Report

# Computer Vision

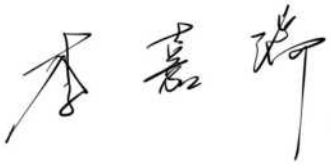

May 11, 2023

Student: Jiarui LI

Student ID: 20216422

Course: COMP3065 Computer Vision

Word Count: 2,142



# Contents

# 1. X-Room

## 1.1 Introduction

Due to the past COVID-19 pandemic, several meetings were transferred from offline to online [5]. Educational meetings are especially affected [3]. Because of the travel limitation during the pandemic, several lectures were held online to make sure all the students can attend. Although, the COVID-19 has been past tense, online lectures are still preferred by students and lecturers, because it can push the boundaries of location and schedule[4]. Everyone can attend and hold a lecture any time and any where. It reduces barriers to knowledge and its dissemination.

Therefore, several companies provided their solutions, such as Google Classroom, Microsoft Teams, and ZOOM [2]. However, these software require the users install application on their laptop and learn how to share screen and change background. It becomes barrier for the lecturers who are not good at computer and caused a lot of troubles.

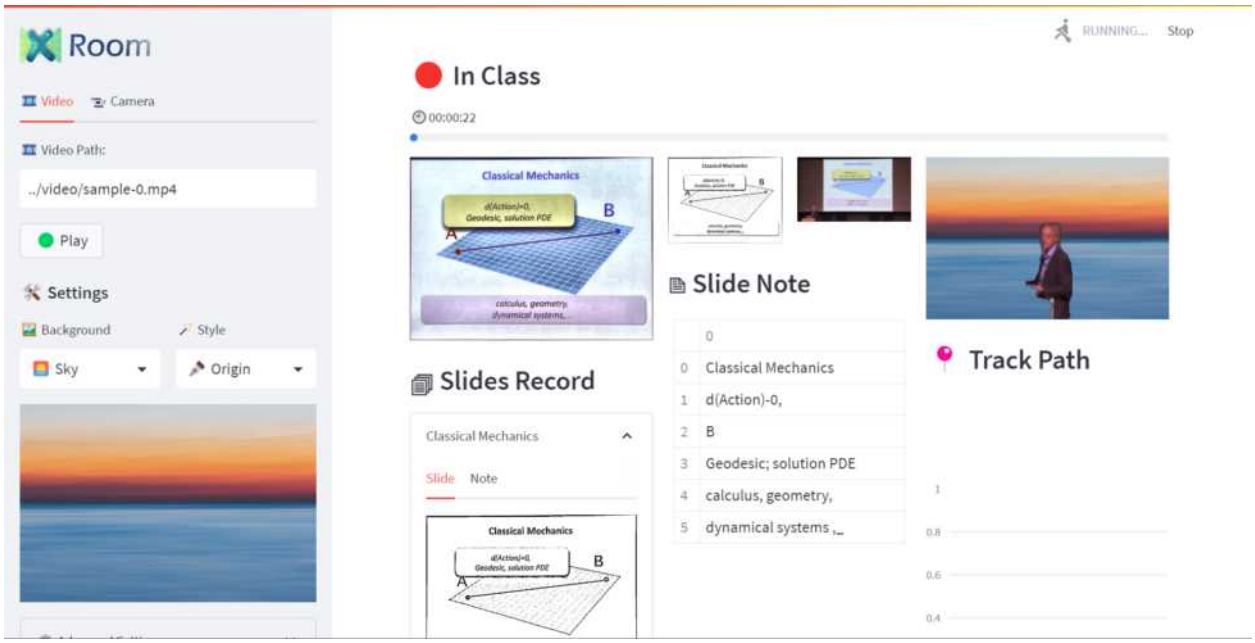



Figure 1: The preview of the X-Room processing existed lecture video.

According to the inconvenience, we propose a new application named X-Room a virtual classroom software. It can provide several advanced functions to make sure it can be easily used by everyone. Compared to the existed software, X-Room only requires video input and analysis the slides and lecturers automatically. This also allows user to process existed lecture video with X-Room. It will automatically analysis the frame, provide users slides with visual enhancement and super resolution process. Also, it will recognize the text on the slides and show plain text to users to assist them make notes. With identify the difference of the slides, it will record the different slides. X-Room can also provide lecturer track and background replacement functions. It will track the lecturers and plot their paths. The lecturers can be segmented from the original video and replace a background they want. Finally, X-Room also provides style filters for users, to make the video as the style they want.

Generally, the features X-Room included are:

1. Lecturers track and plot their paths
2. Lecturer background replacement
3. Slides extraction and enhancement
4. Recognize note from slides and record key slides
5. Provides style filters for lecturer
6. Fetch data from both video and camera

## 1.2 Method

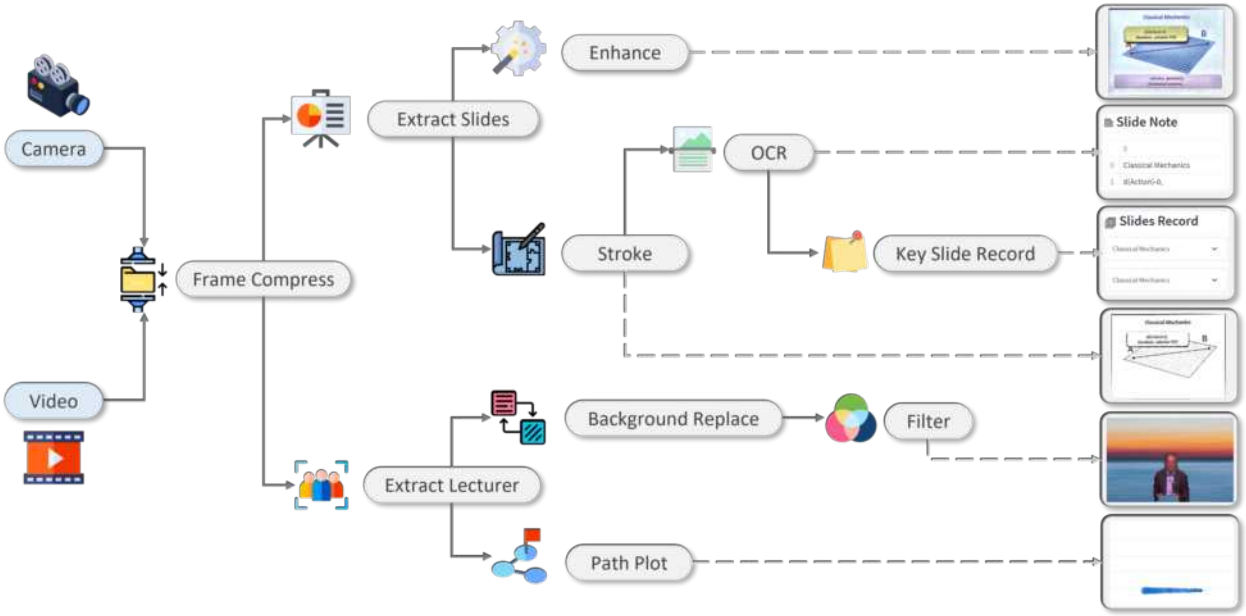



Figure 2: The general method and function flow of X-Room.

The general composition of X-Room is shown in Figure.2. X-Room can read video stream from file or camera. Then, to make sure computers with different capabilities can process the stream smoothly, the stream will be compressed by frame extraction and size zoom. Next, the stream will be passed to two streams for slide and lecturer image separately. For slides stream, the slide will be detected and segmented from the original frame. The segmented slide will be enhanced by sharpen and automatically brighten balance. Also, optical character recognition is applied to make notes from the slides. With difference comparison, key slides will be recorded. For lecturer stream, firstly, we adapt YOLOv8 model to identify and segment the lecturer and then plot the movement path of the lecturer. Also, a user selected background with be fused with the lecturer image and super resolution technique is applied to make the image clearer. Finally, the selected style filter is applied to the image to satisfy the customization requirement of users. All the parameters for these methods can be adjusted by users from the user interface. Also to make sure the program can run smoothly, we created a flow model, which can support asynchronous execution and automatically multi-process.

# 2. Features

The features and contributions of this project are described following the method organization order and are introduced. These features can be generally summarized below:

1. Various video input sources (file and camera)
2. Customized frame extraction and frame resize
3. Recognize and track slides
4. Slides anti-jitter
5. Slides exposure adaptive adjustment and sharpen
6. Recognize slides text and record key slides
7. Lecturer detection and segment
8. Track lecturer and plot path
9. Super resolution process lecturer image
10. Background replacement
11. Customized style filter
12. Adaptive multi-process and asynchronous execution
13. User Interface

## 2.1 Input Video Stream and Compression

The X-Room can accept video stream from both file and camera. This is supported by OpenCV video capture function, which supports both ways to fetch videos. For video compression part, frame extraction and frame resize are applied.

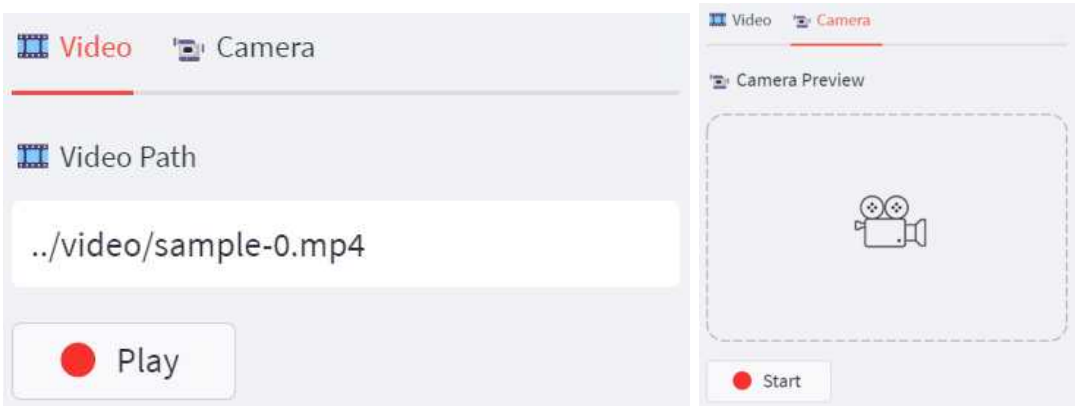



Figure 3: Different video stream input sources.

### 2.1.1 Frame Extraction

To make sure the program can run smoothly on different computers with various capabilities. User can select whether apply frame extraction and the gap for frame extraction. To avoid

the possible stuttering problem, the limitation for this parameter is 10 and to accelerate the program running, it is default set to 5.

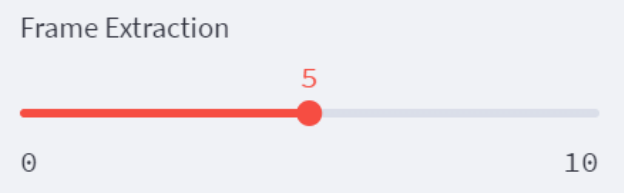



Figure 4: Feature extraction adjustment bar.

### 2.1.2 Frame Resize

The original frame input size is commonly large, which may lead to low frames per second (FPS). Therefore, we allow user to choose scale down the frame to accelerate the processing. It is implemented by OpenCV resize function with apply a magnification to the original image size, range of which is from 0.1 to 1.0 and step by 0.05.


Figure 5: Frame resize adjustment bar.

## 2.2 Slides Process Stream

As an important part for a lecture, this stream is utilized for slides extraction and making notes. Firstly, the frame is processed by canny edge detection and then extract contours from the processed image. The shape which is similar to slides are selected and then tracked with IOU computation. To avoid the possible sharp size wave, the slide anchor with low IOU will be dropped and the transformation of the anchor is applied with a weight to make the change smoothly.

### 2.2.1 Enhanced Slides View

Due to the various of the lecture video quality. The slides may be not clear in some situation. Therefore, we applied a combination of methods to enhance the view of the slides. It includes exposure adaptive method and sharpen.

For exposure adaptive method, firstly, the image is transformed to YCrCb space and the channel Y for illumination is extracted. Then, CLAHE histogram equalization is applied on this channel to adjust the brightness in adaptive. Finally, the updated Y channel is merged back to the image.

To sharpen the images, we utilized a filter with kernel to transform the original input. It can be denoted as:

$$k = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5+\gamma & -1 \\ 0 & -1 & 0 \end{bmatrix}, \quad \gamma \in [-0.5, 0.5] \tag{1}$$

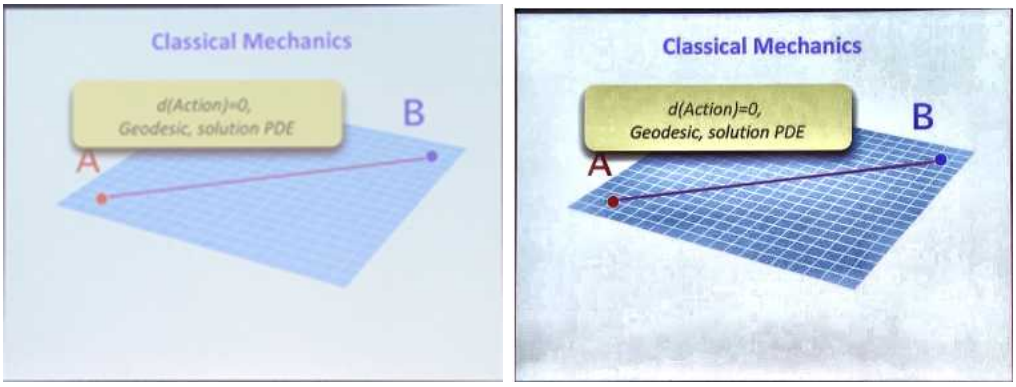



Figure 6: Origin view and enhanced view of the slides.

$$\hat{I}_{x,y} = \sum_{0 \le x' < 3, 0 \le y' < 3} k_{x',y'} \times I_{x+x'-1, y+y'-1} \tag{2}$$

where $\gamma$ denotes the strength of the sharpen.

### 2.2.2 Slides Note and Record

Making note requires user to type the text read from the slides, which is time-consuming and inconvenience. Therefore, we adapt optical character recognition to translate the slides from image to text that can be copied. Then, record the key slides for after class review.

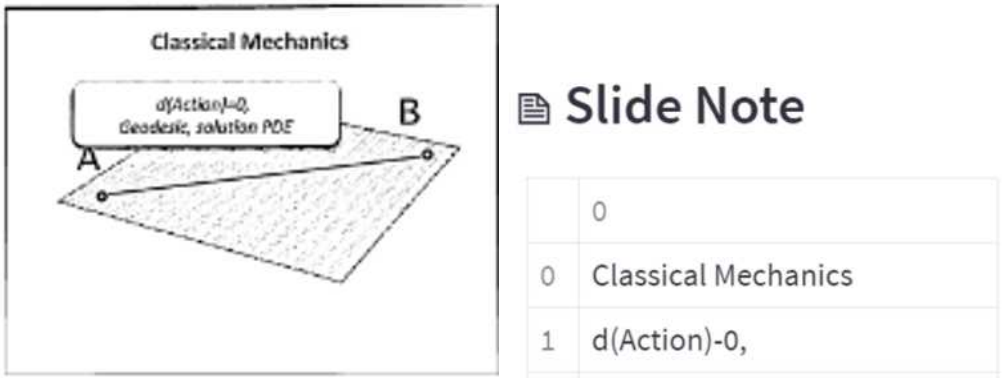



Figure 7: The stroked slides and related optical character recognition result.

Firstly, the image of the slide is stroked to reduce noise for continuous recognition. Then, a VGG-BiLSTM-CTC model of optical character recognition (OCR) is applied to identify the information from the slides [1].

Because of the speed operating OCR, we will only apply OCR on different slides instead of each frame. The difference is determined by mean squared error (MSE) which denotes as:

$$MSE = \frac{\sum_{i=0}^{n} (I_i - \hat{I}_i)^2}{n} \tag{3}$$

where $I$ and $\hat{I}$ denote two images. Only the MSE of two frame surpass the set threshold, the OCR will be applied to the slides. With the difference compared according to MSE, we record the slide and is text note, when MSE surpass the threshold and recognize the slide as a new slide. Also, both the threshold and OCR language can be modified by users. The default threshold is 800 which can distinguish the frames best and OCR default accept English.

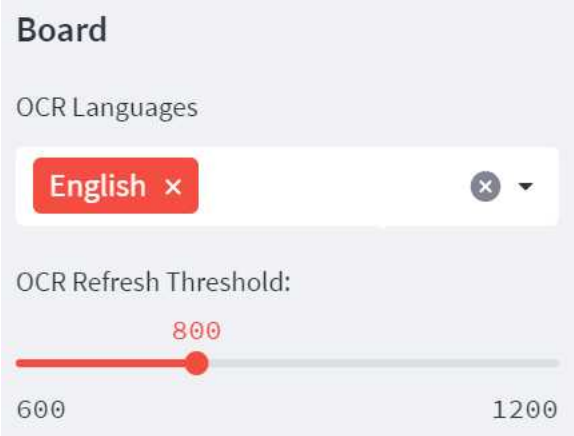



Figure 8: Slide change threshold and optical character recognition language setting.

## 2.3 Lecturer Process Stream

This stream is designed to process the image of the lecturer. Firstly, the lecturer is detected and segmented with YOLOv8 model [7]. Then the mask will be applied on the frame to extract the lecturer part and then clip lecturer part with the anchor produced by the model.

### 2.3.1 Lecturer Track and Path Plot

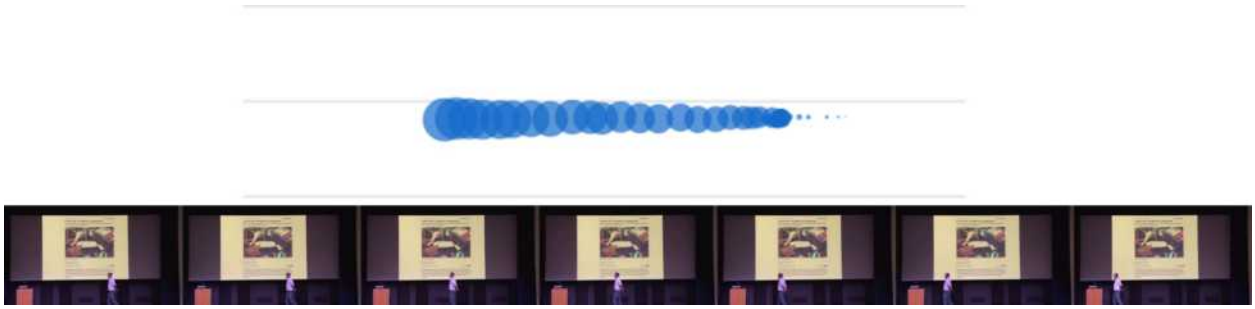


Figure 9: Lecturer path tracking plot.

With the detection result of YOLOv8, we match the anchor to specify person with IOU comparison. And then, the path of the lecturer is drawn on a figure. The figure is plotted with the center of the lecturer anchor and the size of the scatter will decrease following the frames order.

### 2.3.2 Super Resolution

Because of the limitation of the clarity, we utilized Deep Recursive Residual Network with 9 residual blocks (DRRN_B1U9) to apply super resolution on the lecturer image with trained weights from the internet [6].

### 2.3.3 Background Replacement and Style Filter

After enhanced by the super resolution processing, users can select backgrounds and automatically replace the background of the original frame. This is implemented by a procedure for image overlapping. Firstly, a mask for the lecturer will be produced with extract the

Figure 10: Lecturer fused with different background including sky, forest, mountain, night, and valley.

lecturer from its gray image with a threshold set. Then the mask will be applied on the background image to remove the part for the lecturer. Finally, the lecturer will be attached to the part removed and comprised the fused image.

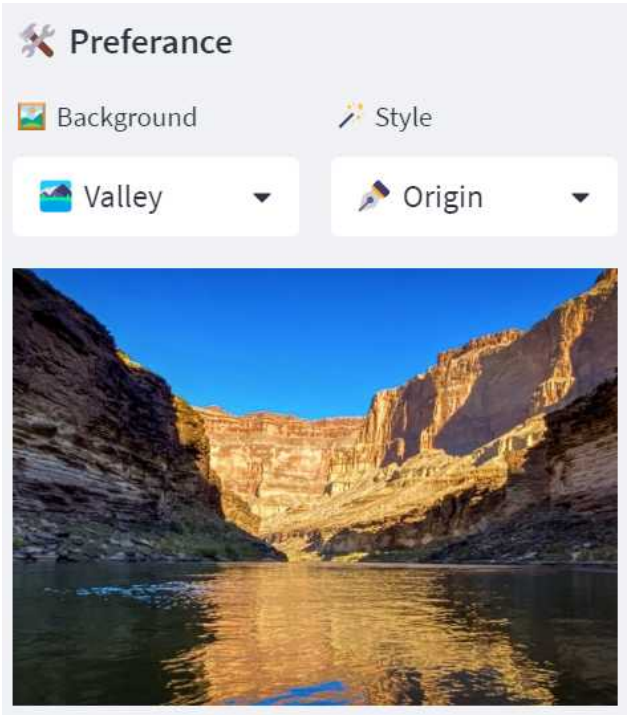



Figure 11: The background and style filter selector.

Then the users selected style filter will be applied to the image to generate the final output. The warm filter is applied a kernel to transform the image. The gray style is implemented by transform the frame to gray scale. And the sketch style is implemented by adaptive threshold with constant subtracted from the mean is set to 1.

## 2.4 Flow Model and Multi-process

As a project, to help build the image processing model easy, maintainable, and can be accelerated. A flow model architecture is created.

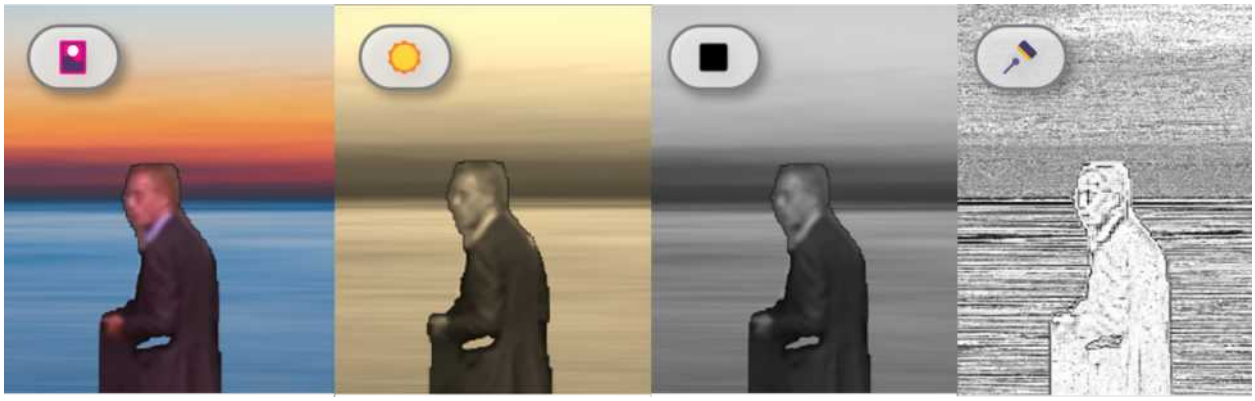


Figure 12: The effects of different filters including origin, warm, gray, and sketch.

### 2.4.1 Flow Model

All the flow model blocks is constructed from a template ($FlowModule$), which requires init and forward methods. Then, two basic organization block is inherited from it, including flow list ($FlowModuleList$) and flow branch ($FlowModuleBranch$). The flow list can accept a list of flow models and run them in order and count the time each model consumed. The flow branch allows split the stream into different parts and execute them parallel. Also, user can select whether use multi-process and the limitation of the thread numbers.

### 2.4.2 Multi-process

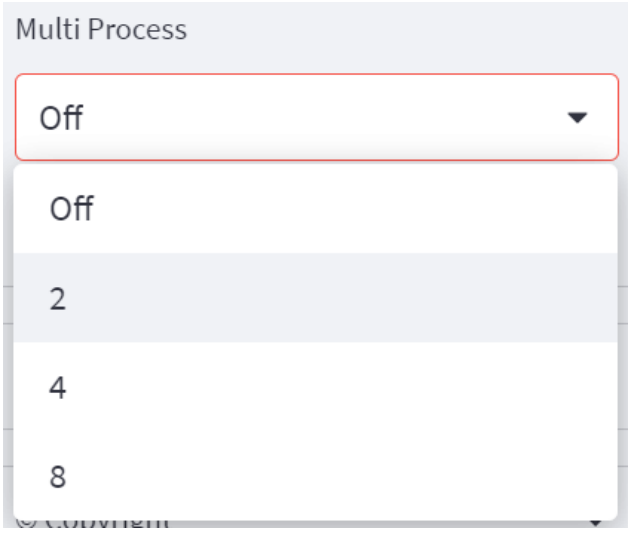



Figure 13: The options for multi-process.

Due to the speed for processing one frame is slow, to accelerate the processing, we add multi-process function to the flow branch block. Because the branch block holds streams parallel, results of which will not affect other streams, these streams will be executed by different threads to accelerate.

### 2.4.3 Asynchronous Execution

Because some of the tasks such as OCR are extremely time consuming and do not require real-time processing. We provide a asynchronous flow model block ($FlowModuleAsync$) to

process these tasks to avoid they stuck the whole procedure. To utilize this block, just need to wrap a normal flow model module with asynchronous block. The program will not wait for the result from the block any more, but check whether the task finished turn by turn. And return the result after it finished.

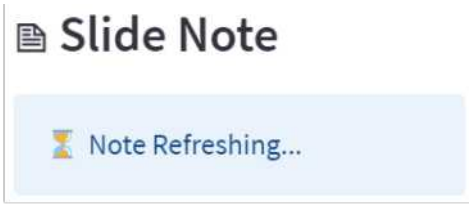



Figure 14: The slides note making block is executed asynchronous.

## 2.5 User Interface

Finally, all the functions are wrapped with a user interface implemented with streamlit. It can provide visible interaction and view of the functions that are mentioned previously. Also it supports both light and dark modes and can change according to the system setting.

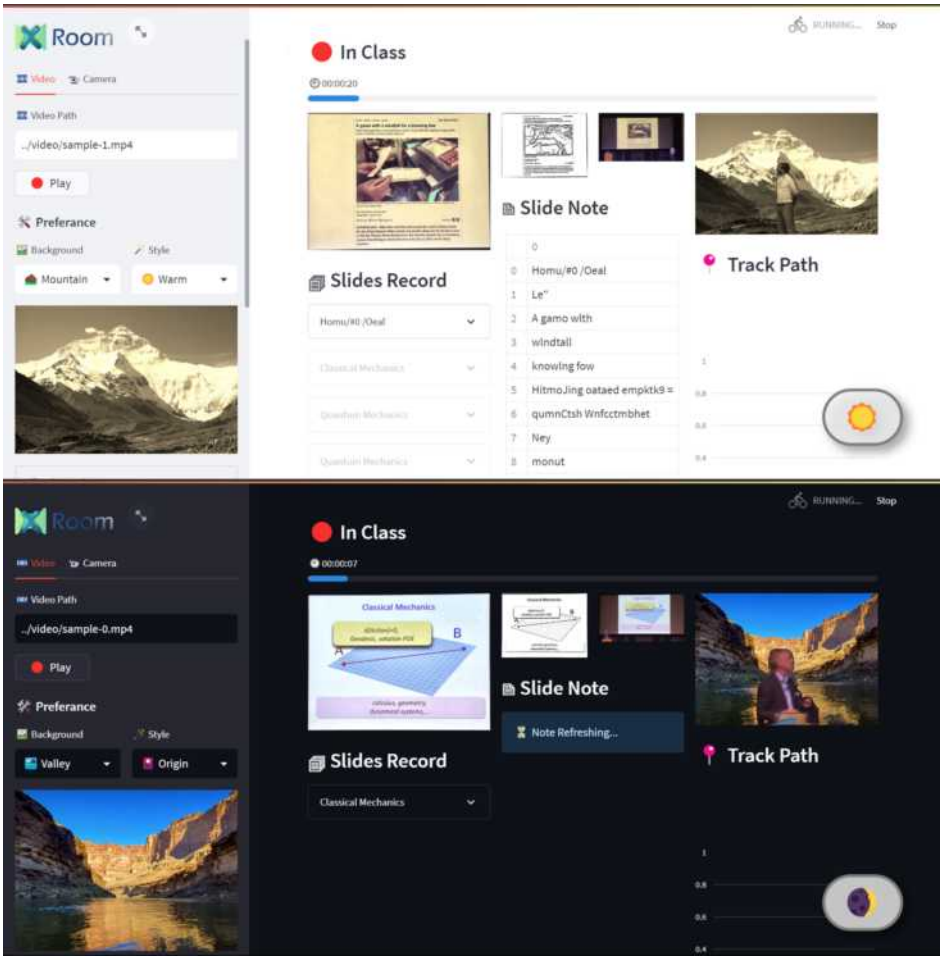



Figure 15: The different color themes for the user interface.

# 3. Results and Discussion

Based on the methods mentioned, we are going to discuss the effects of the methods and analyze the weakness and advantages of these methods.

## 3.1 Frame Compress

Firstly, we are going to discuss the affects caused by frame compression methods.

### 3.1.1 Frame Extraction

With different gap of frame extraction, the frequency of the video is influenced. Although, frame extraction can help reduce the computational resource requirement, but it will also affect the viewers' feeling.


Figure 16: The influence caused by frame extraction to the frequency.

### 3.1.2 Frame Resize


Figure 17: The frames per second and the clarity of the different scale down magnification.

To accelerate the computation and preserve the performance of computers. We allow user to select the magnification for frame scale down. It can significantly accelerate the computation speed, however, it will make the clarity of the outputs reduced.

## 3.2 Slides Track and Note

Then, the improvement and side-effects of the slides processing stream is going to be discussed.

### 3.2.1 Slides Anti-jitter

With smooth transformation with a gradient factor and drop threshold for sharp change, the slides can be tracked more stably and smoothly. However, this may cause the algorithm cannot track the slides when it changed sharply.

### 3.2.2 Slides Enhancement

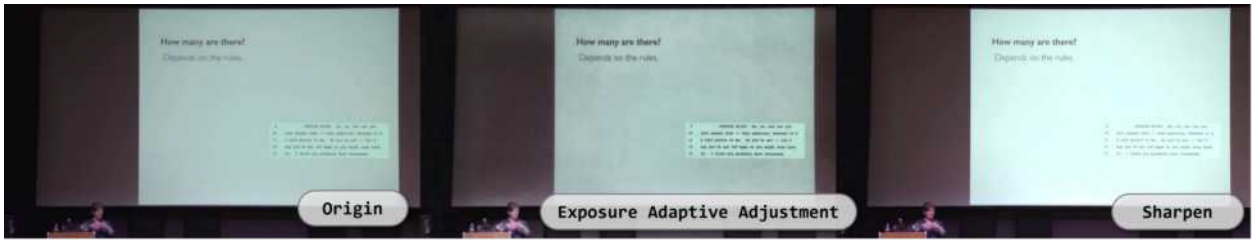
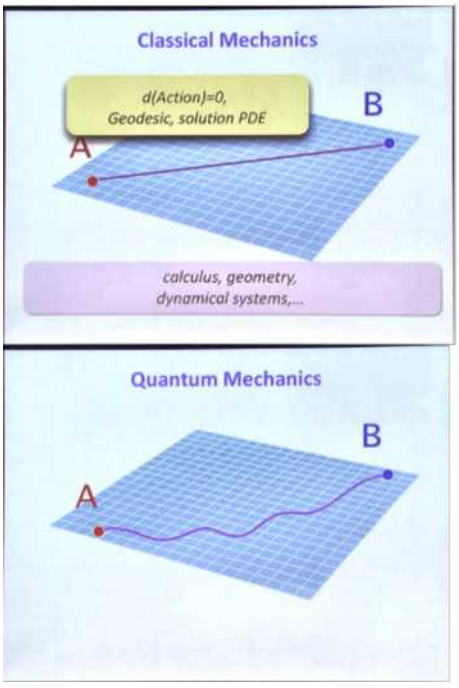



Figure 18: The results from exposure adaptive adjustment and sharpen.

Obviously, with the enhancement methods, the qualities of the images are improved significantly. Especially the exposure adaptive adjustment, it balanced the exposure pixel by pixel instead of whole image.

### 3.2.3 Slides Note Making




| | 0 |
|---|---|
| 0 | Classical Mechanics |
| 1 | d(Action)-0, |
| 2 | B |
| 3 | Geodesic; solution PDE |

| | 0 |
|---|---|
| 0 | Quantum Mechanics |
| 1 | B |

Figure 19: The text recognition and slides record results.

According to the result, the text recognized from the slides is generally accurate. However, it is time-consuming and requires high computer capabilities, which may not be able to handle by all computers.

## 3.3 Lecturer Track and Background Replacement

With the YOLOv8, the lecturer can be detected and segmented accurately. However, it is time-consuming and computational resource thirst. With the demonstration, the detection and segmentation works correctly and accurately.

Figure 20: The segmentation and detection result.

### 3.3.1 Super Resolution

Due to the limitation of the original frame clarity. The super resolution methods are adopted. It can significantly improve the quality of the image, however, it is time-consuming and also computational resource thirst.


Figure 21: Comparison between original image and super resolution processed image.

# 4. References

Bibliography


[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4715–4723, 2019.

[2] Suvashis Dash, Saugata Samadder, Anurag Srivastava, Rajendra Meena, and Piyush Ranjan. Review of online teaching platforms in the current period of covid-19 pandemic. *Indian Journal of Surgery*, 84(Suppl 1):12–17, 2022.

[3] Daniel Hermawan. The rise of e-learning in covid-19 pandemic in private university: challenges and opportunities. *IJORER: International Journal of Recent Educational Research*, 2(1):86–95, 2021.

[4] Francesco Porpiglia, Enrico Checcucci, Riccardo Autorino, Daniele Amparore, Matthew R Cooperberg, Vincenzo Ficarra, and Giacomo Novara. Traditional and virtual congress meetings during the covid-19 pandemic and the post-covid-19 era: is it time to change the paradigm? *European urology*, 78(3):301, 2020.

[5] Hendri Pratama, Mohamed Nor Azhari Azman, Gulzhaina K Kassymova, and Shakizat S Duisenbayeva. The trend in using online meeting applications for learning during the period of pandemic covid-19: A literature review. *Journal of Innovation in Educational and Cultural Research*, 1(2):58–68, 2020.

[6] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[7] Juan Terven and Diana Cordova-Esparza. A comprehensive review of yolo: From yolov1 to yolov8 and beyond. *arXiv preprint arXiv:2304.00501*, 2023.