THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**High-Resolution Images and Knowledge Distillation in Deep Metric Learning with Vision Transformers**
Author's name:	**Yongpan Fu**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Cybernetics
Thesis reviewer:	Bill Psomas
Reviewer's department:	Remote Sensing Lab, National Technical University of Athens

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

The assigned project is an *extension* of the resolution-wise asymmetric metric learning introduced in [2] for transformers. Resolution-wise asymmetric metric learning involves processing images at varying *resolutions* during retrieval. Database images are stored and processed at a *high* resolution to encapsulate greater detail, while query images are processed at a *low* resolution to limit the computational cost. Instead of using solely ViT [4], this project exploits the flexibility of FlexiViT [3], which is able to work at multiple resolutions and patch sizes, without being explicitly trained on these. This way, it presents an extensive study for optimizing the performance-versus-complexity trade-off. The project is characterized by its intricate blend of *theoretical knowledge*, *understanding* and *practical application*. The student had to synthesize existing components and methodologies, but this demanded a significant level of *intellectual engagement* and *technical expertise*.

Fulfilment of assignment **fulfilled with minor 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.

The thesis has effectively met the primary *objectives* set out in the guidelines. Specifically, it achieves the following:

1. The *extension* of the implementation/methodology from [2] to incorporate FlexiViT [3] in the training of both the teacher and the student networks. This adaptation successfully leverages the versatility of FlexiViT to function at multiple resolutions and patch sizes, aligning well with the first goal.

2. The final report presents insightful strategies for optimizing the *performance-versus-complexity trade-off* with transformers. This exploration aligns with the second stated objective, providing valuable *insights* into efficient model optimization.

However, there are minor areas where the thesis deviates slightly from the ideal balance. In section 7, which delves into asymmetric metric learning experiments, the content appears somewhat more condensed compared to the comprehensive treatment seen in other sections. On the other hand, section 3, dedicated to background information, seems slightly overextended. While providing a thorough foundation is commendable, a more concise treatment might have sufficed, allowing for a more balanced distribution of content across the thesis.

Methodology **correct**

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

The thesis extends the methodology from [2] to transformers, with some modifications. It successfully incorporates FlexiViT in the training of both the teacher and the student networks. In equation 13, the total loss calculation differs from the weighted sum in [2], opting for a simple summation, likely for simplification. A relevant comment about equations is that the numbering of equations, such as 10, 11, 12, and 13, groups multiple formulas under a single number, complicating specific references. A distinct numbering for each formula would improve clarity. Furthermore, Figure 12, sourced from [2], only partially reproduces the original, omitting details on different loss components. Including the full figure would provide a more comprehensive understanding of these components and their roles.

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?

Based on the writing style, it appears that the student has a *solid understanding* of deep metric learning and image retrieval, as well as the CNN and transformer architectures. For the most part, the student manages to convey their points clearly.

Some sections create a bit of *confusion*, but this is within a kind of normal range for such complex subjects. In areas requiring explanations or analysis, such as captions or in commenting on results, the student could be more methodical and meticulous or, alternatively, more detailed. For instance, in figure 13, it would be beneficial to have two subfigures: one for (a) Comparison with FlexiViT-S and another for (b) Comparison with FlexiViT-B. In the same figure, the labeling of the models should be clearly specified, enabling the reader to easily correlate that FlexiViT-S/15p refers to FlexiViT-S using a patch size of 15. Additionally, it might be helpful to indicate that the *upper left part* of the cross-dashed lines likely signifies instances where FlexiViT outperforms ResNets both in terms of accuracy and complexity. Continuing with the same figure 13, it would have been beneficial to include commentary on the observation that FlexiViT doesn't seem to gain significantly in accuracy for patch sizes smaller than 20, while the GFLOPS correspondingly increase quite noticeably. Similar remarks apply to figure 15, where, for example, there's no clear reason for patch sizes to be denoted as '/24'. In figure 16, the lines representing ViT are continuous, which might be confusing as it lacks clear points corresponding explicitly to the patch sizes. Regarding table 1, it's not immediately apparent where exactly the model IDs are being utilized, etc.

These observations and suggestions aim to enhance the clarity and interpretability of the visual representations in the thesis. It's important for graphical elements like figures and tables to not only present data but also to elucidate and reinforce the textual content. By addressing these aspects, the thesis could provide a more intuitive and accessible understanding of its findings and conclusions, making it easier for the reader to draw meaningful insights from the visual data presented.

Overall, the student displays commendable comprehension and articulation in their exploration of these advanced topics, yet the inclusion of these finer points would greatly enhance the clarity and impact of their work.

Formal and language level, scope of thesis **B - very 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 thesis demonstrates a mostly proper use of formalisms and notations. However, there is (for example) an inconsistency noted in section 3.2 (page 12), where a PyTorch-like tensor shape notation (B, D, H, W) is used alongside a different, more conventional format (3x3) for dimensions. Consistency in notation, particularly in adopting the more standard (3x3) format across relevant sections, would enhance clarity and professional rigor.

In terms of organization, the thesis is logically structured into sections and subsections. A few recommendations for improvement include: in section 4, it would be beneficial to incorporate comprehensive details about all datasets used, such as ImageNet-1K and ImageNet-21K. Additionally, renaming section 5 to "Methods" and shifting implementation details to section 6, which could be retitled "Experiments," might offer a clearer delineation of content. Within this revised section 6, sub-sections could be introduced to separately address experimental results and their analysis. The current section 7 appears somewhat isolated in its brevity and might be more effectively integrated into section 6.

The presentation quality of the thesis, especially in the initial sections (1, 2, 3), is commendable. It's evident that significant effort has been invested in these parts, resulting in an effective conveyance of the core message. The language used throughout the thesis is clear, and the English is more than satisfactory. It is particularly impressive how the student employs *complex terms* accurately, demonstrating a strong command of the subject matter and its technical vocabulary.

In summary, while the thesis excels in language use and initial presentation, adopting a more consistent approach to formalisms and considering a slight restructuring could further elevate its scientific rigor and overall coherence.

Selection of sources, citation correctness **B - very 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?

The thesis commendably makes adequate reference to earlier work in the field, and the bibliographic citations are up to the required standards. However, there is some room for improvement in the way certain claims are substantiated. A notable example can be found in subsection 5.1.1 (page 25), where the student states: "Various aggregation methods have been explored, including the CLS token, the mean of the patch features alone, and the mean including the CLS token. In practice, it is found that using the output CLS token alone provides the best image representation." This assertion could be strengthened by referencing relevant studies. For instance, [5] presents an extensive study which suggests that, there are cases, in which discarding the CLS token entirely and utilizing the mean of the patch features (global average pooling) might yield a superior representation.

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.

Please insert your comments here.

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

Driven by the imperative to mitigate the test-time complexity inherent in symmetric metric learning, this thesis adeptly explores symmetric and asymmetric metric learning [1]. It delves into resolution-wise [2] approaches, which were previously limited to CNNs.

The extension of this concept to transformers is a non-trivial task, given that Vision Transformers (ViT) [4] are traditionally trained with fixed image resolutions and patch sizes, and tend to underperform when these parameters vary. This challenge is adeptly addressed through the incorporation of FlexiViT [3], a model specifically designed to enhance flexibility in terms of patch size. FlexiViT innovates by utilizing randomized patch sizes during training and a novel technique for resizing patch embeddings, thereby maintaining performance despite varying patch sizes and resolutions.

This thesis presents itself as a sophisticated and well-conceived extension of the work in [2], now applied to transformers, leveraging the adaptability of FlexiViT [3]. It conducts a thorough investigation into the performance-versus-complexity trade-off of FlexiVit, ViT, and ResNet in symmetric and less in asymmetric metric learning scenarios. The approach is characterized by comprehensive experiments that provide valuable insights into these models' behavior in varied learning contexts.

The drawbacks of this thesis might include: inconsistencies in formalisms and notation, insufficient clarity of visual elements (like figures and captions), drawbacks in organizational aspects, particularly the structuring of sections and the delineation of methods and results.

In concluding, I would also like to maybe suggest an expansion of the analysis section in the thesis. While the experiments conducted are comprehensive, they are primarily focused on the performance-versus-complexity trade-off. Additional visualizations could provide deeper insights. For instance, the inclusion of attention maps, which are readily available given the use of ViT models, or t-SNE visualizations for the representations, could be highly beneficial. Examples from query and retrieval results would also be advantageous. These elements are common in both deep metric learning and image retrieval papers, and their inclusion could offer the reader additional insights.

Such enhancements would not only enrich the thesis but also align it more closely with standard practices in the field, thereby augmenting its academic value and applicability.

[1] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. CVPR 2021.
[2] Pavel Suma and Giorgos Tolias. Large-to-small image resolution asymmetry in deep metric learning. arXiv.
[3] L. Beyer, P. Izmailov, A. Kolesnikov, M. Caron, S. Kornblith, X. Zhai, M. Minderer, M. Tschannen, I. Alabdulmohsin, and F. Pavetic. Flexivit: One model for all patch sizes. arXiv.
[4] Dosovitskiy, Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2020.
[5] Psomas, Bill, et al. Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit? ICCV 2023.

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

Date: **24.1.2024** Signature: