Introduction to Computer Vision

Spring 2022 Final-term Exam

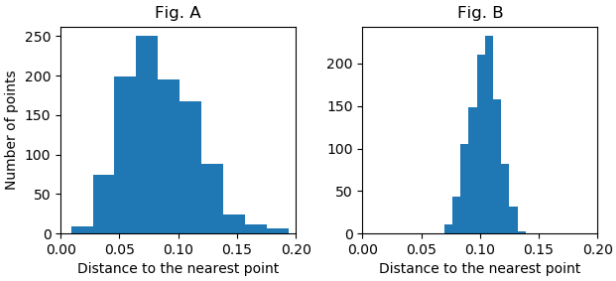
Name_____ Student ID_____

I. Multiple-Choice Questions (3*10 = 30 points)

Note that there **may be more than one correct answer**, please consider each of the choices separately and select all that apply.

1. Which of the following choices of depth cameras is reasonable to obtain depths:
   A. Use LiDAR cameras in the outdoor environment.
   B. Use passive stereo vision based cameras under low-lighting conditions.
   C. Use LiDAR cameras in the indoor environment.
   D. Use structural light cameras to sense a textureless wall.

2. We want to sample a point cloud from a unit sphere mesh. We first uniformly sample 5120 points from the mesh and then use furthest point sampling (FPS) or random selection (without duplication) to sample 1024 points from the 5120 points, which gives us point cloud A and B. Here we show two histograms: for each point in A, we compute the distance to its nearest neighbor in A and plot the statistics in Fig A; we do the same for point cloud B. Which of the following statements is true:



   A. When uniformly sampling the 5120 points, we need to compute the areas of each face and then use them as the weights to decide which faces are to be sampled from.
   B. Fig. A and Fig. B are equally likely from FPS.
   C. Fig. A is more likely from FPS than from random sampling.
   D. If we only sample 10 points from the 5120 points, the results of FPS can dramatically change when we change the first point of FPS.

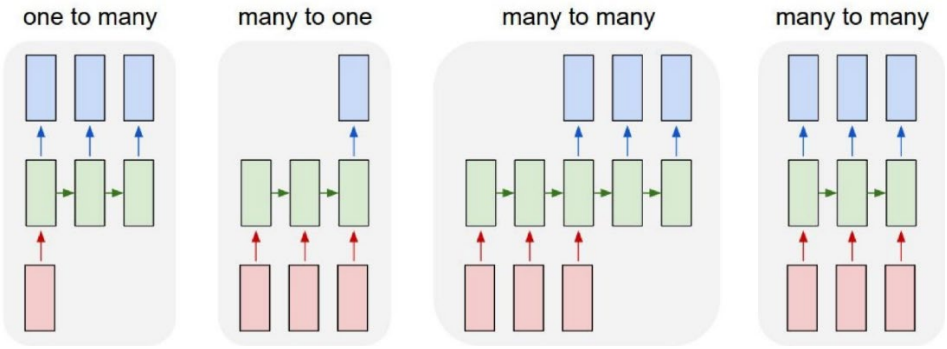
3. Tom wants to build his own Mask-RCNN for instance segmentation of hand-writing digits. Please help him to choose the appropriate modules/designs from below to build the Mask-RCNN:
   A. Detector: Fast-RCNN
   B. Cropping Features: RoI Pool
   C. Non-Maximal Suppression
   D. Uniform rotation augmentation from $[0, 2\pi]$.

4. Which of the following statements about rotation is true:
   A. Under Euler representation, if a rotation is $[\alpha, \beta, \gamma]$, then the inverse of this rotation is $[-\alpha, -\beta, -\gamma]$.
   B. The unit quaternions have four degree-of-freedoms and can represent rotations.
   C. Under axis angle representation, if a rotation is $[\hat{e}, \theta]$, then the inverse of this rotation is $[\hat{e}, -\theta]$.
   D. Under unit quaternion representation, for two rotations q_A and q_B, if we first rotate using q_A and then using q_B, we can represent the composed rotation as $q_B q_A$.

5. Which of the following statements is true about 3D SparseConv:
   A. Compared to PointNet, SparseConv has a larger spatial receptive field.
   B. Compared to dense 3D Conv, 3D SparseConv is more efficient.
   C. Similar to dense 3D Conv, SparseConv has translation equivariance.
   D. Similar to PointNet, SparseConv is invariant to rotation transformations.

6. Choose the correct correspondence between different RNN types and tasks:

RNN types:



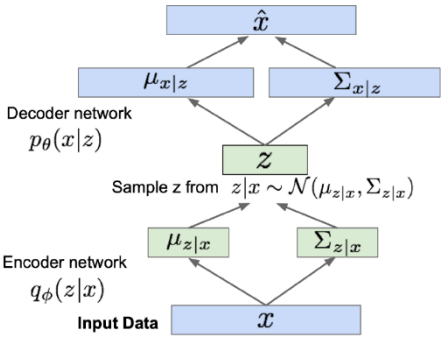
Tasks:

a. Image captioning; b. Video captioning;
c. Action prediction; d. Video classification on frame level;

   A. a - Ⅰ, b - Ⅲ, c - Ⅱ, d - Ⅳ
   B. a - Ⅰ, b - Ⅲ, c - Ⅳ, d - Ⅱ
   C. a -Ⅳ, b - Ⅰ, c - Ⅲ, d - Ⅱ
   D. a -Ⅳ, b - Ⅰ, c - Ⅱ, d - Ⅲ

7. Choose the correct statements about optical flow:
   A. Lighting changes may lead to apparent motions.
   B. Actual motions always lead to apparent motions.
   C. Lucas-Kanade Flow assumes that neighboring pixels have consistent optical flow.
   D. The optical flows estimated from corner regions are usually better than from edges.

8. Assume that we want to train a VAE with an image set $X = \{x^i\}_{i=0,1,\ldots,n}$ for training and the architecture of VAE is shown in the right figure. Choose the correct statements below:



   A. The VAE loss exactly maximizes $\sum_{i=0}^{n} \log p_\theta(x^i)$.
   B. $p_\theta(x)$ can't be greater than 1.
   C. The encoder is used to approximate $p_\theta(z|x)$ and is not needed for sampling new images.
   D. In practice, we can omit $\Sigma_{x|z}$ and simply use L2 loss as reconstruction loss.

9. The loss function of VAE is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right],$$

   where the first term is called KL loss and the second term is called reconstruction loss. Choose the correct statements below:

   A. The reconstruction term tends to decrease the variance $\Sigma_{z|x}$ of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$ during training.
   B. The reconstruction term has no effect on the variance $\Sigma_{z|x}$ of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$ during training.
   C. The KL term tends to decrease the variance $\Sigma_{z|x}$ of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$ to zero during training.
   D. The KL term has no effect on the variance of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$ during training.

10. Choose the correct statements about GAN:

   A. If the generator has a low loss but the generated images are fake, it may because the discriminator is too weak.
   B. If the generator has a high loss and can't converge, it may because the discriminator is too strong.
   C. If we use the non-saturating GAN loss, then the generator is very hard to grow at the very beginning of the training.
   D. Compared to the reconstruction loss used in VAE, GAN loss is more favored in generating photorealistic images.

II. True or False (2*10 = 20 points)

1. Chamfer distance (CD) is more sensitive to changes in sampling patterns than Earth Mover's distance (EMD). (False)
2. One property of signed distance function is that at almost everywhere the length of its gradient with respect to the change in positions is 1. (True)
3. Since the architecture of PointNet++ decoder doesn't rely on the skip links from the encoder side, the decoder can work without the skip links. (False)
4. When solving an Orthogonal Procrustes Problem, two pairs of 3D-3D correspondences would be sufficient for fitting a 3D rotation. (False)
5. Normalized Object Coordinate Space (NOCS) build a category-level reference frame by jointly normalizing the rotation, translation and scale of objects in the same category. (True)
6. The M in LSTM stands for memory, which refers to the cell state. (True)
7. When sampling an RNN, to obtain a fixed-length sequence with a higher probability, beam search is generally better than greedy sampling. (True)
8. One advantage of PixelCNN over PixelRNN is that PixelCNN can run in parallel when generating new images. (False)
9. Regarding categorizing different generative models, VAE belongs to tractable methods under modeling explicit density while GAN belongs to methods that implicitly model density. (False)
10. The reparameterization trick makes z differentiable with respect to the encoder, enabling VAE to be trained end-to-end.(True)

III. Short-Answer Questions (12+8=20 points)

3.1 (12 points) PointNet is one of the most important baselines for point cloud learning. Here, for point cloud data $\{x_1, x_2, \dots x_n\}$ $(x_i \in \mathbb{R}^3)$, we have two MLP networks h_1 and h_2, each of them is composed of two layers of FC and a ReLU in between. For example, if h_1 takes input k dimension, maps to p dimension, activates using ReLU, and finally maps to q dimension, then we denote it as $h_1 = MLP(k_1, p_1, q_1)$.

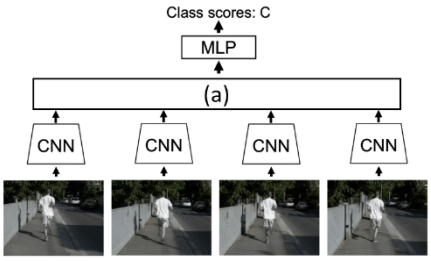
1) Please build a vanilla PointNet with 1024D in bottleneck for a 40-class classification problem using the two MLP networks. You need to explain how your h_1 and h_2 work together and decide the parameter k_i, p_i, q_i $(i = 1,2)$. You can also use MaxPool and SoftMax Layer, if needed. (4')

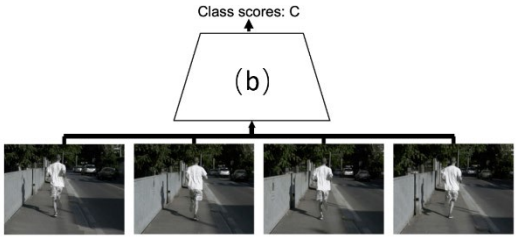
2) Prove that your PointNet is permutation invariant. (2')

3) To segment a point cloud into multiple parts, we also consider building our network based on PointNet. You are allowed to further use another MLP $h_3 = MLP(k_3, p_3, q_3)$. Please explain your network architecture and specify any other operators you need to use. (4')

4) Prove your PointNet-Segmentation network is permutation equivariant.(2')

3.2 (8 points) For the task of video classification, some popular kinds of methods are late fusion, early fusion, 3D CNN, etc. In the following questions, please point out one drawback of each design **briefly**.





1) The left figure shows the architecture of late fusion. If layer (a) is “flatten”, the feature after layer (a) has the shape of TDH’W’; If layer (a) is “average pooling over space and time”, the feature after layer (a) has the shape of D. Please point out one drawback of each architecture with different designs of layer (a). (4’)

2) The right figure shows the architecture of early fusion if layer (b) is “2D CNN”. Please point out one drawback of early fusion. (2’)

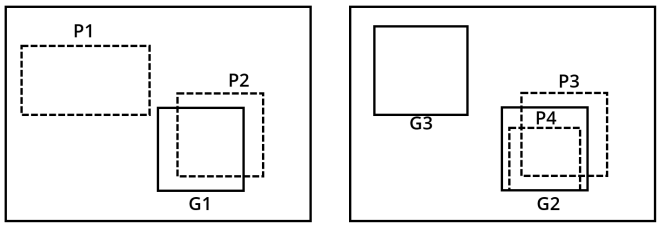
3) The right figure shows the architecture of 3D CNN if layer (b) is “3D CNN”. Please point out one drawback of 3D CNN. (2’)

IV. Calculation and Proof Questions (10+8+12=30 points)

4.1 (10 points) In this question, we want to evaluate a single-class object detector. Here we have two test images I1 and I2, where I1 contains 1 GT object (G1) while I2 contains two GT objects (G2, G3). Our detector outputs two predicted bounding boxes P1 and P2 for I1 and two bounding boxes P3 and P4 for I2. Their positions and confidence are shown in the table below.
Please answer the following questions:

1) Compute the IoUs of G1-P2, G2-P3, G2-P4 pairs. (5’)

2) With an IoU threshold of 0.5, please draw the Precision-Recall curve and calculate the AP. (The recall is defined from 0 to 1 at a step size of 0.1, and no interpolation is required.) (5’)



Name	Type	Top-left coordinate	Bottom-right coordinate	confidence
G1	Ground-Truth	(200, 300)	(400, 500)	-
G2	Ground-Truth	(200, 300)	(400, 500)	-
G3	Ground-Truth	(10, 10)	(220, 200)	-
P1	Prediction	(40, 40)	(180, 180)	60%
P2	Prediction	(180, 320)	(380, 520)	78%
P3	Prediction	(180, 320)	(380, 520)	88%
P4	Prediction	(220, 310)	(400, 490)	98%

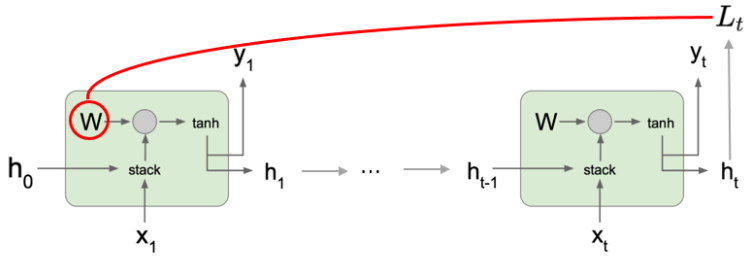
* Coordinates point from top left to the bottom right, and follow the order of (row, cols).

4.2 (8 points) Quaternion is an efficient rotation representation. Please prove the following properties of the quaternions:

1) For a unit quaternion q, its conjugate q' is equal to its inverse q^{-1}. (4')

2) The unit quaternion q and $-q$ represent the same rotation. (4')

4.3 (12 points) A standard RNN is shown below.



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$W = (W_{hh} \quad W_{hx})$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \text{(a)}$$

$$\frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T}\boxed{\text{(b)}}\frac{\partial h_1}{\partial W} \quad \text{(assume no non-linearity)}$$

1) Calculate (a) and (b). Note that for (a) you can directly use *tanh'* and don't have to solve the detailed expression of it. (4')

2) Based on 1), point out two possible problem of the gradient $\frac{\partial L_T}{\partial W}$ (4')

3) Simply describe a solution for each problem in 2). Explaining the Key idea is enough and no need to write down the specific equation. (4')