THESIS REVIEWER’S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Andrii Zakharchenko**
Author’s name:	**Vizuální vyhledávání obrazů a geolokaliace**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Torsten Sattler
Reviewer’s department:	Czech Institute of Informatics, Robotics and Cybernetics

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

The goal of the thesis is to develop a geo-localization system based on image retrieval. Geo-localization is the task of determining where an image was taken, e.g., in the form of predicting GPS coordinates. Geo-localization is an important sub-problem in many applications, including self-driving cars and other autonomous vehicles. Using image retrieval to tackle the localization task allows to develop light-weight and efficient algorithms as opposed to more complicated approaches based on using 3D models. The thesis thus tackles a problem of practical importance. As part of the thesis, the student had to familiarize himself with the topic, create a novel dataset, re-implement an approach from the literature, evaluate it on the created dataset, and, based on the evaluation, improve the approach by extending it, e.g., via learnable components. Given the breadth of the tasks, the project is clearly suitably challenging for a master thesis.

Fulfilment of assignment **fulfilled**

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.

The thesis fulfils all individual tasks of the project: The thesis presents a new dataset by downloading images taken in the Czech Republic from the photo sharing website Flickr, which also provides the geo-tags necessary for the localization task. The dataset is suitably split into training, validation, and testing. Further, the thesis describes a geo-localization system that combines a retrieval approach based on an existing image-level descriptor with four different strategies to derive a geo-tag for a query image: simply using the geo-tag of the most similar database image, using the mean pose of the top-k retrieved database images, using a weighted version of the mean, and using a kernel density estimate based on the geo-tags of the top-k retrieved images and their descriptor similarity to the query image. The last strategy is based on the Deep IM2GPS approach [Vo et al., Revisiting IM2GPS in the Deep Learning Era, CVPR 2017]. The thesis provides a detailed evaluation of the four strategies on the new dataset. The thesis then modified the kernel density estimate-based strategy, which performed best, by introducing trainable layers that adjust the image representation used for image retrieval such that the descriptor similarities can be better used for kernel density estimation. To this end, the thesis shows how to re-create the estimation process in a trainable / differentiable way and explains how to train the system. The resulting approach is then evaluated on the newly proposed dataset. The evaluation shows that the proposed modifications do not lead to a consistent improvement over the Deep IM2GPS baseline over all accuracy thresholds on the proposed dataset. Still, the results also show that the trainable version of the baseline can improve the results for some thresholds. Since hyperparameters were chosen in a way that fits to the thresholds for which improvements can be observed, I would assume that adjusting these hyperparameters will enable the proposed approach to outperform the baseline on the other thresholds as well. Overall, I thus think that all the primary goals of the thesis have been achieved.

Methodology **correct**

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

The methodology chosen for the individual tasks is appropriate and correct: Using Flickr as a source of data is a common approach in the literature as Flickr is a source of diverse photographs that allows to easily obtain images from large geographical regions (in this case the Czech Republic). The chosen approach to avoid potential strong correlations between images in training and test sets is appropriate. The chosen strategies for geo-localization are meaningful and technically sound. The provided experimental evaluation analyzes the impact of the hyperparameters of each strategy. Adapting the best performing strategy by introducing learnable components that allows training image descriptors that are suitable for kernel density estimation is a natural approach and the chosen implementation is technically sound.

My main point of criticism is about the experimental evaluation and the level of detail at which the new dataset is analyzed. The experimental evaluation summarizes the results, but does not always discuss why we observe the results that are reported. Examples include: is the reason why results are better for images taken in Prague because there is more training data for Prague or due to a lower spatial density of images in other regions (the density at which images are available on Flickr varies per region)? What is the best performance that a nearest neighbor classifier can achieve (i.e., what is the distance to the nearest database image for each query) and how close are the proposed approach and the baselines to this bound? This information would provide intuition about how well these methods work on the dataset. Yet, analyzing experimental results on this level of detail is not absolutely necessary in a master thesis.

Technical level **B - very 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 thesis is technically sound and shows that the student is familiar with (some) techniques from the field of geo-localization and image retrieval. There are some minor errors though: 1) The rankings obtained with the Euclidean distance and the cosine distance should not be “nearly identical” but completely identical as the squared Euclidean distance ||x-y||^2 between two unit vectors x and y is related to the cosine distance c(x,y) via ||x-y||^2 = 2 – 2 * c(x, y), resulting in identical rankings. 2) Given the large standard deviations in Fig. 4.6 and 4.7, I do not think that it makes sense to say that one value is better than the other. 3) I don’t see why using the chain rule is impractical to train neural networks. After all, backpropagation is using it.

The thesis explains what has been done on a higher level, but some details necessary to reimplement the developed approaches are missing. For example, how is the retrieval approach from Radenovic et al. trained (on which dataset, with which settings)? What is the batch size used when training the proposed approach? Which type of search index from the FAISS library is used for nearest neighbor search? Yet, these details do not take away from the contributions of the thesis.

Formal and language level, scope of thesis **C - good.**

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 organization of the thesis is logical and the language is clear and understandable. While there are some typos, the level of English is satisfactory.

My main point of criticism is that while the thesis clearly describes what was done in the project, the motivation for some decisions and parts are missing: Why only use the tag “Czech Republic” and not more specific tags such as “Prague”, “Brno”, etc.? Given the comparably small size of the proposed dataset, does it make sense to train on it from scratch? Why not pre-train the trainable layers on larger datasets and fine-tune on the new dataset? Why was the approach from Tolias et al. used and not other descriptors, e.g., AP-GeM or NetVLAD. Why only evaluate the best model for the 100m threshold in Chapter 4.5? Why were the parameters for the D2W and KDE layers chosen as they were in Chapter 5.3.1? Explaining the motivations behind the decisions taken in a thesis to me is very important, even if they end up as simple as “we tried different versions and this is what worked best in preliminary experiments”, as they give insights into the problem that is being studied. Without them, some of the choices made seem ad-hoc.

In the conclusion, I am missing an outline of future research directions based on the results of this thesis: given the results of the thesis, what would be promising directions for future research on improving geo-localization performance? Which parts could be improved to obtain better results?

Selection of sources, citation correctness **C - good.**

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?

There is quite a lot of literature on the geo-localization problem and it would have been good if the thesis would have discussed this literature. For example, [Pion et al., Benchmarking image retrieval for visual localization, 3DV 2021] also evaluate the task of interpolating geo-tags between the top-retrieved images, taking multiple different image-level descriptors into account. [Thoma et al., Geometrically Mappable Image Features, IEEE RA-L 2020] learn descriptors for image retrieval for which the descriptor distance reflects the geographical distance between the images, which seems relevant. There are multiple datasets commonly used for evaluating the geo-localization task, e.g., Tokyo 24/7, Pittsburgh, San Francisco Landmarks, Mapillary Street-Level Dataset, that make all images available through a central download (in contrast to Flickr-based datasets, where individual images might not be available anymore). It would have been good to motivate the proposed dataset compared to these existing ones. References are missing for the vanishing gradient problem (together with a short explanation of what the problem; given that the thesis explains neural networks on a basic level, this concept should also be explained), the backpropagation algorithm, PyTorch, Adam, and a link to the Flickr website is missing. The approaches from Radenovic et al. and Vo et al., which form the basis of the thesis, could have been explained in more detail (especially given that the thesis explains the basic concepts behind neural networks, I don't think it can be expected of a reader to know these two methods in detail). This would have helped readers without knowledge of both papers to better understand how the proposed approach is related to both and would have made the thesis self-contained. It would have been better to cite the conference versions of papers (which have undergone peer review) rather than the arXiv version of the papers (which have not undergone peer review) and some references are missing the venues where they were published (e.g., [6,9]).

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.

Overall, this is a good thesis. I particularly like the detailed experimental evaluation of the different strategies on the proposed dataset. This high level of detail is important for achieving the goals of the thesis as tuning current approaches is central to understanding their performance on this new dataset. The results are then used to develop an improved and technically sound version of the best-performing method. While I am not concerned that the proposed approach does not outperform the baseline in all situations, I am missing an analysis on why this is the case, leading to a discussion of potential future work. Together with missing motivations for some decisions and missing related work, I believe this is what separates this thesis from the grade "B - very good".

Questions for the presentation and the defense:

1. Is the reason why results are better for images taken in Prague because there is more training data for Prague or due to a lower density of images in other regions (i.e., the closest database images are simply farther away in other regions compared to Prague, thus limiting the accuracy that can be achieved)?
2. Is the proposed dataset large enough to be used for training the learnable layers of the proposed approach? Or would it be better to pre-train them on existing larger datasets?
3. Given your experience in this thesis, what do you think are promising extensions / modifications of your method for improving geo-localization performance?

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

Date: **19.1.2022** Signature: