Q1. [1 mark] Image formation. Answer ANY ONE of (a) or (b).

a) Name or describe two Gestalt principles observed commonly in scenes around us. Through examples, explain their significance for computer vision tasks (e.g., recognition, segmentation).

b) In a pinhole camera system, what happens to the image when the size of the pinhole is too big or too small? Briefly explain why?

Q2. [2 marks] Image segmentation.

a) Describe a simple approach for segmentation based on thresholding. When would it fail?

b) What are **two** drawbacks of the GraphCut segmentation approach?

Q3. [3 marks] Stereo camera system.

a) [0.5 marks] Draw and mark the similar triangles in a stereo camera system.

b) [1 mark] Derive the relationship between the object depth z; the distance between the two cameras (baseline, b); their common focal lengths f, and the location of the object points on the image plan (x_l and x_r). Hint: you may assume that the camera centers are at $(0, 0, 0)$ and $(b, 0, 0)$ respectively.

c) [0.5 marks] For a system with baseline b=5cm, where on the image plane are the epipoles?

d) [1 mark] When does a stereo camera system fail to estimate the object's depth? Provide **two** reasons.

Q4. [3 marks] In a cluttered hostel room with a hundred unique objects, how could you use computer vision to try and find your favourite book? State your assumptions. Please provide a detailed explanation of the process, however, no math is required.

Q5. [3 marks] Convolutional Neural Networks.

a) In a grayscale, normalized image (pixel values in 0-1), you are trying to find sharp discontinuities like shown below. What is the size of the convolutional filter that you should use? What are the best kernel values? Please state whether your filter is already flipped before applying it.

...	...	...	...	...	...	...	...	...
...	0	0	0.8	1	0.8	0	0	...
...	0	0	0.8	1	0.8	0	0	...
...	0	0	0.8	1	0.8	0	0	...
...	...	...	...	...	...	...	...	...

b) What is the difference between translation invariance and translation equivariance? Explain with an example.

c) What technical aspect of Residual Networks (ResNets) allows to train very deep networks that are harder to do with older architectures (e.g., AlexNet, VGG, Inception). Why is this aspect important?

Q6. [3 marks] Neural attention and Transformers.

a) In a sequence-to-sequence machine translation system based on an encoder and decoder Recurrent Neural Network, what is the problem with encoding the source language input into a single 1024-dimensional vector? Would increasing the number of dimensions solve this problem?

b) You are training a vision transformer. At the beginning of training (before the first backward pass), you notice that the attention scores after softmax are spiky (far from uniform distribution). Why is this a problem? Briefly state **two** approaches to fix this issue.

c) When you feed in image patches directly to a Transformer encoder, the model architecture is not sensitive to permutations – shuffling the image patches shuffles the output tokens in the same order. How can this problem be solved?