Midterm Exam

Graded

Student

Jonas Li

Total Points

59 / 73 pts

Question 1

Q1. True / False **10** / 10 pts

1.1 **1.1** **1** / 1 pt

✔ **– 0 pts** Correctly chose False

– 1 pt Incorrect

1.2 **1.2** **1** / 1 pt

✔ **– 0 pts** Correctly chose False

– 1 pt Incorrect

1.3 **1.3** **1** / 1 pt

✔ **– 0 pts** Correctly chose True

– 1 pt Incorrect

1.4 **1.4** **1** / 1 pt

✔ **– 0 pts** Correctly chose False

– 1 pt Incorrect

1.5 **1.5** **1** / 1 pt

✔ **– 0 pts** Correctly chose False

– 1 pt Incorrect

1.6 **1.6** **1** / 1 pt

✔ **– 0 pts** Correctly Chose False

– 1 pt Incorrect - The mapping from 3D to 2D loses depth information, meaning that multiple 3D points along the same ray can correspond to the same 2D point. As a result, it is not possible to uniquely recover a 3D point from a single 2D projection using the projection matrix alone.

– 1 pt No attempt

1.7 **1.7** **1** / 1 pt

✔ **– 0 pts** Correctly Chose True

– 0 pts Incorrect

1.8 **1.8** **1** / 1 pt

✔ **– 0 pts** Correctly chose True

– 1 pt Incorrect

1.9 **1.9** **1** / 1 pt

✔ **– 0 pts** Correctly Chose True

– 1 pt Incorrect

– 1 pt Did not attempt

1.10 **1.10** **1** / 1 pt

✔ **– 0 pts** Correctly chose False

– 1 pt Incorrect: parallel vertical lines in 3D will not converge to a vanishing point

Question 2

Q2. MCQ (Single Answer) **8** / 8 pts

2.1 **2.1** **2** / 2 pts

✔ **– 0 pts** Correctly chose A

– 2 pts Incorrect

2.2 **2.2** **2** / 2 pts

✔ **– 0 pts** Correctly chose C

– 2 pts Incorrect

2.3 **2.3** **2** / 2 pts

✔ **– 0 pts** Correctly chose C

– 2 pts Incorrect

2.4 **2.4** **2** / 2 pts

✔ **– 0 pts** Correctly chose B

– 2 pts Incorrect

Question 3

Q3. Select All that Apply **7** / 14 pts

3.1 **3.1** **1** / 2 pts

– **0 pts** Correctly chose AB

✔ – **1 pt** Selected either only A or B (and no incorrect options were chosen)

– **2 pts** Didn't choose AB, or chose A or B but also chose an incorrect option, or didn't choose A nor B

3.2 **3.2** **1** / 2 pts

– **0 pts** Correctly chose ABD

✔ – **1 pt** Selected 1 or 2 of ABD (and no incorrect options were selected)

– **2 pts** Didn't choose ABD, or chose A or B or D but also chose an incorrect option

3.3 **3.3** **1** / 2 pts

– **0 pts** Correctly chose AB

✔ – **1 pt** Selected either only A or B (and no incorrect options were chosen)

– **2 pts** Didn't choose AB, or chose A or B but also chose an incorrect option, or didn't choose A nor B

3.4 **3.4** **1** / 2 pts

– **0 pts** Correctly chose CD

✔ – **1 pt** Selected either only C or D (and no incorrect options were chosen)

– **2 pts** Didn't choose CD, or chose C or D but also chose an incorrect option, or didn't choose C nor D

3.5 **3.5** **0** / 2 pts

– **0 pts** Correctly chose CD

– **1 pt** Selected either only C or D (and no incorrect options were chosen)

✔ – **2 pts** Didn't choose CD, or chose C or D but also chose an incorrect option, or didn't choose C nor D

3.6 **3.6** **2** / 2 pts

✔ – **0 pts** Correctly chose AC

– **1 pt** Selected either only A or C (and no incorrect options were chosen)

– **2 pts** Didn't choose AC, or chose A or C but also chose an incorrect option, or didn't choose A nor C

3.7 **3.7** **1** / 2 pts

– **0 pts** Correctly chose CDE or DE

✔ – **1 pt** Selected a subset of CDE and no incorrect options were chosen

– **2 pts** Didn't choose CDE or DE, or chose a subset of CDE but also chose an incorrect option, or didn't select anything

Question 4

Q4. Flow Matching **6** / 10 pts

4.1 **Flow Matching Objective** **4** / 4 pts

- ✔ **– 0 pts** Correct
- **– 4 pts** wrong answer/ no answer
- **– 2 pts** partially correct

4.2 **Update Rule** **0** / 4 pts

- **– 0 pts** Correct
- **– 2 pts** used x_0 despite being told not to use x_0
- ✔ **– 4 pts** wrong answer/ no answer
- **– 2 pts** missing 2t

4.3 **Dimensionality Requirement** **2** / 2 pts

- ✔ **– 0 pts** Correct
- **– 2 pts** wrong
- **– 1 pt** partially correct

Question 5

Q5 Absolute Positional Encoding **5** / 7 pts

5.1 **APE value matrix** **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt mixed up rows and columns (e.g. put low-freq on the right instead of left)

– 2 pts incorrect values (e.g. applied an extraneous function like sin/cosin, put values in incorrect order)

– 3 pts blank / did not attempt

5.2 **HF encoding** **1** / 1 pt

✔ **– 0 pts** Correctly chose left column

– 1 pt incorrect

5.3 **LF encoding** **1** / 1 pt

✔ **– 0 pts** Correctly chose column 2

– 1 pt incorrect

5.4 **why HF is needed** **0** / 2 pts

– 0 pts Correctly mentioned:

- the low-frequency values among adjacent patches will be indistinguishable (1/10000 ≈ 2/10000), thus we need high-frequency values to better differentiate them

✔ **– 2 pts** incorrect

the LF is used for modeling larger context window, not shorter ones

Question 6

VAE **10** / 11 pts

6.1 **limitations of vanilla autoencoders** **1** / 2 pts

– **0 pts** Correct

✔ – **1 pt** Minor Error

– **2 pts** Incorrect

6.2 **difficulty in learning true posterior** **5** / 5 pts

✔ – **0 pts** Correct

– **2 pts** Does not mention intractability

– **2 pts** Other minor error

– **5 pts** Incorrect

6.3 **VAE objective** **4** / 4 pts

✔ – **0 pts** Correct

– **2 pts** Does not mention Reconstruction Loss

– **2 pts** Does not mention KL Divergence

– **4 pts** Incorrect

– **1 pt** Minor Error 1

– **1 pt** Minor Error 2

Question 7

optical flow **13** / 13 pts

7.1 **flow equation** **3** / 3 pts

✔ **– 0 pts** Correct

– 3 pts both answers are wrong

– 1.5 pts one answer is wrong

7.2 **flow equations under focal length conditions** **3** / 3 pts

✔ **– 0 pts** Correct

– 3 pts 3 answers are wrong

– 2 pts two answers are wrong

– 1 pt one answer is wrong

7.3 **f value** **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt one answer is wrong

– 2 pts two answers are wrong

– 3 pts three answers are wrong

7.4 **camera movement** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts one answer is incorrect

– 4 pts both are incorrect

CS 280
Spring 2025 Computer Vision Exam

- **You have 80 minutes to complete the exam.** The exam starts at 14:10 PM.
- **The exam is closed book, closed notes except your one-page two-sided cheatsheet.**
- **Mark your answer on the exam itself in the space provided specifically for each problem.** When asked for, explain your response and reasoning succinctly, but clearly and convincingly. Please write neatly and legibly, because *if we can't read it, we can't evaluate it.*
- **Sign the Pledge of Academic Integrity.** By my honor, I affirm that I will not cheat. **Please sign your name in the Signature below and write your SID.**

Signature	Yun Li
Name	Yunzhe Li
SID	3040802664

For Staff Use Only:

Total	/ 73

Q1. [10 pts] True / False

Circle True (T) or False (F) for each statement. **You get +1/0 points for each correct/incorrect answer.**

1. [T / F] In the Transformer, causally-masked self-attention ensures that a token can attend to future tokens during training.
2. [T / F] The computational complexity of the self-attention mechanism is log-linear with respect to the sequence length.
3. [T / F] The Gaussian filter is equivalent to convolving an image with a Gaussian kernel, which is the solution of the heat equation.
4. [T / F] Anisotropic diffusion leads to increased smoothing near edges.
5. [T / F] The main reason for using convolutional layers in CNNs is to achieve invariance to rotations.
6. [T / F] A projection matrix maps a 3D point in the world into a 2D point on the camera imaging plane, and vice versa.
7. [T / F] One can transform an image by a homography H by multiplying each pixel coordinates p by the homography, as in $p' = Hp$.
8. [T / F] Walking around a table made of Lambertian surfaces, the table would appear equally bright from every viewing direction.
9. [T / F] Filters approximating the oriented cells in V1 discovered by Hubel and Wiesel emerge naturally if you learn a sparse code on a dataset of natural images
10. [T / F] Any two parallel lines in 3D converge to a vanishing point.

Q2. [8 pts] Multiple Choice Questions

1. [2 pts] What is the main advantage of using masked autoencoders for pretraining ViTs over other methods like DINO?

   A. They drop a lot of tokens during training to make learning efficient.

   B. They utilize a contrastive learning objective to directly align different augmented views of the same image.

   C. They require task-specific labels during pretraining, which are necessary for learning highly discriminative features.

   D. They construct negative pairs through heavy data augmentation.

   Your answer: A

2. [2 pts] Which of the following is a key mathematical property of self-attention that differentiates it from traditional convolution?

   A. The attention weights are computed from pre-learned filters applied identically to all tokens.

   B. Self-attention restricts each token to attend only to a small, fixed neighborhood in the sequence.

   C. The attention weight matrix is computed dynamically based on pairwise similarity among tokens.

   D. Self-attention has strictly fewer parameters and cannot be scaled up for large inputs.

   Your answer: C

3. [2 pts] Which filtering technique is the most similar to the self-attention mechanism in transformers?

   A. Gaussian filtering.

   B. Bilateral filtering.

   C. Non-local means.

   D. Anisotropic diffusion.

   Your answer: C

4. [2 pts] When computing the Fréchet Inception Distance (FID) between two sets of real images—one with 10,000 samples and another with 20,000 samples—the resulting FID is not zero. What is the most likely reason for this outcome?

   A. The FID metric inherently assigns a non-zero value even when comparing identical distributions.

   B. Estimation errors in the sample mean and covariance of the Inception features occur due to different sample sizes.

   C. The FID calculation uses a non-Euclidean distance metric that is sensitive to image quality differences.

   D. Larger sample sizes always produce higher FID values regardless of distribution similarity.

   Your answer: B

Q3. [14 pts] Multiple Choice Questions: Select All tha. Apply.

You will receive +2 points for selecting all the correct answers, without any wrong answers. You will receive +1 points for selecting some of the correct answers, without any wrong answers. You will receive +0 points if any incorrect answers are selected.

1. [2 pts] Why is it challenging to apply vision transformers directly to vision tasks? (Select all that apply)
   A. Image resolution leads to an explosion in the number of tokens, making self-attention computationally expensive.
   B. Transformers lack an inherent spatial inductive bias, unlike CNNs.
   C. Vision transformers can not be pre-trained without large labeled datasets.
   D. Attention mechanisms construct Q, K from same tokens.

   Your answer (write all letters that apply): A

2. [2 pts] The KL divergence term in the VAE objective serves which of the following purposes? (Select all that apply.)
   A. It ensures that encoded representations $q(z|x)$ remain close to the prior $p(z)$ preventing arbitrary latent space organization.
   B. It prevents the encoder from collapsing into a deterministic mapping by enforcing stochasticity in the latent space.
   C. It directly maximizes the likelihood $p(x)$, ensuring better reconstructions of the input data.
   D. It regularizes the model by constraining the overall distribution of latent variables, making sure all samples are meaningful.

   Your answer (write all letters that apply): A

3. [2 pts] Select all statements that correctly describe characteristics of DINO:
   A. DINO teacher and student have the same number of trainable parameters.
   B. DINO does not require any negative pair sampling.
   C. DINO does not need data augmentation.
   D. DINO features are not semantically meaningful because it is not trained with labels.

   Your answer (write all letters that apply): B

4. [2 pts] Select all statements that correctly describe Masked Auto Encoders:
   A. Masked Autoencoders need lots of labeled data for training.
   B. For ImageNet 50% masking is the optimal to get best representations.
   C. During training MAEs are trained with patch normalized loss.
   D. MAE decoder has less number of parameters than encoder.

   Your answer (write all letters that apply): C

5. [2 pts] In flow matching model generation, comparing stochastic sampler with deterministic sampler, which of the following are true statements?

   A. Stochastic sampler requires more steps
   B. Stochastic sampler gives better generation quality than deterministic sampler
   C. Stochastic sampler helps explore the distribution
   D. Stochastic sampler may introduce more errors than deterministic sampler

Your answer (write all letters that apply): A C

6. [2 pts] Which of the following approaches could help mitigate the truncation errors in flow matching sampling?

   A. Replace a Euler sampler with a 2nd Heun sampler
   B. Start with a Gaussian noise with a smaller std than training
   C. Use more sampling steps and smaller step sizes
   D. Using a stochastic sampler

Your answer (write all letters that apply): A C

7. [2 pts] Which of the following statements regarding Rotary Positional Encoding (ROPE) are true? (Select all that apply.)

   A. ROPE is not flash attention friendly.
   B. ROPE reduces the memory complexity of transformers.
   C. ROPE has to be extended to 2D positional encoding by applying separate rotations along the horizontal and vertical axes.
   D. ROPE inherently encodes multiple frequency components within a single rotational transformation.
   E. ROPE allows relative positional information to be encoded naturally.

Your answer (write all letters that apply): E

Q4. [10 pts] Flow Matching

1. **[4 pts]** [Flow matching derivation] We define a time-dependent interpolation between two data points x_0 and x_1 as follows:

$$x_t = (1-t^2) \cdot x_0 + t^2 \cdot x_1 \quad t \in [0,1]. \tag{1}$$

We aim to learn a velocity field $v_\theta(x_t, t)$, with a neural network. Write the conditional flow matching objective for this setup.

Your answer:

$$\nabla_\theta \| k(x_1 - x_0) - v_\theta(x_t, t) \|^2$$

$$\frac{dx_t}{dt} = -2x_0 t + 2t x_1$$

$$= 2t(x_1 - x_0)$$

2. **[4 pts]** Using the same time-dependent interpolation (eq. 1 above), suppose we train a model $f_\theta(x_t, t) = x_1$ that always predicts the ground truth clean image x_1. To sample from this model at test time, we using the following pseudocode, Fill in the missing part, and **leave your answer in terms of $\hat{x}_1, x_t$ and t.**

Algorithm 1 Sampling pseudocode

Input: s: Total number of sampling steps
Output: Generated image x_1

// Initialization: sample the noise
$x_0 \sim \mathcal{N}(0, I)$
for $t = 0, \frac{1}{s}, \frac{2}{s}, \ldots, 1$ **do**
  // Use the model to predict the clean image
  $\hat{x}_1 = f_\theta(x_t, t)$ // Compute the velocity and then update
  $x_{t+\frac{1}{s}} =$ <answer here>
return x_1

Your answer:

$x_{t+\frac{1}{s}} = x_t + \frac{1}{S} \hat{x}_1$

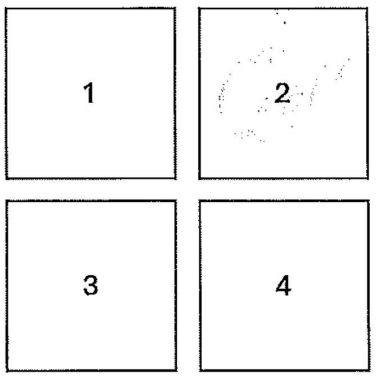
3. **[2 pts]** What is the reason why flow based generative models require the source distribution and target distribution to be of the same dimensionality?

Your answer:

Flow-based model denoise the source image step by step
In every step, the input and the output should be of the same dimensionality,
the scale to the source and the target distribution, they should be the same.
If not, we can not generate the expected image

Q5. [7 pts] Absolute Positional Encoding (APE)

1. We want to design a system that automatically recognizes images where the Cal logo appears in the top-right corner. To keep things simple, we will divide each image into 2×2 tokens, ordered as shown below:



Each token i is assigned a 2D Absolute Positional Encoding (APE) defined by $P_i = (P_{i1}, P_{i2})$:

$$P_{i1} = (-1)^i, \quad P_{i2} = \frac{i}{10000}$$

Write out the 4×2 matrix of APE values for the four tokens.

Your answer:

$$P = \begin{bmatrix} -1 & \frac{1}{10000} \\ 1 & \frac{2}{10000} \\ -1 & \frac{3}{10000} \\ 1 & \frac{4}{10000} \end{bmatrix}$$

2. **[1 pts]** Which column represents high-frequency encoding?

Your answer: left column

3. **[1 pts]** Which column represents low-frequency encoding?

Your answer: right column

4. **[2 pts]** In two sentences, explain why relying only on the low-frequency component could cause problems when determining whether the Cal logo is in the top-right token.

Your answer:

Low-frequency can not cover much position and will just focus on nearby neighbours.

Q6. [11 pts] Variational Autoencoders (VAEs)

Limit your answer to 2 sentences.

1. **[2 pts]** What is the limitation of standard/vanilla autoencoders as generative models?
   Your answer:

   It can not generate meaningful new outputs

2. **[5 pts]** Ideally, we want to learn the true posterior $p(z|x)$ Why is this distribution difficult to optimize directly?
   Your answer:

   If we want to learn $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$
   we don't know exactly $p(x)$
   so we have to learn $q(z|x)$ and make it close to $p(z)$ $D_L(q(z|x)||p(z))$

3. **[4 pts]** What are the two competing terms in the VAE objective, and what role does each play?
   Your answer:

   objective: $E_{z\sim q(z|x)} \log p(x_i|z) + \log p(z) + H(q(z|x_i))$

   Two competing terms: ① $p(x|z)$ ② $q(z|x)$
   We want to maximize $p(x|z)$ while regularize $q(z|x)$
   $q(z|x)$ is used to be close to $p(z)$

Q7. [13 pts] Optical Flow

1. [9 pts] Below is the optical flow equation, where f is the focal length, V is the translational velocity and W is the angular velocity.

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \frac{1}{Z}\begin{bmatrix}-f & 0 & x\\ 0 & -f & y\end{bmatrix}\begin{bmatrix}V_x\\ V_y\\ V_z\end{bmatrix} + \frac{1}{f}\begin{bmatrix}xy & f^2-x^2 & fy\\ -f^2-y^2 & -xy & -fx\end{bmatrix}\begin{bmatrix}W_x\\ W_y\\ W_z\end{bmatrix}.$$

Now assume, we are in a scene and our $V_x = V_y = V_z = 0$, and $W_z = W_x = 0$. We are only rotating along the y axis.

(a) Write the flow equation for the above conditions.

Your answer:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \frac{1}{f}\begin{bmatrix}xy & f^2-x^2 & fy\\ -f^2-y^2 & -xy & -fx\end{bmatrix}\begin{bmatrix}0\\ W_y\\ 0\end{bmatrix} = \frac{1}{f}\begin{bmatrix}f^2-x^2\\ -xy\end{bmatrix}$$

(b) Write the flow equations for the cases where focal length f is very small, $f = 1$, and f is very large. Write any assumptions that you made.

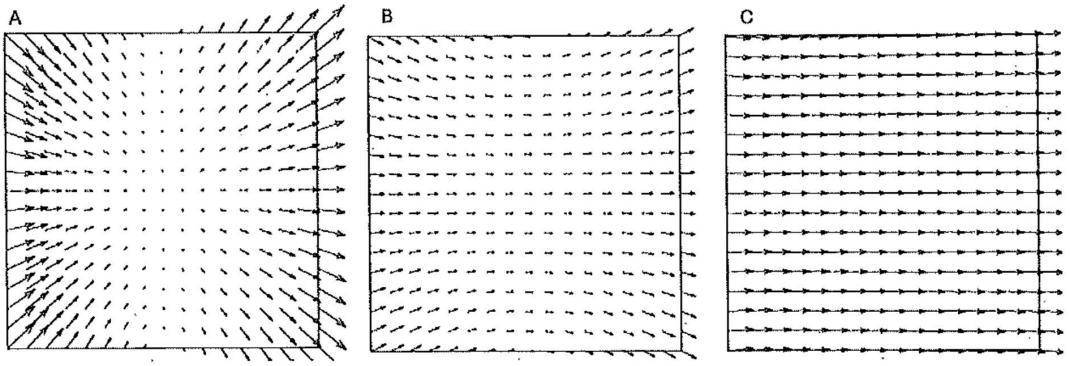
Your answer:

If $f = 1$ then $\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \begin{bmatrix}1-x^2\\ -xy\end{bmatrix}$

If f is very large $\to \infty$ then, $\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \lim_{f\to\infty}\begin{bmatrix}\frac{f^2-x^2}{f}\\ \frac{-xy}{f}\end{bmatrix} = \begin{bmatrix}f\\ 0\end{bmatrix}$

If f is very small $\to 0$ then $\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \lim_{f\to 0}\begin{bmatrix}\frac{f^2-x^2}{f}\\ -xy\end{bmatrix} = \frac{1}{f}\begin{bmatrix}-x^2\\ -xy\end{bmatrix}$ $f\to 0$

(c) Now, match the optical flow observations below with the focal length conditions.

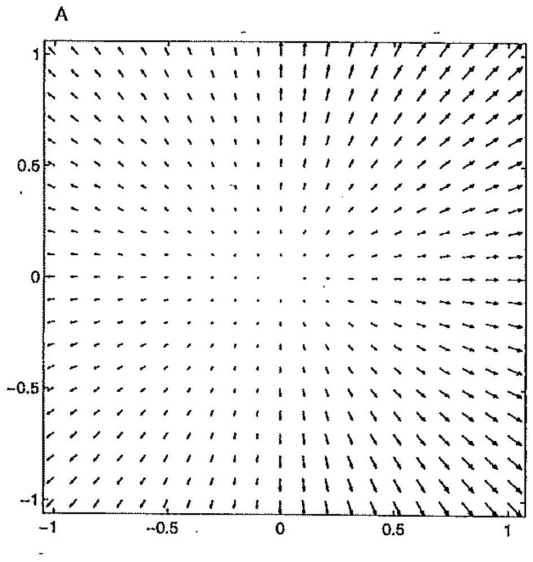


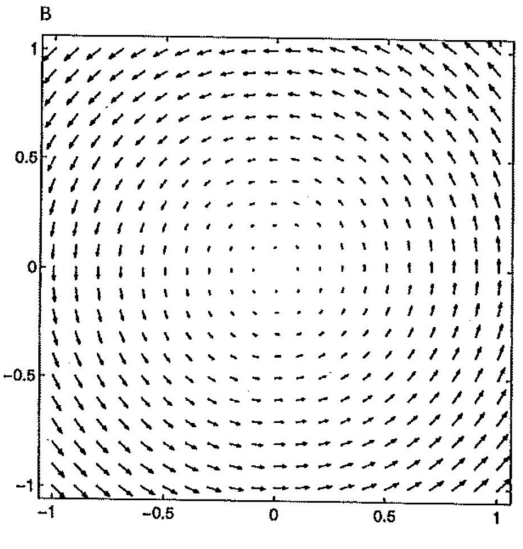
f is very small **Your answer:** A

$f = 1$ **Your answer:** B

f is very large **Your answer:** C

2. [4 pts] Describe the camera movement in a static scene which would result in the following optical flows. Th x-axis is horizontal, y-axis is vertical and z-axis is perpendicular to the plane of the paper. Your answer should be in terms rotations and translations with their associated directions





optical flow	A	B
camera movement	Move forward along the Z-axis	Rotation along z-axis counter-clockwise

THIS PAGE IS FOR SCRATCH WORK

THIS PAGE IS FOR SCRATCH WORK