Exercises

1	2	3	4

Surname, First name

Computer Vision 1 (52041COV6Y)
CV1 practice exam 2

1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9
0	0	0	0	0	0	0	0

UNIVERSITEIT VAN AMSTERDAM

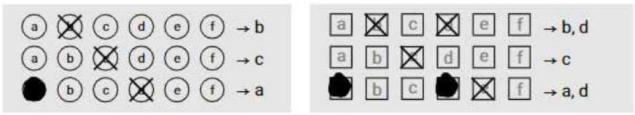
Before you start

1. Do not open this booklet until the supervisor signals the start.
2. Fill out your student number and name on this page. Notify the instructors immediately if you made a mistake.
3. The exam needs to be finished within 3 hours unless an extension is given by the exam committee.
4. You are allowed to use a calculator and one A4 memo sheet. Cell phone and laptop are strictly prohibited unless you have special permission from the exam committee.
5. This exam has in total 45 points.

During your exam

1. Keep a good manner and try not to disturb other students.
2. If you find any typo or ambiguity in the questions, raise your hand and let the supervisors know.
3. If you want to use the toilet, raise your hand.
4. Use the scrap paper for drafting.
5. We recommend a pencil for multiple choice and figure-related questions to allow for corrections.

The following scheme shows how answers are interpreted automatically which can also be used for corrections. If you need to do further corrections please ask for a sticker.



Fill in your answer(s) to the multiple-choice questions as shown above (circles = one correct answer, boxes = multiple correct answers possible).

After the exam

1. Before you submit, double check that your name/student ID are filled out correctly and you did not miss any questions.
2. Please fill out the course evaluation.
3. Submit everything at the front desk and sign the attendance sheet.
4. Leave quietly.

Good Luck!

This page is left blank intentionally

Question 1: Low Level Vision

Camera Model

Figure 1.1:

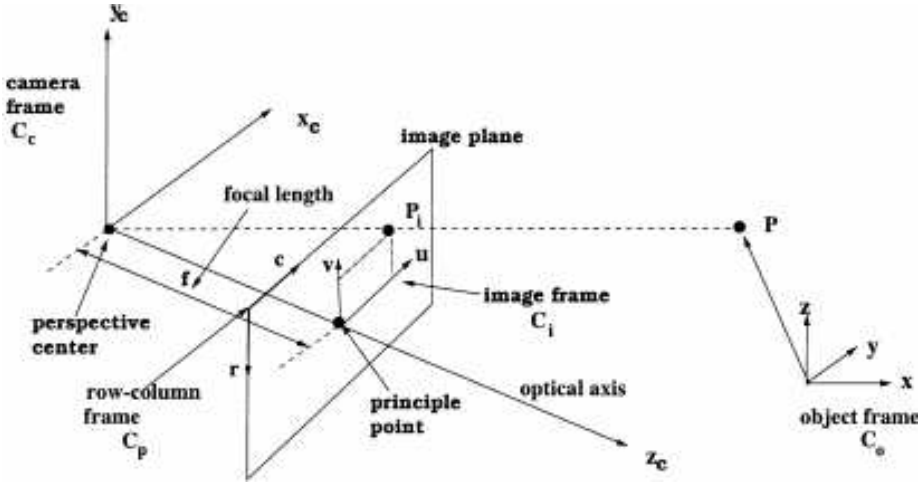


Figure 1.1 illustrates the projection from 3D space to 2D image using a pin-hole camera. The equation can be written as:

$$\mathbf{x} = \mathbf{K}\begin{bmatrix}\mathbf{R} & \mathbf{t}\end{bmatrix}\mathbf{X}$$

where **X** is the coordinates of the 3D point in homogeneous coordinates and **x** is the homogeneous coordinates on the 2D image.

0.5p **1a** What is the name of matrix **K**?

- (a) Intrinsic matrix
- (b) Extrinsic matrix
- (c) Rotation matrix
- (d) Translation matrix
- (e) Projection matrix

1p **1b** Provide the expanded form of **x** and **X** :

2p **1c** The matrix K takes the form of a 3 by 3 matrix. Four elements have been provided. Can you provide the missing 5 elements for a pinhole camera model supporting non-square sensor pixels and a skew parameter:

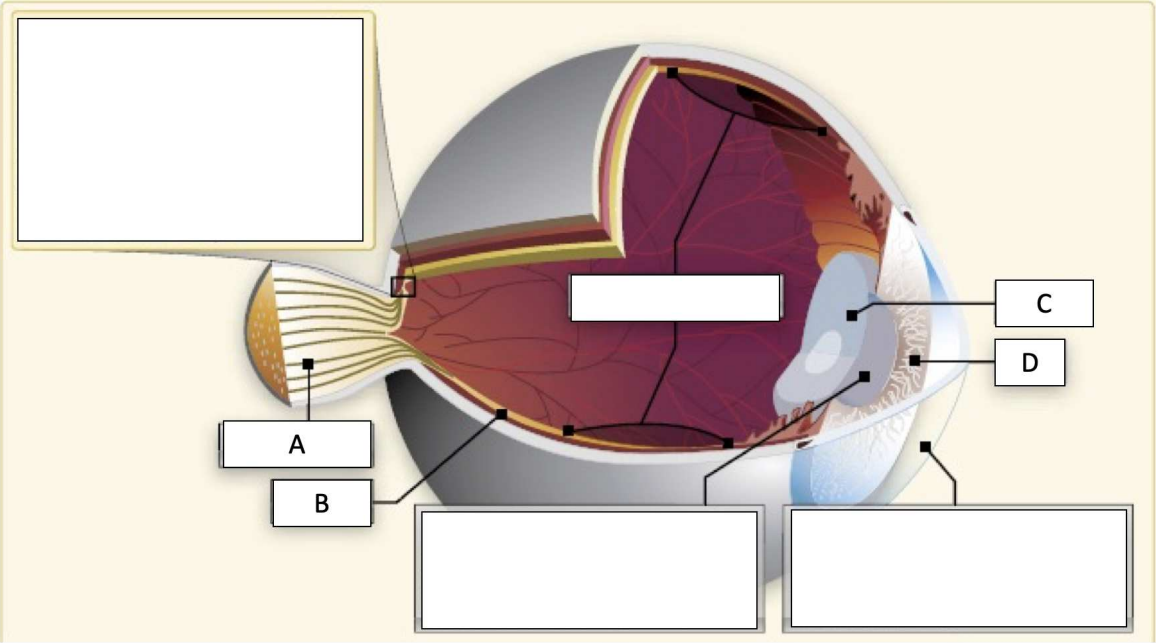
$$K = \begin{bmatrix} \square & \square & \square \\ 0 & \square & \square \\ 0 & 0 & 1 \end{bmatrix}$$

Also, provide the meaning of the 3 elements in the first row in their correct order:

0.5p **1d** Given the current form of K, can we perform an arbitrary rotation using K? If yes explain how, if not explain why?

Human Vision and Color

0.5p **1e** This is an anatomy of human eye.



Which one is the retina?

(a) A

(b) B

(c) C

(d) D

(e) None of them

0.5p **1f** Here are two statements regarding human perception. Which are correct?
A. There are more rods than cones on the human retina.
B. Rods are more sensitive in low light (darkness) than cones.

(a) only A

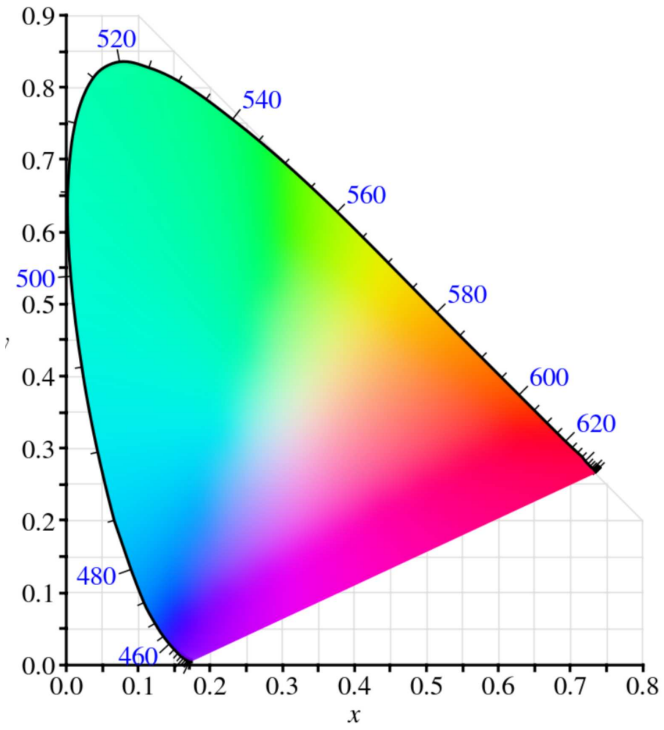
(b) only B

(c) Both of them

(d) None of them

1.5p **1g** CIE systems are commonly used to study color. Assume the sunlight (ideal white) has CIE values $X_S = Y_S = Z_S = 100$. Further, let $X_A = 100$, $Y_A = 300$ and $Z_A = 100$ be the values for a given artificial lamp A. Calculate the chromaticity values *x,y* for both S and A and plot them on the CIE-xy chart in figure 1.2 (0.5pt). (use pencil in case of correction)

Figure 1.2:



Use your plot and calculate the hue (0.5pt) and saturation (0.5pt) of A

Reflection Model

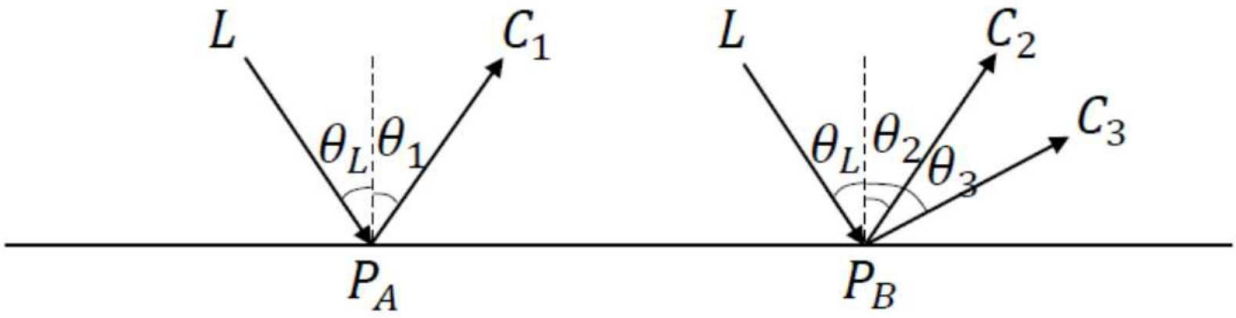


Figure 1.3:

Figure 1.3 shows an image of a flat plane. L is the only light source, which is a uniform and parallel light pointing to the image plane. There is no other light source or object in the setting. There are two points P_A and P_B. And they are observed by three cameras C_1, C_2 and C_3. The angle between light source and surface normal is θ_L, and the angle between camera and surface normal are θ_1, θ_2 and θ_3. The three cameras are identical except there positions.

0.5p **1h** Denote the reflected light intensity to C_1 from point P_A as I_1 provide the simplified *Lambertian* reflection model. Please provide an equation to explain how I_1 can be determined. Give extra symbols and explain them if necessary.

1p **1i** Using the conditions provided in Fig. 1.3 and its descriptions but not the conditions in other questions.

In this question, assume the plane material is *Lambertian* (ideally diffusing).
We know the intensity observed by camera 1 is $I_1 = 300$.

Assume $\theta_1 = \frac{\pi}{6}$, $\theta_2 = \frac{\pi}{6}$ and $\theta_3 = \frac{\pi}{3}$.
Is the information provided so far sufficient to derive the intensity I_2 and I_3 captured by C_2 and C_3.
If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

1p **1j** Using the conditions provided in Fig. 1.3 and its descriptions but not the conditions in other questions.

In this question, assume the plane material is *glossy*, which works like a mirror (ideally diffusing).
We know the intensity observed by the camera 1 is $I_1 = 200$.

Assume $\theta_1 = \frac{\pi}{6}$, $\theta_2 = \frac{\pi}{6}$ and $\theta_3 = \frac{\pi}{3}$.
Is the information provided so far sufficient to derive the intensity I_2 and I_3 captured by C_2 and C_3.
If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

↳

1p **1k** Using the conditions provided in Fig. 1.3 and its descriptions but not the conditions in other questions.
In this question, assume the plane material is *Lambertian* (ideally diffusing) and *uniform*,

We know the intensity observed by the camera 1 is $I_1 = 100$.

Assume $\theta_1 = \frac{\pi}{6}$, $\theta_2 = \frac{\pi}{6}$ and $\theta_3 = \frac{\pi}{3}$.
Is the information provided so far sufficient to derive the intensity I_2 and I_3 captured by C_2 and C_3.
If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

Question 2: Image Processing

In Place Processing and Morphology

Consider the following image patches:

$$P = \begin{array}{|c|c|c|c|}\hline 0&0&0&0\\\hline 0&0&0&0\\\hline 1&1&1&1\\\hline 1&1&1&1\\\hline\end{array} \qquad Q = \begin{array}{|c|c|c|c|}\hline 0&0&1&1\\\hline 1&1&1&1\\\hline 0&0&1&1\\\hline 0&0&0&0\\\hline\end{array}$$

2p **2a** Lets do some binary morphology. Given a construction element

$$U = \begin{array}{|c|c|c|}\hline 0&1&0\\\hline 1&1&1\\\hline 0&1&0\\\hline\end{array}$$

Compute the results after performing *dilation* and *corrosion* separately on image *P* by using template *U*.
All elements at the image boundaries (i.e. outside image *P*) are mirrored. The elements outside filter *U* are all zeros.
P after dilation

P after corrosion

1p **2b** Show the filtering result on *Q* by first do a *corrosion* and then do a *dilation* using *U*. Use zero-padding for both *U* and *Q*.

Image Filtering

2p **2c** We have a sequence

s = [0, 0, -2, 0, 0, 0, 0, 1, 1, 1].

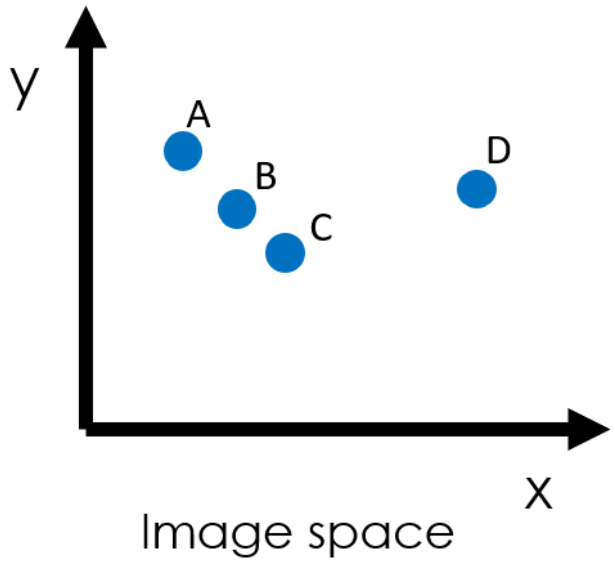
Show the self correlation result by first applying a difference filter h = [1, 0, -1]. Then apply the Gaussian filter g = [1, 2, 1] as defined in the previous question. Use zero padding.

1p **2d** Following the above question, explain what do these operations do? Why do we need to combine a Gaussian filter with a differential filter?

0.5p **2e** Explain the concept of a Bilateral filter:

Edges and Lines

2p **2f** Figure 2.1:



In figure 2.1, we have 4 points in an image. The coordinates are A = (A_x, A_y) = (1, 3), B = (2, 2), C = (3, 1), D = (5, 2).
Show how you can fit a line using Hough transform. For simplicity, use the linear version: y=ax+b.
To get the full points, all steps of Hough transform need to be provided.

1.5p **2g** Following the above question, this time, show how you can find the line using RANSAC.

Corners

2p **2h** Harris corner detection is based on analyzing the second order differentials.
For the patch below:

0	0	1	1
1	1	1	1
0	0	1	1
0	0	0	0

compute its M matrix for the 2 * 2 pixels in the lower right.

$$M = \begin{pmatrix} \sum f_x^2 & \sum f_x f_y \\ \sum f_x f_y & \sum f_y^2 \end{pmatrix}$$

To compute f_x use a simple derivative filter h_x= [-1, 1] in the x-direction and $h_y = [-1, 1]^T$ in the y-direction. The center of h_x is at the first element, idem for h_y. Use cross-correlation for simplicity. Handle the out-of-boundary pixels with mirroring. To save time, assume the window size for summation over the neighborhood Σ is 1x1, i.e. you can ignore the summation.

1p **2i** Name two **photometric** image transformations that the SIFT descriptor is invariant to.

Optical Flow

1.5p **2j** Which statements about the Lucas-Kanade optical flow method are correct?

- ☐ The method is not robust to handle large motions (larger than the window size).
- ☐ The color or brightness of pixels is assumed to remain unchanged during motion.
- ☐ The method favors neighboring pixels to have the same flow vectors.
- ☐ The structure tensor is used to enforce local smoothness of the estimated flow field.
- ☐ The structure tensor is required to contain gradient information in at least one direction to uniquely estimate motion.
- ☐ The method fails to estimate a flow vector for homogeneously colored image regions.

Linear Transformations

1p **2k** Different types of transformations have a different degree of freedom. What is the degree of freedom of a 3D rigid body transform $T \in SE(3)$?

(a) 2 (b) 3 (c) 4 (d) 6

1p **2l** What is the degree of freedom of a 2D rotation $R \in SO(2)$?

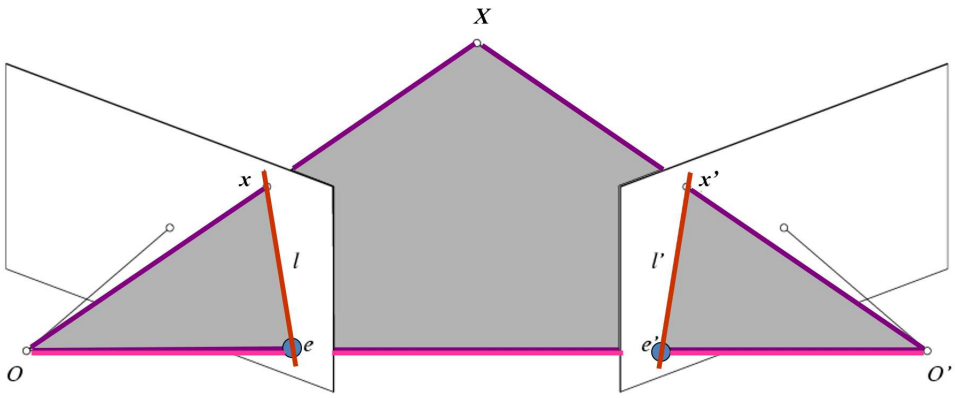
(a) 1 (b) 2 (c) 3 (d) 6

1p **2m Transformation composition:** Consider a kinematic chain of a human skeletal model where joint locations are described relative to each other. In particular, the joint for the hand is described by point $H = (h_x, h_y, h_z) \in \mathfrak{R}^3$ in local coordinates relative to the elbow $E = (e_x, e_y, e_z) \in \mathfrak{R}^3$, while the elbow is defined relative to the shoulder $S = (s_x, s_y, s_z) \in \mathfrak{R}^3$, which in turn is relative to the pelvis $P = (p_x, p_y, p_z) \in \mathfrak{R}^3$ that acts as the main anchor point for the entire skeleton.
All body joints also have locally defined relative rotations $R_J \in SO(3)$ associated with them, where $J \in \{H, E, S, P\}$ is a place holder for any body joint. Therefore, coordinate transformations from a local joint coordinate frame into the corresponding parent coordinate are described by a rigid body transformation $T_J = [R_J | J] \in SE(3)$.
Select all correct statements from the following options (multiple correct answers possible):

- ☐ To obtain the position of the hand in the global coordinate frame one only reads out point H.
- ☐ To obtain the position of the hand in the global coordinate frame one has to compute $T_P T_S T_E T_H \vec{0}$
- ☐ To obtain the position of the hand in the global coordinate frame one has to compute $T_H T_E T_S T_P \vec{0}$
- ☐ To move the entire skeleton by a given displacement vector $D \in \mathfrak{R}^3$ one has to update the relative translation only for joint P.
- ☐ To move the entire skeleton by a given displacement vector $D \in \mathfrak{R}^3$ one has to update the relative translation for all joints J.

Multiview Geometry and Reconstruction

2p **2n** The **epipolar geometry** is a key concept in multi-view stereo.



The figure above shows the camera centers O and O' and a 3D point X which projections on the image planes are x and x'.

Please mark correct all correct statements about epipolar geometry (multiple correct answers are possible).

- ☐ The epipolar plane always contains the epipolar lines.
- ☐ The line through the points x and e is called epipolar line.
- ☐ The epipoles are always part of the the epipolar lines.
- ☐ The epipolar lines are always orthogonal to the line defined by the points o and o'.
- ☐ The line through the camera center o and the point X defines the principal axis.

1p **2o Rectified Stereo:**

- ☐ Image rectification reduces the correspondence search between corresponding pixels in the two input images from a 2D to a 1D search problem.
- ☐ In a rectified stereo setting all epipolar lines are parallel.
- ☐ The principal axes of two cameras are orthogonal after stereo rectification.
- ☐ Stereo rectification can be achieved by transforming both images with a rigid-body transform (i.e. SE(3)).

2p **2p** The following groups are generalizations of each other. Please order them from the most to the least general by entering numbers from 1 to 4.

( ) General linear group

( ) Special orthogonal group

( ) Orthogonal group

( ) Set of square matrices

0.5p **2q** **Intrinsic camera calibration.** Consider the following camera model:

$$w \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} x \\ y \\ y \\ 1 \end{bmatrix}$$

How many parameters (DOF) need to be estimated for intrinsic camera calibration for the shown camera model?

(a) 3 (b) 4 (c) 6 (d) 9 (e) 12

0.5p **2r** **Extrinsic camera calibration.** Using the same camera model as in the previous question: How many parameters (DOF) need to be estimated for extrinsic camera calibration?

(a) 5 (b) 6 (c) 9 (d) 12

Question 3: Image Understanding

Traditional Classification and Retrieval

1p **3a** Consider a binary classifier with the following classification results.

		Predicted class labels	
		Positive	Negative
Actual class labels	Positive	5600	40
	Negative	1900	2460

Please compute the precision and recall values for this classifier using the provided confusion matrix.

1p **3b** Mark all correct statements.

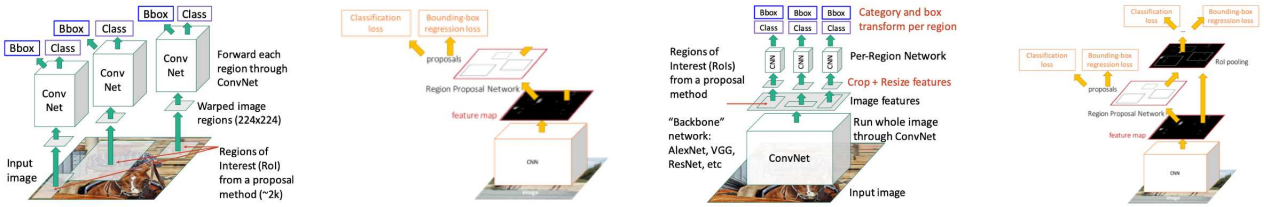
- ☐ The maximum IoU score is 1.
- ☐ F1 score is not normalized and requires additional normalization to ensure a value range of [0,1].
- ☐ A bag of words representation does not account for the frequency a word occurs.
- ☐ F1=$2 \times \frac{Precision+Recall}{Precision \times Recall}$
- ☐ The bag of visual words representation describes an ordered set of visual words.

Object Detection

1p **3c** Mark all correct statements about object detection methods and their individual steps.

- [ ] For given object proposals, Fast R-CNN predicts absolute bounding box coordinates.
- [ ] All R-CNN methods require object proposals as input.
- [ ] The major difference between Fast R-CNN and (slow) R-CNN are learned vs. non-learned region proposals.
- [ ] Both Fast R-CNN and Faster R-CNN have a global feature extraction stage which is jointly computed and then used for all region proposals.

1p **3d Object detection architectures.**



A **B** **C** **D**

Match the correct name with the architectures **A, B, C, D** depicted above:

( ) R-CNN

( ) Fast R-CNN

( ) Faster R-CNN

( ) Single Stage Detector

Neural Networks

1p **3e** Select all correct statements about neural network architectures.

- ☐ Convolutional layers have fewer parameters than fully connected layers if the kernel size is smaller than the image size.
- ☐ Fully connected layers are naturally invariant to object translations in the image.
- ☐ Convolutional layers are naturally invariant to object rotations in the image.
- ☐ For patch sizes smaller than the image, convolutional layers require less computing operations and occupy less memory.

2p **3f** Consider a 2D convolutional layer with RGB-D input of size 64 x 64. We apply 3 convolutional filters of size 4 x 4. All input channels are convolved together, not separetely, no padding, and stride = 2.
1. What are the dimensions of the output activation layer ?
2. How many weight parameters have to be trained in this layer ?

2p **3g** On the result of the convolutional layer of the previous question we apply a 3 x 3 max pooling layer using stride = 1.
1. What are the dimensions of the output layer ?
2. How many weight parameters have to be learned in this layer ?

1p **3h** Mark all correct statements about 2-stage and single shot object detectors.

- ☐ 2-stage detectors require substantially less object proposals than single shot detectors.
- ☐ Non-maximum suppression is only required for 2-stage detectors.
- ☐ Single shot detectors are typically faster than 2-stage detectors.
- ☐ Single shot detectors require much less training data than 2-stage detectors.
- ☐ Single shot detectors use a small set of fixed regions as object proposals.

Guest Lectures

Prof. dr. Javier Civera

0.5p **4a** Which topic was part of the lecture?

- (a) Vision Language Models
- (b) Medical Image Segmentation
- (c) Methods for Image Retrieval
- (d) Network pretraining for Computer Vision Models
- (e) Style Transfer Methods
- (f) Generative Classification Methods
- (g) Morphological Segmentation Methods

Dr. Jose Alvarez and Ms. Maying Shen

0.5p **4b** What was the theme/topic of ms. Ma Shen's part of the lecture

- (a) Generative AI
- (b) Network pruning/compression and module distillation
- (c) NVidia Cuda API and AI chipset
- (d) Latent Diffusion Model
- (e) Training Large Language Model

Dr. Sezer Karaogru

0.5p **4c** What was the theme/topic of Dr. Karaogru's lecture

(a) ChatGPT and Visual Language Model

(b) Robust Vision in Bad Weather

(c) Autonomous Driving and Modern Robotics

(d) Media synthesis and identify synthetic media

Extra empty page

This page is left blank intentionally