The Australian National University
College of Engineering and Computer Science Final Examination, First Semester 2021

ENGN6528 Computer Vision

Question Booklet

Instructions on next page

Allotted Time

You will have xx hours to complete the exam plus 15 minutes of reading time. An additional 15 minutes has also been allowed to accommodate the additional task of uploading your completed exam to the final exam turnitin submission portal on the ENGN6528 Wattle site. Thus, you have xx hours to complete the exam. NO late exams will be accepted. You may begin the exam as soon as you download it.

Minimal requirements:

You may attempt all questions

You SHOULD NOT include an assignment cover sheet

You must type your ANU student identification number at the top of the first page of your submission

You must monitor your own time (i.e. there is no invigilator to tell you how many minutes are left).

Your answers must be clear enough that another person can read, understand and mark your answer. 11 or 12 point font with 1.5 spacing is preferred. Scanned images of handwritten equations or diagrams must be legible and of a suitable size.

Numbering questions

- You must specify the question you are answering by typing the relevant question number at the top the page
- Each question should begin on a new page
- Multi-part questions (e.g. question 1 parts a and b) may be addressed on the same page but should be clearly labelled (e.g. 1a, 1b)
- Questions should be answered in order

You must upload your completed answers in a single document file within the allotted time using a compatible file type for Turnitin (Preference: MS Word's .doc or .docx format) **It is the student's responsibility to check that the file has uploaded correctly within Turnitin. No late exams will be accepted.**

Academic integrity

Students are reminded of the declaration that they agree to when submitting this exam paper via Turnitin:
I declare that this work:

- upholds the principles of academic integrity as defined in the University Academic Misconduct Rules;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;

- is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

There are 5 questions in total.
(Q1-Q5)

Please name your submission as
ENGN6528_exam_u1234567.docx

Questions on the next page

Q1: (21 marks) [3D SFM and Image formation question]

Answer the following questions concisely. Write down working, and if you are unsure about some part along the way, state your best assumption and use it for the remaining parts. Similarly, if you think some aspect is ambiguous, state your assumption and write the answer as clearly as you can.

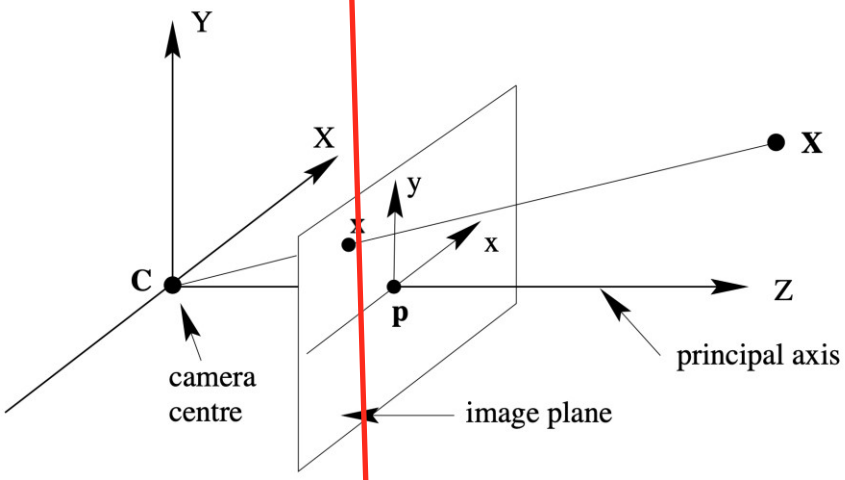
(a) Given two calibrated cameras, C1 and C2, C1 has focal length of 500 in x and 375 in y, # pixels / meter. (in pixel unit) the camera has resolution 512x512, and the camera centre projected to image is at (249, 249), with no skew. Suppose C2 has the same image resolution and focal length as C1, but the camera centre projected to image is at (251, 252). Write down the calibration matrix K1 and K2 for C1 and C2 respectively. (Hint: please only write down the final two 3x3 matrices.) [3 marks]

Answer:

$$\mathrm{K}_1 = \begin{pmatrix} 500 & 0 & 249 \\ 0 & 375 & 249 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathrm{K}_2 = \begin{pmatrix} 500 & 0 & 251 \\ 0 & 375 & 252 \\ 0 & 0 & 1 \end{pmatrix}$$

(b) Suppose that a 3D world coordinate system ((X,Y,Z) coordinates as in the below diagram from the lecture notes) is defined as aligned with the camera coordinate system of C1. More specifically, the world origin is at the camera centre of C1, the Z axis is aligned with the optical(principal) axis and the X and Y world coordinate systems aligned parallel with the x and y axes of the image of C1. Write down the matrices K[R|t] which define the projection of a point in world coordinate system to the image of C1. (Hint: please only write down the final 3x4 matrix.) [3 marks]

no extrinsic



$$\mathrm{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathrm{t} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \text{therefore,}$$

cam 1 $$\mathrm{K}[\mathrm{R} \quad \mathrm{t}] = [\mathrm{K}_1;\ \mathbf{0}] = \begin{pmatrix} 500 & 0 & 249 & 0 \\ 0 & 375 & 249 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

(c) Suppose that the scene has a point, P1, that in the world coordinate system defined above that lies at (39, 35, 100). Note that the points in world coordinate system are measured in cm. What location (to the nearest pixel) will that world point (P1) map to in the image of C1? [2 marks]

$$\mathrm{K}[\mathrm{R} \quad \mathrm{t}]\mathbf{P}_1 = \begin{pmatrix} 500 & 0 & 249 & 0 \\ 0 & 375 & 249 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 39 \\ 35 \\ 100 \\ 1 \end{pmatrix}$$

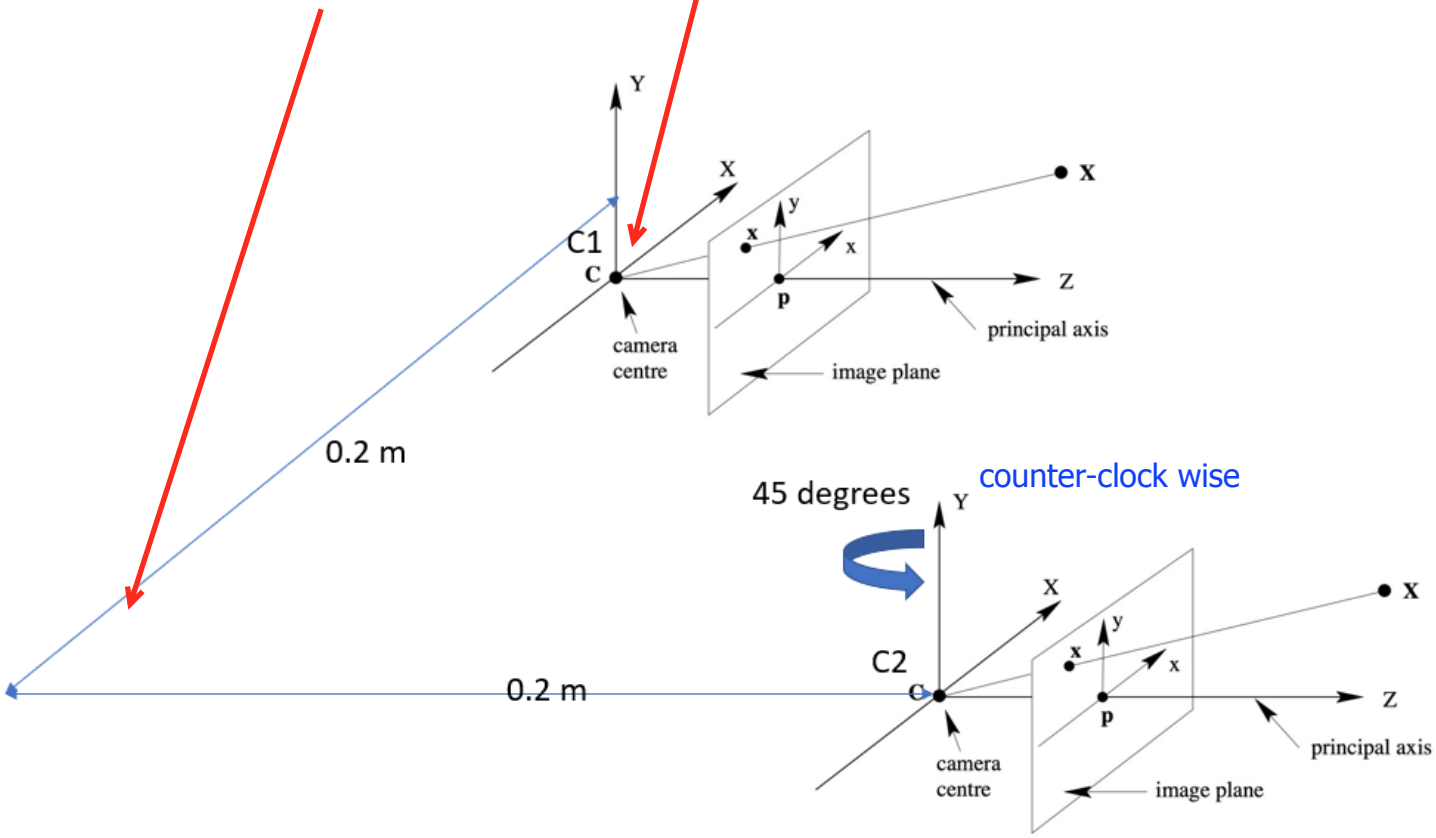
$$\tilde{x} = 5 * 39 + 249 = 444$$
$$\tilde{y} = 375 * 35 + 249 = 380.25$$

3.75

We thus have the rounded pixel coordinate as

$$\tilde{x} = 444$$
$$\tilde{y} = 380$$

(d) Suppose that with respect to the world coordinate system that is aligned with camera C1, camera C2 begins being aligned to C1, and is then rotated by 45 degrees about its vertical axis (Y-axis)(as shown below), and subsequently the centre of C2 is translated by 0.2 m to the left of C1 (along the X axis of C1), then moved forward by 0.2 m parallel to the optical axis of C1.



Write down the matrices K[R|t], which define the projection of points in the world system (i.e, the same coordinate system of C1) to the image of C2. (Hint: please only write down the final 3x4 matrix.) [3 marks]

$$K_2 = \begin{pmatrix} 500 & 0 & 251 \\ 0 & 375 & 252 \\ 0 & 0 & 1 \end{pmatrix}$$

intrinsic keeps the same

sin(-45) = - sin(45), counterclock-wise

$$R_2 = \begin{pmatrix} cos(45°) & 0 & -sin(45°) \\ 0 & 1 & 0 \\ sin(45°) & 0 & cos(45°) \end{pmatrix}$$

$$t_2 = R_2 \begin{pmatrix} 20 \\ 0 \\ -20 \end{pmatrix}$$

$$K_2[R_2|t_2] = \begin{pmatrix} 531.037 & 0 & -176.07 & 14142.1 \\ 178.191 & 375 & 178.191 & 0 \\ 0.707 & 0 & 0.707 & 0 \end{pmatrix}$$

0.2 is changed to 20 because "the points in world coordinate system are measured in cm".

(e) What is the location (to the nearest pixel) that P1 maps to in the image of Camera C2? (Hint: Please write down only the final result.) [2 marks]

$$w_2 \begin{pmatrix} u_2 \\ v_2 \\ 1 \end{pmatrix} = K_2[R_2|t_2]P_1 = \begin{pmatrix} 531.037 & 0 & -176.07 & 14142.1 \\ 178.191 & 375 & 178.191 & 0 \\ 0.707 & 0 & 0.707 & 0 \end{pmatrix} \begin{pmatrix} 39 \\ 35 \\ 100 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 17245.6 \\ 37893.5 \\ 98.2878 \end{pmatrix}$$

$w2 = 98.2878$

$$u_2 = \frac{17245.6}{w_2} = 175.46$$

in pixels

$$v_2 = \frac{37893.5}{w_2} = 385.53635$$

(f) Define the term epipole. [2 points]

The epipole is the point of intersection of the line joining the optical centres, that is the baseline, with the image plane. Thus the epipole is the image, in one camera, of the optical centre of the other camera.

Given a setup consisting of two cameras, the epipole is defined as the projection of one camera centre at the image plane of the other one.

(g) For camera C1, there is an epipole (or epipolar point) that relates to Camera C2. For the two-camera setup for predicting structure from motion, what is the position of the epipole in camera C1 of camera C2? (Hint: It is a point in the image coordinates of Camera C1). [2 points]

Denote X_1 as the coordinates under camera C1, and X_2 as the coordinates under camera C2.

According to (d),

$$X_2 = R_2 X_1 + t_2.$$

Then, we have:

$$X_1 = R_2^{-1}(X_2 - t_2) = -R_2^{-1} t_2$$

$$w_1 \begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} = K_1(R_1 X_1 + t_1) = K_1(-R_1 R_2^{-1} t_2 + t_1)$$

Solve u_1, v_1 as in (e).

(h) Given a point P2 that appears in camera C1 at image location (x1, y1), and in camera C2 at image location (x2, y2). How would you find the world coordinates of point P2? [4 points]

Plz refer to the slides of "ENGN6528-Week09-3DVision.pdf", page 28.

- Direct analogue of the linear method of camera resectioning.
- Given equations

$$\mathbf{x} = \mathbf{P}\mathbf{X};\ \mathbf{x}' = \mathbf{P}'\mathbf{X}$$

- $\mathbf{p}^{i\top}$ are the rows of P.
- Write as linear equations in $\mathbf{X}$

$$\begin{bmatrix} x\mathbf{p}^{3\top} - \mathbf{p}^{1\top} \\ y\mathbf{p}^{3\top} - \mathbf{p}^{2\top} \\ x'\mathbf{p}'^{3\top} - \mathbf{p}'^{1\top} \\ y\mathbf{p}'^{3\top} - \mathbf{p}'^{2\top} \end{bmatrix} \mathbf{X} = 0$$

- Solve for $\mathbf{X}$.
- Generalizes to point match in several images.
- Minimizes no meaningful quantity – not optimal.

Replace x →x1, y→y1, x'→x2, y'→y2.

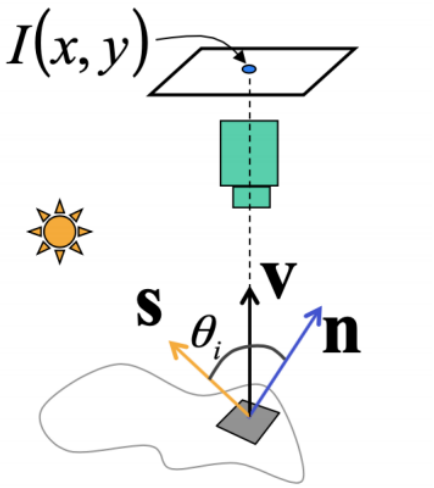
P is the matrix computed in (b), and P' is the matrix computed in (d).

Q2: **(7 marks)** [Shape-from-X, Stereo]

(a) Shape-from-Shading approaches predict the brightness of an image pixel. Given a point light source at infinity (distant light source), write down the equation that defines the brightness at an image pixel assuming that the camera views a Lambertian surface, Please also define the terms of the equation. [2 marks]

Please refer to our slides ENGN6528-3DVision-SFS-PMS-updated.pdf, page 41.

Image formation for Lambertian Surface



- Relate image irradiance I(x,y) to surface orientation
(p,q,-1) for a given light source direction (p_s,q_s,-1) an
surface reflectance.
- Lambertian case:

k : light source brightness
ρ : surface albedo (reflectance)
c : constant (optical system)

- Image irradiance:

$$I = \frac{\rho}{\pi} kc \cos\theta_i = \frac{\rho}{\pi} kc\mathbf{n} \cdot \mathbf{s}$$

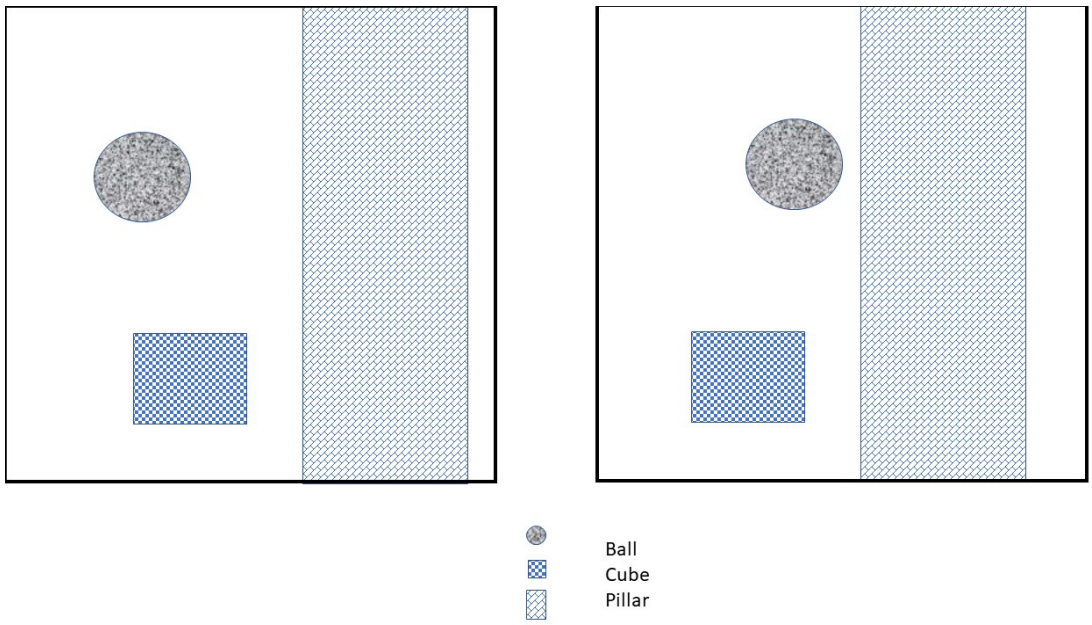
If $\frac{\rho}{\pi} kc = 1, I = n \cdot s = cos\theta_i$

If the computed value is less than 0, then set it to 0.

(b) Suppose that we have used some other methods to know the brightness of the lighting, its direction and the reflectance properties of the surface in the above scenario, but we only have intensity information about this particular pixel for this surface, what can we say about the surface orientation? [2 marks]

The surface normal is lying on a cone.

(c) The images (a and b) shown below are the left, and the right image of an ideal stereo pair, taken with two identical cameras (A and B) mounted at the same horizontal level and with their optical axes parallel. [3 marks] same intrinsic



Draw a planar-view (i.e., a top-down bird-eye's view) of the scene showing roughly what the spatial arrangements of the three objects are. Only relative (rather than accurate) positions are required.

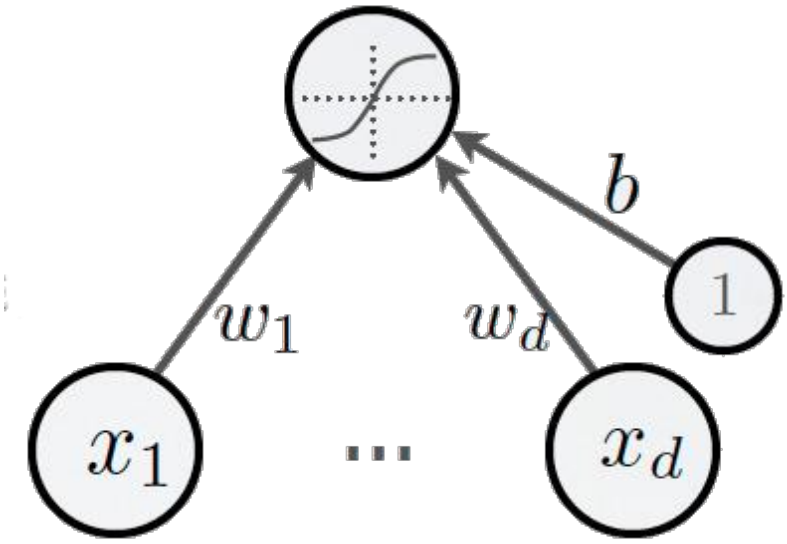
more shift in two views, closer the object is

The larger of the disparity, the nearer of the object.

Camera Ball cube pillar

Q3: (8 marks) [Deep neural network]

Given below is a single node in a neural network. Supposing that d is 4, x={2,1,2,3}, and w={0.3,0.4,0.1,-0.4}, b=0.1, and that the activation function is a standard ReLU, that is =max(0,x), where x is the input to the activation function.



(a) What is the output of this node? [2 marks]
2*0.3+1*0.4+0.1*2-0.4*3+0.1 = 0.1

(b) Describe the difference between, recognition and detection in terms of how you would use a Deep Convolutional Network to solve the problem? [2 marks]

The output of a deep CNN for image recognition task is the probability of one image belonging to one class. By contrast, the output of the deep CNN for object detection is the class probability associated to each bounding box as well as the bounding box center and its dimension.

(c) Two cascaded 3x3 layers, or a single 5x5 layer result in the same number of pixels in the input image impacting the result. So why might you prefer one representation over the other? [2 marks]

Two cascaded 3x3 layers can model more complex functions by inserting the non-linearity activation function between two 3x3 layers.

Q4: (2 marks) (questions with short answers) Given a dataset that consists of images of the Eiffel Tower, your task is to learn a classifier to detect the Eiffel Tower in new images. You implement PCA to reduce the dimensionality of your data, but find that your performance in detecting the Eiffel Tower significantly drops in comparison to your method on the original input data. Samples of your input training images are given in the following figures. Why is the performance suffering? [hints: describe in two sentences.]

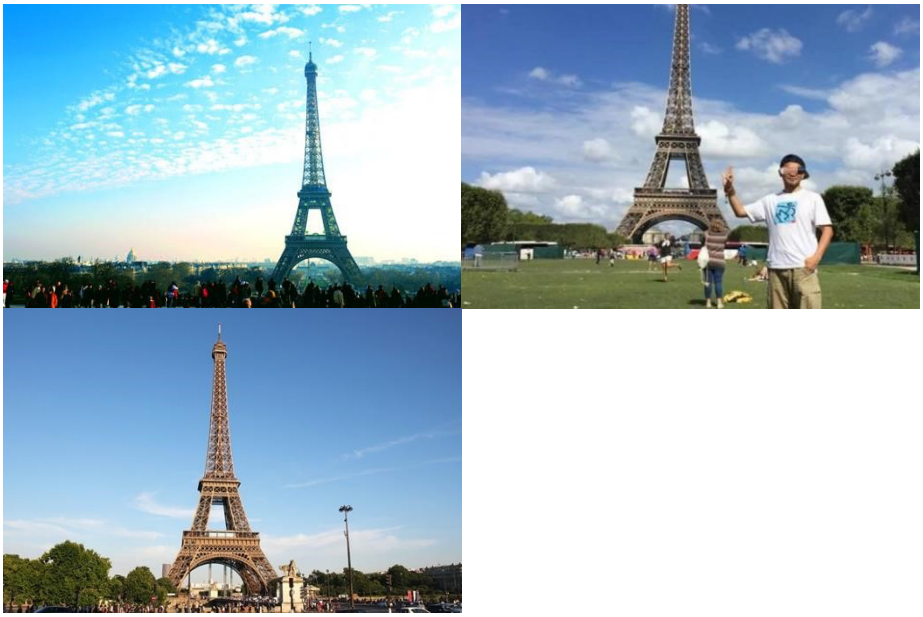


Figure 1. Images in the dataset

All images have very diverse background. The Eiffel tower is not aligned across images. This will introduce noise to image data samples. The images are then cannot be represented by low dimensional data.

Q5: (10 marks) [algorithm design] Turn your phone into a GPS in an art museum or a library. GPS usually does not work well in an indoor environment. The goal of designing this algorithm is to localize your position by taking a few images around you in the museum. Please Briefly describe the key steps of your method.

Localize yourself

Step 1: Collect an image dataset for the scene. Here we use a camera to capture images within this museum. Assume that the image dataset consists of N images

We can calibrate the sensor namely we assume that we know the intrinsics which allows us knowing the focus length, principal points.

Step 2: We aim to build a visual map using these images. In particular, we are going to adopt the structure from motion pipelines (details listed here, plz refer to slides ENGN6528-3DVision-MVS-SFM-OF.pdf). Output the camera poses of all images relative to a reference one.

Step 3: When the user is inside a building to find out its location, the user is required to take one image with textures.

Step 4: This image is used to retrieve similar images from the dataset. We then perform local structure from motion by including the image taken at the user's current location.

======= END of ALL QUESTIONS in the EXAM ==========