Solution cv3 s22 retake

Computer Vision III: Detection, Segmentation and Tracking (Technische Universität München)

E0032
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 / Retake	**Date:**	Friday 30th September, 2022
Examiner:	Prof. Dr. Ing. Laura Leal-Taixe	**Time:**	10:45 – 12:15

	P 1	P 2	P 3	P 4	P 5	P 6
I						

Working instructions

- This exam consists of **12 pages** with a total of **6 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 65.5 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 __________

□ Exam empty

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 following statements is/are true for YOLO versions and SSD (Single shot multibox detector)?

□ YOLOv2 completely removes anchor boxes and this enables faster inference.

□ SSD makes predictions at different resolution scales.

□ YOLO is fully convolutional but not end-to-end trainable.

□ SSD and YOLO struggle with small objects.

b) Which of the following applies for 2-Stage Detectors.

□ The region proposal in the R-CNN network does not contain trainable weights; this first stage is manually-designed, not trainable.

□ Fast-RCNN includes the proposal generator to significantly speed-up the proposal generation process that was previously handcrafted.

□ The Region Proposal Network uses a fixed number of anchor boxes per location where for each anchor box a classification scores (object / no-object) is predicted. Coordinate regression prediction is done later.

□ During training of the Region Proposal Network, if the IoU value of an anchor bounding box with the ground truth bounding box is ≥ 0.5 we use the anchor box as positive sample for the object/non-object classification in the Region Proposal Network.

c) Check all that apply for Multi-Object Tracking

□ Tracktor uses PointNet Detector to directly regress bounding boxes from the current frame to the next frame.

□ Linear motion models do not work well in Multi-Object Tracking scenarios because of the pedestrians complex motion behaviour.

□ We can add extra nodes during Hungarian matching to account for possible missing detections.

□ Viewing tracking as a retrieval problem, we want to learn a distance function between two images / bounding boxes to determine if the two images contain the same person or not.

d) Check all that apply for Transformers:

□ Q, K, V must have the same dimensions.

□ Q and K must have the same dimension, but V can have a different dimension.

□ Q and V must have the same dimension, but K can have a different dimension.

□ V and K must have the same dimension, but Q can have a different dimension.

e) In YOLACT, the Protonet generates k protomasks. Check all that apply:

□ k is the number of classes in the dataset.

□ k is a hyperparameter.

□ k is the number of classes in the minibatch.

□ k is the number of classes in the dataset, plus 1 (for the background class).

f) Check all that apply for Video Object Segmentation:

- □ In VOS, by contrast to multi-object tracking, we always assume only a single object is present throughout the sequence.
- □ In unsupervised (zero-shot) VOS, we assume strictly no labeled training data is provided at any training stage.
- □ In VOS, we are asked to estimate both instance and semantic labels for objects.
- □ By contrast to multi-object tracking, VOS methods often resort to test-time fine-tuning / optimization.

Problem 2 Various Topics (13 credits)

0 □ □ 1 □ □ 2 □ □ 3 □ □ 4 □ □ 5 □ □ 6 □ □ 7 □

a) You are given a deep network that is trained to detect objects in an image. A forward pass through this model gives you bounding-box predictions with their location (x, y, w, h) and confidence (ranging from 0 to 1). Your goal is to choose the best predictions and evaluate the performance of your model.

- Initially you are using all predictions from the network. After qualitative analysis, you realize that some predictions are of low quality. Suggest a way to eliminate these by only using confidence scores (0.5 p).
- Even after employing your answer to the previous question, you still end up with many boxes trying to explain one object. Name the algorithm that would solve this problem by eliminating overlapping predictions (0.5 p). Explain how this algorithm works step-by-step by clearly stating what the thresholds of this algorithm correspond to for this specific task (2p).

Now you are ready to quantitatively analyze the performance of your model on the validation set. For this, you want to implement mean average precision (mAP) metric.

- How does one define a "positive match (true positive)" between a prediction box and a ground truth box? (0.5p) How does one define a "negative match (false positive)" (0.5p)
- Explain step-by-step how mAP metric is computed (3 p).

b) DeepLab (first version) proposes three solutions to solve three challenges of semantic segmentation: reduced feature resolution, objects existing at multiple scales and poor localization of edges. Name and briefly explain DeepLab's three propositions to solve each of these challenges (1p each)

□ 0
□
□ 1
□
□ 2
□
□ 3

c) What is the goal and expected output of human pose estimation task in 2D domain? (1p) List two challenges related with this task. (0.5p each) Explain heatmap prediction approach for this task (0.5p) and compare it with direct regression approach (0.5p).

□ 0
□
□ 1
□
□ 2
□
□ 3

□ Page empty

Problem 3 Two-Stage Detectors and Multi-Object Tracking (12.5 credits)

0 □ 1 □ 2 □ 3 □

a) How did SPP speed up the slow test time of R-CNN? Name (1P) and explain the algorithm (1P) What is happening during backpropagation? (1P)

0 □ 1 □ 2 □ 3 □

b) Describe to goal of deep metric learning (1P). Explain the intuition behind the triplet loss (1P). Name and explain one training trick to improve the training of the Siamese network (1P).

0 □ 1 □ 2 □

c) Suppose you want to build a Multi-Object Tracker based on deep metric learning, i.e., re-identification. Name two problems that often occur that are especially challenging for re-identification (1P). If we want to stick with an online tracking approach, what could we do to overcome this? Name two proposals. (1P)

d) Considering an offline graph-based approach where nodes represent the detections. Name and describe the minimization problem corresponding to Multi-Object Tracking with graphs (1P).

0 1 2 3 4

Suppose we want to solve graph-based Multi-Object Tracking with message passing networks which features can be used to represent geometry information (name 2) and where could we use them in the graph (1.5P)? Also, name and give the formula of the two required update steps (2P).

□ Page empty

Problem 4 Transformers and Semantic Segmentation (8 credits)

0 □ 1 □ 2 □ 3 □

a) In self-attention, there is the three matrices Query (Q), Key (K), and Value (V) are given as input to the self-attention layer. How are these matrices generated in the first place? (2p)

0 □ 1 □ 2 □

b) Describe how DETR can be adjusted to get used for panoptic segmentation. Please describe in high terms the architecture of the segmentation network (1p), and the input to it (1p).

0 □ 1 □ 2 □ 3 □

c) What type of convolutions are used in the semantic head of UPSNet (1p)? Describe the difference between them and dilated convolutions (1p). Why they are needed (1p)?

Page empty □

Problem 5 Learning from 3D data (11 credits)

We typically represent our data and feature maps as (dense) tensors when working with images.

a) Say we would like to represent 3D data, such as point clouds, as dense tensors. What do we need to do before we can start learning representations (1p)? What challenges do we face (list at least two)? (2p)

0 1 2 3

b) Instead of using dense tensors, we can represent data using an alternative representation using different data structures that we have discussed in the lecture. Write how we could represent a 2D matrix:

$$A = \begin{bmatrix} 1 & \text{NaN} \\ 3 & \text{NaN} \end{bmatrix} \tag{5.1}$$

using such a representation (explain and write down names of data structures and the content). Assume NaN signals unobserved region (no data).

0 1 2 3 4 5

c) Describe two key differences between (lidar-based) 3D multi-object tracking and 4D lidar panoptic segmentation (1.5p each).

0 1 2 3

□ Page empty

Problem 6 Video Object Segmentation (9 credits)

In lectures on video-object segmentation (VOS), we discussed an approach for VOS via test-time optimization, OSVOS (one-shot video object segmentation).

0 1 2 3 4

a) Describe the key ideas behind this approach (two key ideas 2p+2p) .

0 1 2

b) Explain why this approach has issues with shape consistency, *i.e.*, occasionally producing segmentation results that do not conform to the boundaries of objects (2p).

0 1 2 3

c) Briefly explain an approach (discussed in the lecture) on how the shape consistency can be improved *wrt.* vanilla OSVOS (1p for the key idea, 2p for explanation).

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

☐ Page empty