Question 1)

In a Structure-from-Motion problem with m=3 camera views and n=20 3D points, how many equations are available if each point is visible in all views?

Antwort:

Question 2)

Given f/8 and an aperture diameter of 16mm, what is the focal length of a lens in mm? Type in only number (no "mm").

Antwort:

Question 3)

Select all statements that are correct wrt. **capturing digital images**.

- a. Color values in an RGB image measured through a Bayer filter are typically obtained by interpolation from neighboring subpixels.
- b. A higher signal-to-noise ratio generally results in more reliable intensity measurements.
- c. CMOS and CCD sensors measure the number of photons arriving at each pixel location over exposure time.
- d. Sensor noise has a stronger relative impact in dark image regions than in bright regions.
- e. In a Bayer filter pattern, each pixel directly measures all three color channels (R, G, and B).

Question 4)

Which statement **best describes the main difference** between the **Lucas-Kanade** and **Horn-Schunck** optical flow methods?

- a. Lucas–Kanade estimates optical flow by solving a local least-squares problem, while Horn–Schunck estimates optical flow by minimizing a global energy function with a smoothness constraint.
- b. Horn–Schunck works only for small motions, while Lucas–Kanade handles large displacements without modification.
- c. Lucas–Kanade produces dense optical flow, while Horn–Schunck produces sparse optical flow by default.
- d. Lucas–Kanade enforces global smoothness, while Horn–Schunck estimates motion independently per pixel.

Question 5)

Which of the following statements about Large Multi-Modal Models (LMMs) are **correct**?

- a. LMMs can be used for high-level vision tasks (e.g. reasoning) but may struggle with precise low-level tasks such precise spatial estimations.
- b. LMMs suffer from poor generalization to novel visual concepts due to overfitting on language data.
- c. LLMs suffer from hallucination of visual content not present in images.
- d. LMMs can only perform tasks for which all input modalities are present at inference time.

Question 6)

In graph-based segmentation, an affinity matrix A of size

N×N is constructed, where

$a_{i,j}$

represents the similarity between pixels i and j.

Suppose you use the *clustering by graph eigenvectors* method and compute the eigenvectors of A in each clustering round (iteration).

Given:

- N=1000 pixels
- The largest eigenvalue after the first round is $\lambda_1 = 85.3$
- The largest eigenvalue after the second round is $\lambda_2 = 42.7$
- The largest eigenvalue after the third round $\lambda_3 = 18.4$

How many clusters will you obtain if you stop when the eigenvalue drops below **20.0**?

Antwort:

Question 7)

Which statement about **convolution in image processing** is **correct**?

- a. Convolution is a non-linear, shift-variant operation.
- b. Convolution requires the kernel to be the same size as the image.
- c. Convolution amplifies high-frequency components only.
- d. Convolution in the spatial domain corresponds to multiplication in the frequency domain.

Question 8)

What is the primary purpose of using a scale-invariant loss function for **monocular depth prediction**?

- a. To improve the accuracy of surface normal prediction.
- b. To eliminate the need for ground truth depth data.
- c. To reduce computation time during training.
- d. To account for the fundamental scale-depth ambiguity in single-image depth estimation.

Question 9)

Which of the following are valid reasons for performing image rectification in stereo vision?

- [ ] a. To make epipolar lines horizontal.
- [ ] b. To simplify correspondence search to 1D and in one direction.
- [ ] c. To avoid projection.
- [ ] d. To estimate camera intrinsic parameters.

Question 10)

Event cameras differ fundamentally from conventional frame-based cameras. Which of the following statements about event cameras are correct?

- [ ] a. Event cameras inherently suffer from motion blur when observing fast-moving objects.
- [ ] b. Event cameras generally offer higher temporal resolution and lower latency than frame-based cameras.
- [ ] c. Event cameras asynchronously report changes in pixel intensity rather than capturing full image frames at fixed intervals.
- [ ] d. Event cameras require significantly more bandwidth than conventional cameras because they continuously stream all pixel values.
- [ ] e. Each event typically encodes the pixel location, timestamp, and the brightness change.

Question 11)

In a Convolutional Neural Network (CNN), the **Rectified Linear Unit (ReLU)** activation function
$f(x) = \max(0, x)$
is primarily used to:

- [ ] a. Introduce non-linearity and help mitigate the vanishing gradient problem.
- [ ] b. Normalize pixel values to a range between 0 and 1.
- [ ] c. Reduce the spatial dimensions of feature maps.
- [ ] d. Compute gradient orientations for feature descriptors.

Question 12)

What does a **Neural Radiance Field (NeRF)** model output for a given 3D point and viewing direction?

- ( ) a. Vertex coordinates and texture.
- ( ) b. Depth and surface normal.
- ( ) c. Occupancy probability and gradient.
- ( ) d. Volume density and RGB color.

Question 13)

A detector produces the following results for a class:

- True Positives (TP): 80
- False Positives (FP): 20
- False Negatives (FN): 40

What is the precision?

Antwort:

Question 14)

In a U-Net architecture for semantic segmentation, what is/are the primary purpose(s) of **skip connections** between corresponding encoder and decoder layers?

- a. Mid-level representations are often formed by clustering low-level features before classification.
- b. To preserve high-resolution spatial information lost during down-sampling
- c. To reduce the number of trainable parameters
- d. To introduce non-linearity into the decoder
- e. To allow the network to bypass the bottleneck layer entirely

Question 15)

In a traditional **model-based classification pipeline** (e.g., using handcrafted features like SIFT), which of the following statements are true?

- a. It typically involves steps such as feature detection, descriptor extraction, and clustering.
- b. Features detectors are manually designed and remain fixed (e.g., edge, corner, and blob detectors).
- c. Mid-level representations are often formed by clustering low-level features before classification.
- d. The feature extraction kernels are learned automatically from labeled training data.
- e. It requires no preprocessing or invariance handling (e.g., scale or rotation invariance).

Question 16)

Which of the following statements about **deep learning–based optical flow estimation** are correct?

- a. Encoder–decoder architectures (e.g., U-Net–like networks) are commonly used to produce dense optical flow fields.
- b. Convolutional neural networks for optical flow inference typically require explicit computation of image gradients.
- c. Deep learning methods estimate optical flow by learning a direct mapping from pairs of images to flow fields.

Question 17)

For **Eigenfaces and Face Recognition**, which of the following statements are incorrect?

- a. The number of Eigenfaces (K) is chosen based on the **largest eigenvalues**, which correspond to the directions of maximum variance.
- b. Face images are projected onto the Eigenface space by computing the dot product between the mean-subtracted image and each Eigenface.
- c. Eigenfaces are the **eigenvectors** of the covariance matrix computed from a set of face images.
- d. A new face image is represented as a weighted linear combination of the mean face and the Eigenfaces.
- e. Eigenface-based recognition is highly robust to extreme variations in lighting, pose, and facial expression between training and test images.

Question 18)

A convolutional layer is designed with an input volume of 32 × 32 × 3 and uses 16 filters, each of size 3 × 3 × 3. Assuming no bias terms are used, how many **trainable parameters** are in this convolutional layer?

Antwort:

Question 19)

An image patch is given by

$$I = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

and the convolution kernel is

$$h = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

What is the **output value at the center pixel** after convolution (assume valid convolution, no padding)?

Antwort:

Question 20)

What is the primary function of the Region Proposal Network (RPN) in Faster R-CNN?

- a. To compute the loss for objectness.
- b. To perform bounding box regression.
- c. To propose regions that likely contain objects.
- d. To classify objects into specific categories.

Answer Question 1)

Die richtige Antwort ist: 120

Answer Question 2)

Die richtige Antwort ist: 128

Answer Question 3)

Die richtigen Antworten sind: CMOS and CCD sensors measure the number of photons arriving at each pixel location over exposure time., Sensor noise has a stronger relative impact in dark image regions than in bright regions., A higher signal-to-noise ratio generally results in more reliable intensity measurements., Color values in an RGB image measured through a Bayer filter are typically obtained by interpolation from neighboring subpixels.

Answer Question 4)

Die Antwort ist richtig.

Die richtige Antwort ist: Lucas–Kanade estimates optical flow by solving a local least-squares problem, while Horn–Schunck estimates optical flow by minimizing a global energy function with a smoothness constraint.

Answer Question 5)

Die Antwort ist richtig.

Die richtigen Antworten sind: LMMs can be used for high-level vision tasks (e.g., reasoning) but may struggle with precise low-level tasks such precise spatial estimations., LLMs suffer from hallucination of visual content not present in images.

Answer Question 6)

Die richtige Antwort ist: 2

Answer Question 7)

Die Antwort ist richtig.

Die richtige Antwort ist: Convolution in the spatial domain corresponds to multiplication in the frequency domain.

Answer Question 8)

Die Antwort ist richtig.

Die richtige Antwort ist: To account for the fundamental scale-depth ambiguity in single-image depth estimation.

Answer Question 9)

Die richtigen Antworten sind: To make epipolar lines horizontal., To simplify correspondence search to 1D and in one direction.

Answer Question 10)

Die richtigen Antworten sind: Event cameras asynchronously report changes in pixel intensity rather than capturing full image frames at fixed intervals., Each event typically encodes the pixel location, timestamp, and the brightness change., Event cameras generally offer higher temporal resolution and lower latency than frame-based cameras.

Answer Question 11)

Die Antwort ist richtig.

Die richtige Antwort ist: Introduce non-linearity and help mitigate the vanishing gradient problem.

Answer Question 12)

Die richtige Antwort ist: Volume density and RGB color.

Answer Question 13)

Die richtige Antwort ist: 0,8

Answer Question 14)

Die Antwort ist richtig.

Die richtige Antwort ist: To preserve high-resolution spatial information lost during down-sampling

Answer Question 15)

Die richtigen Antworten sind: Features detectors are manually designed and remain fixed (e.g., edge, corner, and blob detectors)., It typically involves steps such as feature detection, descriptor extraction, and clustering., Mid-level representations are often formed by clustering low-level features before classification.

Answer Question 16)

Die richtigen Antworten sind: Deep learning methods estimate optical flow by learning a direct mapping from pairs of images to flow fields., Encoder–decoder architectures (e.g., U-Net–like networks) are commonly used to produce dense optical flow fields.

Answer Question 17)

Die richtige Antwort ist: Eigenface-based recognition is highly robust to extreme variations in lighting, pose, and facial expression between training and test images.

Answer Question 18)

Die richtige Antwort ist: 432

Answer Question 19)

Die richtige Antwort ist: 0

Answer Question 20)

Die richtige Antwort ist: To propose regions that likely contain objects.