thesis.

Fulfilment of assignment **fulfilled with major objections**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

1. The review and the state of the art is not convincing, the student does not talk about previous, well-known methods for object detection (Yolo family, R-CNN family, DETR, etc) as well as image segmentation (U-Net, Deeplab, etc). The remaining cited work does not seem to be perfectly understood.
2. Nevertheless, the student did select good and relevant architecture (Mask RCNN for object detection + ByteTrack, SegFormer for road segmentation) using Detectron and MMlab.
3. The student successfully introduces a method to automatically annotate the crossing criteria, using object detection + tracking of the pedestrian and semantic segmentation of the road.
4. The student did create a small dataset on BDD, but not KITTI, using good heuristic and analysis using either the Trapezodial Method or the Segmentation for selecting positive and negative data.
5. Training is done using a pre-trained EfficientNet for transfer learning, and a classifier head composed of stacking of MLP. But initial choices (why 9 layers? Which feature from EfficientNet? Why 2 neuronnes for binary classification?) lack of important information. Some elements are missing (amount of parameters, EffiencientNet layer used, etc.). Nevertheless, there are good ablations on learning rate, dropout, optimizer, initialization, etc.
6. Evaluation lack good metrics for safety oriented evaluation as AuROC, FPR95TPR, etc. but remains convincing.

Methodology **correct**

Comment on the correctness of the approach and/or the solution methods.

The dataset creation is correct. The student effectively shows how to select positive and negative data, and how to filter the initial dataset. For the training of the methods, the student did a lot of ablations and provided a good explanation on how he resolved the problem and how he improved the results.

Technical level **C - good.**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

The student shows strong expertise on the training and the inference of different methods on different tasks (segmentation tracking, classification, detection). He also shows rigorous scientific methods to construct the dataset with nice critiques on the different approaches. He shows a good evaluation, but is not using the right metrics for safety oriented problems.

Formal and language level, scope of thesis **E - sufficient.**

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The structure of the document is not satisfactory, especially for chapter 2. Section 2.4 is composed of SegFormer and EfficientNet while the former is not a segmentation method. Section 2.5 Loss and 2.6 Optimizers are poorly explained. And section 2.8 Image Preprocessing should come earlier. The manuscript contains some redundancy sections 3.2.2 and 2.3.2. Nevertheless, the English is clear and easy to read.

Selection of sources, citation correctness **F - failed.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

No, the review of the literature is nearly absent in the thesis: no mentioning of well-known methods in object detection literature such as Yolo, previous work on RCNN, or more recent one (DETR). For semantic segmentation, no mentioning of U-Net, SegNet, DeepLab. The only references are blog posts [10, 12, 18, 42, 44, 45, 40, 35, etc], code repositories [6, 38], reviews [21,18, 15], or irrelevant elements [11, 34, 37].

Additional commentary and evaluation (optional)

Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

The overall quality is satisfactory, the student shows strong knowledge of data processing, training and evaluation of deep learning neural networks for computer vision tasks. He can clearly identify the problems, resolve them and offer good ablation to improve the final solution.
However, the student lacks overall knowledge about the literature and the structure of the writing can be improved. Some technical choices or architecture design could also be explained in more details.

Pending questions:
1/ The student used padding of 0 to have a square image as input to the EfficientNet Feature extractor. Why not simply use bounding box context? In other words: why not crop a square around the bounding box to get the right shape around the pedestrian and not using padding?
2/ Figure 4.1.a & Figure 4.2.a: why is the accuracy suddenly bound at regular intervals?
3/ For transfer learning the student uses a pre-trained EfficientNet and replaces the head. There is no analysis on which layer of EfficientNet is the best to extract the features. This is something that would be interesting to add here. Moreover, can the student explain why he used 9 to 17 layers of MLP+Relu, while it is known that staking MLP performs poorly (specially without batchnorm, residual connection, etc.)? And finally, it would also be interesting to have more insights on the number of parameters and the shape of the input.
4/ Why does the student have 2 nodes as output for the classification head, while only one is necessary (sigmoid and use "> .5" threshold to distinguish the class)?

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS REVIEWER'S REPORT

Additional notes:
The student is mentioning the vanishing gradient for CNN, but it is not entirely true since residual connection allows ResNet to be trained with thousands of layers (section 2.4.1). Later, he states that the Transformer architecture is better and more suitable thanks to attention layers. It is not clear to me why this is the case: how does the attention layer help to avoid gradient vanishing on large images?

The grade that I award for the thesis is **C - good.**

Date: **11.6.2024** Signature: Victor Besnier