# THESIS REVIEWER'S REPORT

## I. IDENTIFICATION DATA

| | |
|---|---|
| **Thesis title:** | **Text-driven Real-time Video Stylization using Diffusion Models** |
| **Author's name:** | **Bc. David Kunz** |
| **Type of thesis :** | master |
| **Faculty/Institute:** | Faculty of Electrical Engineering (FEE) |
| **Department:** | Department of Computer Graphics and Interaction |
| **Thesis reviewer:** | Ing. Ondrej Texler, PhD |
| **Reviewer's department:** | External reviewer |

## II. EVALUATION OF INDIVIDUAL CRITERIA

**Assignment** **challenging**

*How demanding was the assigned project?*

The assignment poses a very interesting task and research direction that has not been explored very well in the past few years. Exactly as the assignment and thesis motivates; diffusion models have, in a way, revolutionized and democratized the style transfer field and allowed for new usecases, but there is not much research effort being directed towards the approach proposed in this thesis: distilling diffusion model into a real-time img2img translation networks. Thus, I believe this direction is certainly worth exploring.

**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 follows and precisely fulfills all parts of the assignment. Also, I believe that the magnitude of experiments, working implementation, and technical contribution is inline, if not exceeding, the expectations for a master thesis.

**Methodology** **outstanding**

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

The decision to use hybrid approach, using IP2P instead of other diffusion models, and running all keyframes through the IP2P in a single batch make sense. The proposed way to mitigate temporal instability and further increase performance (including all implementation optimization) is sound. The proposed technique/implementation was sufficiently tested and evaluated in the Chapter 6.

**Technical level** **A - excellent.**

*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?*

I believe that the thesis is well aligned with the student's field of study, it solves technically challenging problem, and it coherently describes student's efforts, thought process, and contribution.

**Formal and language level, scope of thesis** **A - excellent.**

*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 motivation and problem proposition are very clear. I appreciate the historical context and overview of the related work in Style Transfer and Diffusion space and the fact that most of the text is "self-sufficient" (reader does not need to read other literature and references). Overall, the entire thesis is extensive enough, reads very well, is easy to follow, and has perfect English.

**Selection of sources, citation correctness** **A - excellent.**

*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?*

I believe the references are adequate and the previous work and student's contribution is clearly separated. I only have two very minor comments/suggestions.

Missing citation for PixelShuffle on page 6; I understand that it is hard to cite every single building block and technique, however, as pixel shuffle is somewhat less known, citation would be useful:
[Shi et al., (CVPR 2016)], Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network

I truly appreciate the in-depth overview of the most relevant related work in Chapter 3. There is only one additional technique I could think of
[Geyer et al.,] TokenFlow: Consistent Diffusion Features for Consistent Video Editing
To appear at ICLR 2024 (however, published on arXiv in Jul 2023 already having numerous citations)

**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.*

N/A

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

*As I already expressed in the sections above, I believe the topic is challenging and interesting from the research point of view with potentially a big impact on the artistic community. I also find the thesis to be in very good shape and it undoubtedly deserves the highest grade.*

*I have two questions:*

*(1) Do you have any thoughts on how to make the style transfer technique (StyleVid) even less computationally expensive so that it can run on lower-end GPUs, CPU, or perhaps even a mobile phone? Do you think that making some model architecture changes (e.g., using observations from MobileNet papers) or simplifying the model architecture in any other way could work?*

*(2) I understand that the temporal noise suppression method (that blends previous input with the current input frame) is very fast, practical, and it effectively reduces the temporal noise. But does not it introduce some "ghosting" artifacts in the output video if the subject undergoes fast motion (especially for lower alpha)? As the "t-1" input frame is added to the "t" input frame without being compensated for motion between the "t-1" and "t" frame, I would expect some artifacts.*

The grade that I award for the thesis is **A - excellent.**

Date: **23.1.2024** Signature: