+1/1/60+

EPFL

1

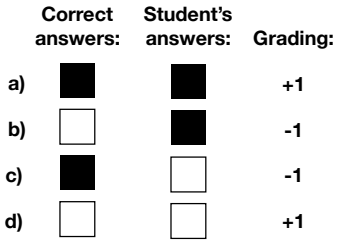
Teacher : Prof. Pascal Fua
CS-442 Computer Vision
27/06/2025
Duration : 90 minutes

Example Name

SCIPER : **123456**

Do not turn the page before the start of the exam. This document is double-sided, has 12 pages, the last ones possibly blank. Do not unstaple.

- Place your student card on your table.
- **A one page two-sided hand-written cheat-sheet** is allowed to be used during the exam.
- Using a **calculator** or any electronic device is not permitted during the exam.
- All questions have one or more correct answers.
- The grading scheme is such that random answering is discouraged:
  - Each answer of a multiple choice question is awarded +1 point if correct and −1 point if incorrect. If the **whole** question is left unanswered no points (positive nor negative) are awarded. Note that "correct" means that a true answer should be ticked and that a false one should be left unticked.



  - The scores for separate questions are **not clipped to 0**, that is, you can get negative score for a question.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question is wrong, the teacher may decide to nullify it.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

choisir une réponse | select an answer
Antwort auswählen

ne PAS choisir une réponse | NOT select an answer
NICHT Antwort auswählen

Corriger une réponse | Correct an answer
Antwort korrigieren

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

Pour votre examen, imprimez de préférence les documents compilés à l'aide de auto-multiple-choice.

First part: Multiple choice questions

For each question, mark the box corresponding to the correct answer. Each question has **at least one** correct answer.

Question 1 Which of the following statements about strategies for improving the performance of deep learning models in computer vision is (are) true?

- [x] Using data augmentation (e.g., random flips, crops, rotations) during training can improve generalization by reducing overfitting.
- [ ] Using a deep and complex network on a small training dataset is a good way to prevent overfitting.
- [ ] Dropout works by randomly removing some training images in each epoch to reduce overfitting in the model.
- [x] Fine-tuning a model pre-trained on a large dataset (like ImageNet) often yields better performance on a smaller target dataset than training a new model from scratch.

Question 2 Which of the following statements about deep learning for semantic segmentation is (are) true?

- [ ] Semantic segmentation models avoid using any pooling or downsampling layers so that the output label map remains at the same resolution as the input image.
- [ ] A convolutional neural network pretrained for image classification cannot be repurposed for segmentation, since classification networks output only a single label per image.
- [x] Deep convolutional networks learn hierarchical feature representations where early layers capture geometry and texture information, and deeper layers capture semantic features.
- [x] U-Net is an encoder–decoder convolutional network with skip connections that merges detailed and high-level features to output high-resolution segments.

Question 3 Which of the following statements about deep neural networks is (are) true?

- [ ] A neural network with no hidden layer (i.e., logistic regression) and a non-linear output activation can represent any arbitrary classification decision boundary.
- [ ] Multi-layer perceptrons are an extension to logistic regression that makes them applicable to linearly separable data.
- [ ] Backpropagation computes weight updates starting from the input layer and moving forward to the output layer.
- [x] Adding one or more hidden layers with non-linear activation functions enables a neural network to model more complex, non-linear decision boundaries.

Question 4 Which of the following statements correctly describe the self-attention mechanism used in Vision Transformers?

- [x] Self-attention computes the relevance between all pairs of input tokens to capture long-range dependencies.
- [x] In scaled dot-product attention, attention weights are computed using the dot product of query and key vectors, scaled by the square root of the key dimension.
- [ ] Self-attention requires fixed-size input and is not permutation-invariant.
- [ ] Self-attention applies a fixed-size kernel over local neighborhoods of tokens.

Question 5 Which statement(s) apply to the Live Wire method?

- ☐ Requires a fully labeled training dataset to function
- ■ Computes local edge costs based on image gradient magnitude
- ■ Uses Dijkstra's algorithm for optimal path search
- ■ Penalizes diagonal moves more heavily

Question 6 You are designing a robot to navigate using ceiling beams in a factory. How might the Hough Transform help?

- ☐ Segmenting beam textures based on color histograms.
- ☐ Classifying the ceiling material based on brightness.
- ■ Estimating vanishing points for path planning.
- ■ Detecting the lines represented by the beams to estimate the movements of the robot.

Question 7 Which of the following is (are) considered fully automated delineation techniques?

- ☐ Deformable Models
- ■ Hough Transform
- ■ Graph-Based Approaches
- ☐ Dynamic Programming

Question 8 What is the main purpose of the Hough Transform in computer vision?

- ■ To detect geometric shapes in images.
- ☐ To perform histogram equalization for image enhancement.
- ☐ To reduce the resolution of an image for fast processing.
- ☐ To compute texture features for classification.

Question 9 Which of the following statements about Sobel filters for edge detection is (are) true, with respect to detecting diagonal edges?

- ☐ A Sobel filter for detecting 45° diagonal edges can be obtained by rotating the standard vertical Sobel filter by 45°.
- ■ A combination of horizontal and vertical Sobel outputs can be used to approximate the response to diagonal edges via magnitude and direction.
- ■ The standard Sobel filters primarily detect horizontal and vertical edges but can be extended to detect diagonal edges by designing new kernels oriented along diagonal directions.
- ☐ Sobel filters can inherently detect diagonal edges without modification because they already include diagonal weights.

Question 10 Why is it important to smooth an image before applying differentiation (e.g., for edge detection)?

- ☐ Smoothing increases the gradient magnitude, making edges more prominent.
- ☐ Differentiation of an image naturally removes noise, so smoothing is not necessary.
- ■ Smoothing ensures that weak but consistent edges are not lost due to noise during differentiation.
- ■ Smoothing suppresses high-frequency noise, which would otherwise be amplified by differentiation.

Question 11 In the Canny edge detector, how do the Gaussian smoothing parameter σ and the choice of high/low thresholds affect the final edge map?

- ☐ Setting both thresholds very low always improves detection of faint edges.
- ■ Decreasing σ increases localization accuracy but may lead to more noise-induced false edges.
- ■ Increasing σ will broaden the Gaussian kernel, which reduces noise but can also blur finer edges.
- ☐ A larger gap between the high and low thresholds makes the hysteresis step more selective, reducing the chance of weak edges being connected to strong ones.

Question 12 Which of the following statements about the Fourier Transform and Discrete Fourier Transform (DFT) of a Gaussian function is (are) true?

- ☐ For a zero-centered Gaussian sampled symmetrically, its DFT produces non-zero real and imaginary components.
- ☐ Fourier Transform of a Gaussian may not always be Gaussian in the frequency domain.
- ☐ The width of the Gaussian in the frequency domain is always equal to its width in the spatial domain.
- ■ The magnitude of the DFT is invariant under spatial shifts of the Gaussian.

Question 13 We are using a model based on the U-Net architecture to segment objects in a room. Which of the following conditions make(s) this task **MORE** challenging?

- ☐ Objects with constant color and no texture
- ■ Varying illumination and shadows
- ■ Occlusions between objects in the scene
- ■ Backgrounds with textures similar to the foreground object

Question 14 We represent a segmentation S as an image composed of zeroes and ones and with the same resolution as an associated image I. Given a video and an initial segmentation S_0 of an object on the first frame I_0, we would like to compute the segmentations $(S_i)_{1\leq i\leq N}$ of that object across all following frames $(I_i)_{1\leq i\leq N}$. Which of the following ideas could be good starting point(s)?

- ☐ Training a Vision Transformer to take I_i as input and output I_{i+1}
- ■ Take the pixel in S_i that minimizes $|I_i - I_{i+1}|$ and apply Region Growing from that pixel
- ■ Compute the minimum/maximum values in I_i of pixels in S_i, and use these values to perform Histogram Splitting in I_{i+1}
- ☐ Training a U-Net to take S_i as input and output S_{i+1}

Question 15 We use shape-from-contour to reconstruct a cup with rich texture. Which of the following statement(s) is(are) true?

- ☐ It provides dense depth maps with high accuracy.
- ■ The normal at the contour are constrained.
- ☐ It assumes a known reflectance model and relies on high-quality texture.
- ■ It requires the camera to be calibrated before reconstruction.

Question 16 How can shape-from-contour complement shape-from-motion?

- ☐ By providing dense texture information.
- ■ By providing boundary constraints where shape-from-motion might fail.
- ☐ By simplifying lighting estimation from shadows.
- ☐ By estimating lighting conditions from the object's appearance.

Question 17 We are performing traditional shape-from-motion (SfM) to reconstruct a building. Which statement(s) is(are) correct?

- [x] Traditional SfM relies on brightness, temporal and spatial consistency of input images.
- [ ] Camera poses are required to run SfM.
- [ ] A single image is enough for SfM.
- [x] A sequence of images from different viewpoints are required for SfM.

Question 18 When implementing shape-from-motion, which of the following statement(s) is(are) true for each step?

- [x] Bundle adjustment can be used to estimate 3D coordinates of feature points.
- [ ] It can only estimate 3D translations of the camera.
- [x] 8 point correspondences between a pair of images can be used to obtain an initial estimation of motion through SVD.
- [x] Feature matching between images impacts the reconstruction quality.

Question 19 In shape-from-stereo, which of the following statements is (are) true about epipolar geometry?

- [ ] Epipoles are always visible in the captured images.
- [x] Rectification transforms the captured images such that all epipolar lines are parallel.
- [x] For a point on one image, its epipolar line in the other image is the line where the corresponding point must lie.
- [x] Epipolar lines may have any orientation.

Question 20 To infer 3D geometry using shape-from-stereo, which of the following statements is (are) true?

- [ ] By searching along epipolar lines to find correspondences, depths can be estimated for all pixels.
- [ ] Disparity is proportional to depth.
- [x] Correlation computation can be very useful to infer the 3D geometry, which is one thing deep neural networks can be used for.
- [x] The uncertainty of 3D geometry estimation depends on the baseline (i.e. distance between cameras) of the camera setup.

Question 21 Which of the following statements regarding technical challenges in shape-from-stereo is (are) true?

- [x] Operating in scale-space is better than using a fixed window size.
- [x] Repetitive patterns, textureless areas, and occlusions can cause problems for correlation computation.
- [ ] Large windows lead to better precision, while small windows yield diminishing precision.
- [x] Dynamic objects in the scene could be reconstructed in shape-from-stereo.

Question 22 Besides Normalized Cross Correlation (NCC), Graph Cut is also a useful approach for solving shape-from-stereo. Which of the following statements is (are) true?

- [x] Graph Cut treats shape-from-stereo as a labeling algorithm.
- [ ] The α-expansion algorithm guarantees the global minimum of the solution in Graph Cut.
- [x] NCC is a normalized measure such that the value does not depend on the mean gray level of the pattern and the image window.
- [x] Graph Cut usually provides less noisy results than NCC.

Question 23 Which of the following is (are) true about shape-from-shading?

- ☐ It cannot be used in conjunction with other shape recovery methods due to having unique boundary conditions.
- ■ It allows for the recovery of shape based on a single image under certain conditions.
- ☐ Specular reflections on the object make recovery easier due to their simple, mirror-like behavior.
- ☐ Given an image of an illuminated Lambertian surface (reflects light equally in all directions), the matching 3D shape that can be recovered from this image is unique.

Question 24 For fun and practice, you decide to apply the reflectance map surface reconstruction you learned in class to reconstruct your face from a picture you wil take. Which of the following would improve the accuracy of your reconstruction?

- ■ Cleaning your face to ensure there are no patches of dirt or discolorations.
- ■ Using a light source far away from your face, such as the sun.
- ■ Incorporating a statistical face shape prior into your reconstruction on top of the smoothness and integrability terms.
- ☐ Applying shiny makeup to make recovery easier.

Question 25 Which of the following is (are) correct about textures?

- ☐ Shape recovery from texture only works when structural textures are present in an image, such as checkerboard patterns.
- ☐ For an image containing multiple types of textures, the DFT (Discrete Fourier Transform) can be used to clearly separate them.
- ■ A texel is usually made up of many pixels.
- ☐ When recovering shape from texture, the first step is to apply the Canny Edge Detector to the image.

Question 26 You are given the following image of a skirt with uniform polka dots, and you will use a shape recovery method to recover its shape. Which of the below methods makes the most sense, along with the stated motivation?

- ☐ Shape-from-stereo, because we see similar patterns repeating in the image. We can split the image into several patches and consider each patch as a view to acquire an accurate reconstruction.
- ■ Shape-from-texture, because we have a clear structure in the polka dots. We can use their scaling and foreshortening to estimate the shape, although the folds will cause difficulties.
- ☐ Shape-from-shading, because the drastic albedo change between black and white will make for an easy reconstruction.
- ☐ Shape-from-contours, because it will allow us to accurately reconstruct the structure behind the folds.

Second part, open questions

Answer in the empty space below each question. Your answer should be carefully justified, and all the steps of your argument should be discussed in details.
Leave the check-boxes empty, they are used for grading.

Shape-from-Stereo

Shape-from-Stereo aims to reconstruct the 3D structure of a scene across images.

Question 27: *This question is worth 4 points.*

☐0 ☐1 ☐2 ☐3 ■4

In the context of shape-from-stereo for 3D reconstruction using two images, describe the key steps involved in matching and reconstructing the corresponding 3D points. (You may consider the following aspects: how to identify corresponding points between the two images, how to compute the camera matrices and the 3D coordinates of matched points, and how to refine the reconstructed 3D points for higher accuracy.)

Answer:
Identify corresponding points between the two images using feature tracking and feature matching (1pt)
Estimate the relative pose (position and orientation) of the second camera with respect to the first by computing the fundamental or essential matrix from the correspondences. (1pt)
Triangulate the matched points using the known camera matrices to recover their 3D coordinates. (1pt)
Refine both camera poses and 3D points using bundle adjustment. (1pt)

Question 28: *This question is worth 4 points.*

☐0 ☐1 ☐2 ☐3 ■4

Explain the principle of triangulation, and write down how to reconstruct the 3D coordinates $\mathbf{X}$ of the corresponding 3D point given the camera intrinsic matrix $\mathbf{K}$, rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$, and the matched feature points $\mathbf{x}_1$ and $\mathbf{x}_2$ in two images.

Answer:
Triangulation utilizes the relationship between the projection positions of a scene point in images taken from different positions by a camera and the camera poses.(1pt)
Through geometric calculations, the coordinates of this point in the 3D space can be determined. (1pt)
Assume $\mathbf{x}_1 = \mathbf{K}[\mathbf{I}|\mathbf{0}]\mathbf{X}$ and $\mathbf{x}_2 = \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{X}$. Convert them into homogeneous coordinate forms $\tilde{\mathbf{x}}_1 = \mathbf{K}\tilde{\mathbf{X}}$ and $\tilde{\mathbf{x}}_2 = \mathbf{K}(\mathbf{R}\tilde{\mathbf{X}} + \mathbf{t})$, where $\tilde{\mathbf{X}}$ is the homogeneous coordinate of $\mathbf{X}$. (1pt)
Solve the linear system of equations $\begin{bmatrix}\lambda_1\tilde{\mathbf{x}}_1 \\ \lambda_2\tilde{\mathbf{x}}_2\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{R} & \mathbf{t}\end{bmatrix}\tilde{\mathbf{X}}$ to obtain $\tilde{\mathbf{X}}$, and then get the 3D point coordinates $\mathbf{X}$. (1pt)

Segmentation

You are applying for an internship at an autonomous driving company. This company works on segmentation in autonomous driving, aiming to identify drivable regions and obstacles. Due to the complexity in real-world scenarios, both hand-crafted and learning-based algorithms could be used. Using insights from the lecture slides and your knowledge, answer the following questions.

Question 29: *This question is worth 3 points.*

☐ 0 ☐ 1 ☐ 2 ■ 3

The Region Growing algorithm defines a metric:

$$\delta(x) = \left|g(x) - \text{mean}_{y\in\mathbf{A}_i(x)}\left[g(y)\right]\right|, \tag{1}$$

where g(x) is the gray value at pixel x.

Explain how this criterion determines which unlabeled pixel x is added to the region $\mathbf{A}_i$ (1pt).

Answer:
The algorithm sorts pixels based on $\delta(x)$, picking the most similar one next to grow the region. (1pt)

In an autonomous driving scenario (e.g., segmenting the road), why might this metric be sensitive to shadows or lane markings (2pts)?

Answer:
Shadows and lane markings may cause large variations in pixel intensities in the same semantic region (road). (1pt)
$\delta(x)$ is purely based on intensity, so it may misclassify these variations as region boundaries, causing fragmentation or incorrect boundaries. (1pt)

Question 30: *This question is worth 3 points.*

0 1 2 3

You move on to K-Means clustering, which minimizes the objective function as

$$\min_{\{\mu_k\}} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \|\mathbf{x}_j^k - \mu_k\|^2 \tag{2}$$

You have deployed this method, using the color (R, G, B) to represent $\mathbf{x}$. However, it fails in certain scenarios, such as incorrectly assigning blue traffic signs and the sky to the same segment.

Please explain why it fails, propose a better solution, and justify why this works better.

Answer:
Reason: color-only clustering can assign similar colors far apart to the same cluster, leading to spatially disjoint segments. (1pt)
Solution: represent $\mathbf{x}$ as a 5D vector (x, y, R, G, B). (1pt)
Explanation: (x, y) penalizes distant pixels from being in the same cluster, promoting spatial continuity. (1pt)
Other reasonable answers are also correct, such as using feature embeddings from pretrained models.

Question 31: *This question is worth 4 points.*

☐0 ☐1 ☐2 ☐3 ■4

Since the hand-crafted algorithms are sensitive to certain challenges, you have decided to investigate deep learning methods. Deep learning methods like U-Net output

$$f_\theta(\mathbf{I}) = \hat{\mathbf{Y}}, \tag{3}$$

where $\hat{\mathbf{Y}}$ is the pixel-wise label map.
Your company has collected a dataset with 100 categories in autonomous driving scenarios. Your task is to train a network on this dataset.

Which loss function can you use? Please write the equation and explain how it works (2pts).

Answer:

$$\mathcal{L}(y, \hat{y}) = -\sum_{c} y_c \log(\hat{y}_c) (1pt) \tag{4}$$

The loss penalizes differences between predicted probabilities $\hat{y}_c$ and true labels y_c. It encourages the model to assign high probability to the correct class for each pixel. (1pt)

Consider the two following cases: (1) rare objects, and (2) unusual weather conditions. Explain whether the trained model would still perform well with your loss function in **EACH** of these scenarios, and justify your reasoning (2pts).

Answer:
The cross-entropy loss relies on balanced training data. For rare objects, the training set has limited examples, which results in poor predictions. (1pt)
Unusual weather changes image appearance, resulting in out-of-distribution data. The predictions become unreliable due to domain shift. (1pt)

Question 32: *This question is worth 3 points.*

☐0 ☐1 ☐2 ■3

During your internship, the major task is to improve the U-Net, making it more robust to the challenges of rare objects and unusual weather. Please come up with three reasonable ideas.

Answer:

- Collect a larger dataset that contains more objects and scenes.
- Use data augmentation to enhance data diversity.
- Use an advanced loss such as focal loss.
- ...