Chair of Computer Vision and Artificial Intelligence
Department of Informatics
Technical University of Munich

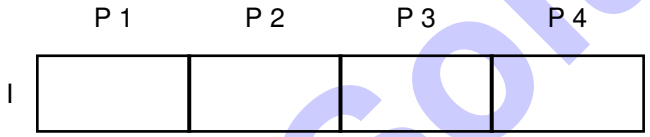
Esolution
Place student sticker here

Note:

- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

Computer Vision III: Detection, Segmentation and Tracking

Exam:	IN2375 / Endterm	**Date:**	Thursday 11th August, 2022
Examiner:	Prof. Dr. Ing. Laura Leal-Taixe	**Time:**	10:45 – 12:15



Working instructions

- This exam consists of **12 pages** with a total of **4 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 60 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
  - one **non-programmable pocket calculator**
  - one **analog dictionary** English ↔ native language
- Subproblems marked by * can be solved without results of previous subproblems.
- **Answers are only accepted if the solution approach is documented.** Give a reason for each answer unless explicitly stated otherwise in the respective subproblem.
- Do not write with red or green colors nor use pencils.
- Physically turn off all electronic devices, put them into your bag and close the bag.

Left room from _____ to _____ / Early submission at _____

Problem 1 Multiple Choice (12 credits)

Mark your answer clearly by a cross in the corresponding box. Multiple correct answers per question possible. For every question, you will either get full credit (if you mark all the correct answers, and not mark all the incorrect answers) or no credit otherwise.

Mark correct answers with a cross ☒
To undo a cross, completely fill out the answer option ■
To re-mark an option, use a human-readable marking ×■

a) Which of the followings are correct for Non-Maximum Suppression (NMS) (check all that apply)

- [x] NMS measures region overlap with the Intersection over Union (IoU) or Jaccard Index.
- [ ] NMS returns a single bounding box per image by eliminating all boxes except the best one.
- [x] Choosing a wider threshold may lead to low precision.
- [ ] Choosing a wider threshold may lead to low recall.

b) Check all that apply for Hourglass (Unet) architecture.

- [x] Hourglass (Unet) architecture is composed of an encoder and a decoder.
- [ ] Encoders are responsible for upsampling, while decoders downsample.
- [x] Residual connections help preserve knowledge at each resolution.
- [ ] Hourglass (Unet) architecture is not well-suited for pixel-precise tasks like segmentation.

c) Check all that apply for video object segmentation.

- [x] Video object segmentation is a task equivalent to multi-object tracking, and the only difference is in object representation (segmentation masks instead of bounding boxes).
- [x] ON-AVOS uses each frame for fine-tuning, relying on past-frame network predictions as a supervisory signal. To this end, it uses all positive predictions from the past frame as positives and all negative predictions as negatives.
- [ ] A significant advantage of ShapeMask is that it can be trained from stationary images.
- [ ] The main idea behind PReM-VOS is to stack several images before feature extraction to learn rich patio-temporal context during backbone feature extraction.

d) Check all that apply for Faster R-CNN object detector.

- [x] In Faster R-CNN, Region Proposal Network does not contain trainable weights; this first stage is manually-designed, not trainable. The second stage (regression, classification heads) is mostly trainable and therefore does contain trainable parameters.
- [ ] Vanilla Faster R-CNN contains two classifiers at different stages of the network.
- [x] Anchor boxes are of various sizes. To cope with that, the second stage (regression and classification heads) employs a Fully Convolutional Network.
- [ ] Faster R-CNN is only compatible with a VGG-based backbone.

e) Check all that apply for 3D multi-object tracking from 3D sensory data.

- [x] Methods presented in the lecture primarily rely on appearance models for the association. However, in contrast to image-based tracking, appearance models are learned from geometric data.
- [x] Due to precise range maps provided by the lidar sensor, localization of objects (as bounding boxes) in 3D space is less channeling compared to the localization of objects (as bounding boxes) in the image domain.
- [x] Kalman filter can be used to learn to predict motion vector from the training data.
- [x] In contrast to image-based tracking, several consecutive point clouds are often stacked before feature extraction for each frame.

f) Which of the following statements is/are true.

- [x] A fully convolutional network can deal with any input size.
- [x] 1x1 convolutions can be used to shrink the number of channels.
- [x] 1x1 convolutions can be used to increase the number of channels.
- [x] 1x1 convolutions further adds a non-linearity, hence the model can learn more complex functions.

Problem 2 Learning from 3D data (14 credits)

In the lecture, we have mentioned several different backbones for learning representations from 3D point cloud data. First, we discussed backbones operating on two different representations of the input data (i.e., point clouds).

0 1 2 3

a) Describe two different types of representations on which these backbones operate, and discuss their pros and cons (two pros, two cons for each) (3p: 1p for specifying, 1+1 for pros/cons).

- : Point sets (1), voxel grids (1).
- Point sets. Pros: no preprocessing needed, no discretization artifacts, directly operate on sensory data. Cons: Less consolidated than regular-structured data, more memory demanding than (sparse!) voxel grids.
- Voxelgrids: Pros: conceptually similar to image data (just extra dim), therefore: well-consolidated, state-of-the-art performance (sparse version). Cons: preprocessing/discretization needed, point refinement needed (obtain only per-voxel classification by default)

0 1 2 3

b) Describe the key idea behind PointNet. Your answer needs to specify key bounding blocks of the PointNet network, as used for point cloud classification (i.e., assign a single semantic label to the input point cloud) (3p: 1p for the key idea, 2p for key components).

- Key idea: learn a (permutation-invariant) representation of a point set by encoding individual points with a shared MLP, followed by max-pooling to obtain fixed-dim point cloud representation (1p).
- Key components: T-Net, per-point shared MLP, max pooling, classification FC layer (2p, 0.5 each).

0 1 2

c) Describe how PointNet can be extended to per-pixel classification (semantic segmentation). (2p).

- Retain learned per-point embeddings (learned via per-point MLP), concatenate per-point the global point cloud feature vector, obtained by max-pooling (1p).
- This is followed by dimensionality reduction (1-d conv) to obtain per-point feature vector, followed by a per-point classification MLP (1p).

d) Briefly describe the key idea behind sparse Convolutional Neural Networks. Describe two types of convolutions we have mentioned in the lecture and describe the key difference between the two. (3p: one for the key idea, 1 for two types of convolutions, 1 for the key difference).

0 1 2 3

- Voxegrid-based representation with convolutional kernels. Only process active, "occupied" cells (1p).
- Sparse convolutions and submanifold convolutions (1p).
- Key difference: sub-manifold convolutions only process sites on which kernels are centered (sparse convos: process whenever kernel touches an active site). Also accept (alternatively): submanifold convolutions, by contrast to standard sparse, do not cause submanifold dilation ("spread" of active sites in deeper layers) (1p).

e) To learn representations from image sequences (video as an image sequence, not 3D point clouds), we could simply stack a sequence of images. This would yield a $W \times H \times 3 \times N$ input tensor, where N denotes the sequence length. Are sparse Convolutional Neural Networks the right backbone for processing such data? Justify your answer. (3p for a well-justified answer)

0 1 2 3

Open-ended question. Aljosa's take: NO, those tensors are not sparse. For each well-justified (and correct) answer, give 3 points. Zero points for answer-only, even if correct. Answer YES can be accepted if the point of view is very well justified.

Problem 3 Transformers (11 credits)

Please answer the following questions on Transformers.

0 1 2

a) In Transformers, why do we mask the input of the decoder, before feeding it to the decoder? (2p).

Answer: In Transformers, the training is done in parallel (unlike in RNNs). Masking is used so that $output_i$ cannot see $outputs_j$, where $i < j$. In other words, masking prevents the model from "peeking into the future output". (2p).

0 1 2 3

b) Can Transformers accept inputs of arbitrary sizes? For example, expect a sentence of size i in the first batch, one of size j in the second batch (where $i \neq j$), and so on (assume we are not allowed to use padding). Justify why this can or cannot be done. (3p: 1 for correct answer, 2 for justification).

Answer: Yes they can handle arbitrary input sizes in different batches. (1P) The computation of the attention matrix only requires the query and key matrices to have the same embedding dimension and the output vectors of the multi-head attention need to be of the same dimension as the input. (2P) Since $Q \in R^{qxd}$, $K \in R^{kxd}$, $V \in R^{kxd}$:

$$att = Q * K^T \rightarrow att \in R^{qxk} \tag{3.1}$$

$$final = att * V \rightarrow final \in R^{qxd} \tag{3.2}$$

the only variable that has to be constant is d.

0 1 2 3

c) How do Visual Transformers (ViT, "An image is worth 16x16 words") divide (1p) and encode (1p) the input image? What are they used for? (1p).

- Answer: ViT divides the input into rectangles (1p).
- Then, they encode rectangular patches using MLP and add positional encoding (1p). Afterward, these units/tokens are fed to the Transformers encoder (optional mention: rectangles are of size 16x16).
- They are used for the task of image classification (this is what we talked about in the lecture). They can also be used as backbones for object detection/segmentation etc. (works that use them for such tasks do exist). (1p)

d) What does masking mean in the context of Masked Auto-Encoders, and how do masked auto-encoders work? What are they used for? (3p, one for each correct answer).

0 1 2 3

- Masking: MAE masks parts of the input such that only a part of the input signal is fed to the encoder network. (1p)
- The idea behind MAEs is to reconstruct the full signal (e.g., image, sentence) from the non-masked) part of the signal that was fed to the encoder. For the network to learn to complete the signal, it needs to learn a good representation that should ideally be transferable to different tasks (1p).
- We use them for self-supervised learning, where the supervisory signal comes from hiding a part of the signal. Therefore, we can use them to learn representations from unlabelled data on large datasets and then fine-tune this representation for downstream tasks (i.e., classification, segmentation) (1p).

Problem 4 Various Topics (23 credits)

0 1 2

a) One of the traditional approaches to object detection is based on template matching and sliding windows. These methods evaluate how much the pixels in the image and template correlate for every position. Name two disadvantages of such an approach. (2p: 1p each)

Any two, 1p each:

- Occlusions: we need to see the whole object
- This works to detect a given instance of an object but not a class of objects
- Sliding window search space is large: Objects have an unknown position, scale and aspect ratio

0 1 2 3

b) How does YOLO (Redmon et al, “You only look once: Unified real-time object detection”, CVPR 2016) produce initial guesses to directly regress from image to box coordinates? (1p) How does YOLOv2 (Redmon et al, "YOLO9000", CVPR 2017) improve this scheme at the initial prediction level? (1p) List two advantages of YOLO (1p: 0.5 each)

Divide the image in a grid, place a box at the center of each cell in the grid, and this will be our initial box guess for that object. Then predict. (1p) Anchor boxes (1p) Any two (0.5 p each):

- Very fast
- End-to-end trainable
- Fully convolutional

0 1 2 3

c) Anchor boxes are commonly used in 2D object detection; however, there is an alternative line of work that tries to eliminate them. Name two one-stage approaches for 2D object detection that are either point-based or transformer-based. Briefly explain their main idea. (1.5p each)

Any two (1.5p each: name (0.5p), main idea (1 p)):

- CornerNet: represent bounding boxes with top left and bottom right corners
- CenterNet: Use the corners as proposals and focus on the center of the object to infer its class
- ExtremeNet: represent objects by their extreme points
- DETR: CNN backbone, transformer for decisions. Bipartite matching based loss for set prediction.

d) What is the problem with one-stage detectors about foreground and background objects? (1p) Why is it easier for two-stage detectors to deal with this problem? (1p) What does RetinaNet (Lin et al, ICCV 2017) propose to solve this? (1p)

0 1 2 3

Foreground-background imbalance: Many locations need to be analyzed densely covering the image (1p) Region proposals: Classification only work on “interesting” foreground regions (1p) Focal loss (1p)

e) What is the difference between online tracking and offline tracking? (1p) Suggest an application in which an online tracker is favorable over an offline tracker. (0.5 p) Suggest an application in which an offline tracker is favorable over an online tracker. (0.5 p)

0 1 2

Online tracking: you have access to past and current frames. Offline tracking: you can peak into future. (1 p) Real-time tasks (0.5 p) video analysis, or semi-automated video labelling (0.5 p)

f) Write down the complete equation of the contrastive loss (Chopra et al, "Learning a similarity metric discriminatively, with application to Face verification", CVPR 2005) (1.5 p) Highlight the components belonging to the positive and negative pair (0.5 p) and explain the purpose of each component (1p)

0 1 2 3

Fully correct loss (1.5 p), identify components (0.5 p). Correct explanation is 0.5 p each; however, "margin" of the negative pair should be explained for full credit.

$$\mathcal{L}(A,B) = y^* ||f(A) - f(B)||^2 + (1-y^*) max(0, m^2 - ||f(A) - f(B)||^2)$$

0 1 2

g) Consider graph representation of multi-object tracking based on network flows. What does a node (0.5 p), an edge (0.5 p), a source (0.5 p) and a sink (0.5 p) represent?

Node = detection, edge = trajectory (hypothesis), source = special node with connections that starts a trajectory, sink = special node with connections that end a trajectory

0 1 2

h) Message passing networks (MPNs) propagate information across the graph for several iterations. What are the two steps of the propagation process? (1p) During message passing, messages are often aggregated with a mathematical operation. Which property should this operation show? (0.5 p) Give a commonly used example of such an operation. (0.5 p)

Node to edge update (0.5 p), Edge to node update (0.5 p). Permutation invariance (0.5 p). E.g. Sum, mean or max (0.5 p)

0 1 2 3

i) DeepLab proposes a solution to increase the size of the receptive field without increasing the number of weights in a network. Name this solution and explain how it works. (0.5p + 1p). DeepLabv3+ uses a special type of solution to reduce the computation needed for convolutions significantly. Name this solution and explain how it works. (0.5 p + 1p)

Atrous (dilated) convolutions (0.5p). The convolutions weights are spread out (dilated), so that receptive field increases without increasing the number of weights or parameters (Convolutions filter with "holes"). (1p) Depth-wise separable convolutions (0.5p) Filters are applied only at certain depths of the features and then 1x1 convolutions are used. (1p)

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.