Homework 2

Release: 10/31/2022 Due: Sun. 11/13/2022, 11:59 PM

- You are **allowed** to consult any external resources but you must cite them. You are also **allowed** to discuss with each other but you need to acknowledge them. However, your submission must be your own work; specifically, you must **not** share your code or proof.
- Your submission has 3 parts. The first part is a single PDF file, containing proof, code, and results of problem 1, as well as the report for problem 2 and 3. The second part is a zip file containing your code for problem 2 and 3. The third part is a submission to the leaderboard.
- This homework is worth 30/100 of your final grade.
- This homework itself contains 26 points and **5** extra credit!

Problem 1 (Deform a shape).[5pt + 5pt] In this problem, we will practice part of what we learned in the image-to-3D lecture for shape deformation.

1. (Laplacian)[3pt] Given a mesh $M = (V, E, F)$, we assume that the adjacency matrix is $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix where $D[i,i]$ is the degree of the i-th vertex. The **Laplacian** matrix is defined as $L = D - A$.

   Prove that:

   (a) $\sum_{(i,j)\in E} \|x_i - x_j\|^2 = x^T L x$ for $x \in \mathbb{R}^n$. [1pt]

   (b) $L \in \mathbb{S}^n_+$, i.e., L is a symmetric and positive semi-definite matrix. [1pt]

   (c) For the data matrix $P \in \mathbb{R}^{n\times 3}$ where each row corresponds to a point in $\mathbb{R}^3$, denote the columns of P as $P = [x, y, z]$ and rows of P as $P = [p_1^T; p_2^T; \ldots; p_n^T]$, show that $\sum_{(i,j)\in E} \|p_i - p_j\|^2 = x^T L x + y^T L y + z^T L z$. (hint: Use the conclusion from 1(a)) [1pt]

2. **Normalized Laplacian**[2pt] is defined as the normalized version of the Laplacian matrix above:

$$L_{norm} = D^{-1} L \tag{1}$$

   (a) Prove that the sum of each row of L_{norm} is 0. [1pt]

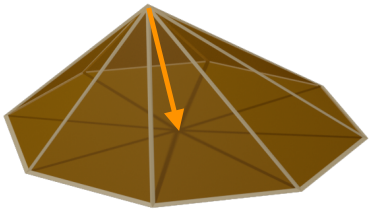

Figure 1: Curvature approximation by discrete Laplacian

   (b) The difference between a vertex x and the average position of its 1-ring neighborhood is a quantity that provides interesting geometric insight of the shape (see Figure 1). It can be shown that,

$$x - \frac{1}{|N(x)|} \sum_{y_i \in N(x)} y_i \approx H\vec{n}\Delta A \tag{2}$$

for a good mesh, where $N(x)$ is the 1-ring neighorhood vertices of x by the mesh topology, $H = \frac{1}{2}(\kappa_{min} + \kappa_{\max})$ is the mean curvature at x (in the sense of the underlying continuous surface being approximated), $\vec{n}$ is the surface normal vector at x, and ΔA is a quantity proportional to the total area of the 1-ring fan (triangles formed by x and vertices along the 1-ring).

Define $\Delta p_i := p_i - \frac{1}{|N(p_i)|}\sum_{p_j \in N(p_i)} p_j$. Prove that $\Delta p_i = [L_{norm}P]_i$, where P and p_i are defined as in 1(c), and $[X]_i$ is to access the i-th row of X. [1pt]

3. (programming, **extra credit**) [5pt] Please load the `source.obj` and `target.obj` files using the trimesh library of Python, and optimize to deform the vertices of the `source.obj` to match `target.obj`. Plot the source object, target object, and deformed object.

   (a) Warm-up by Chamfer-only loss: Use the provided Chamfer distance function as a loss to deform the source towards the target. Optimize the position of the vertices in the source mesh by torch (you can just use the Adam optimizer). Show the result and describe what has happened in the deformation process by language.

   (b) Curvature and normal-based loss: We observe that Chamfer loss alone is not able to deform the source mesh properly, and additional loss needs to be added to regularize the process. Here we introduce a simple idea that would work for the provided instances by matching $\{\Delta p_i\}$:

   - First, we compute the Δp_i for each vertex of the source and the target meshes using 2(b).
   - Then, when we compute Chamfer-loss as in 3(a), we actually know about the correspondences across the source and target mesh vertex sets. For each pair of correspondences found in computing the Chamfer loss, we can use the L_2-norm square difference between the corresponding Δp_is' as the loss. Note that this loss computation can be bidirectional (from source to target and from target to source).

   Implement this loss in pytorch and combine it with the Chamfer loss to deform the source shape. Show your final results and deformation process (*e.g.*, the deformed mesh at every 100 steps).

   Note: `source.obj` and `target.obj` have similar amount of vertices, so the ΔA for vertices in the two meshes do not differ much, and matching Δp_is' corresponds to match mean curvatures and normal direction. If there is significant difference in vertex numbers, we need to compensate for the effect caused by significantly different ΔA in the two meshes.

Problem 2[10pt] In this problem, you need to implement the ICP algorithm by yourself to align two point clouds (`banana.pose0.ply` and `banana.pose1.ply`). Please report the rotation and translation error against the ground-truth `banana.pose.txt`.

Problem 3[10pt]

In this problem, you will use ICP to predict object poses in a dataset. The training data (`training_data.zip`) and testing data (`testing_data_pose.zip`) are provided here. You can download the data or directly use it in Google Colab by mounting the folder. You can use any third-party implementation of ICP, like Open3D in this problem.

Please read the following instructions carefully and ask on Piazza immediately if you have questions.

1. The task is to predict 6D poses for all the objects of interest in the scene given the image and the depth map. There are 79 different object classes in total. In each scene, there is only one instance for each object class. Across different scenes, instances of an object class might be scaled differently (in fact, only a discrete range, e.g. 0.5 or 1.0). This scale is provided in both training and testing datasets.

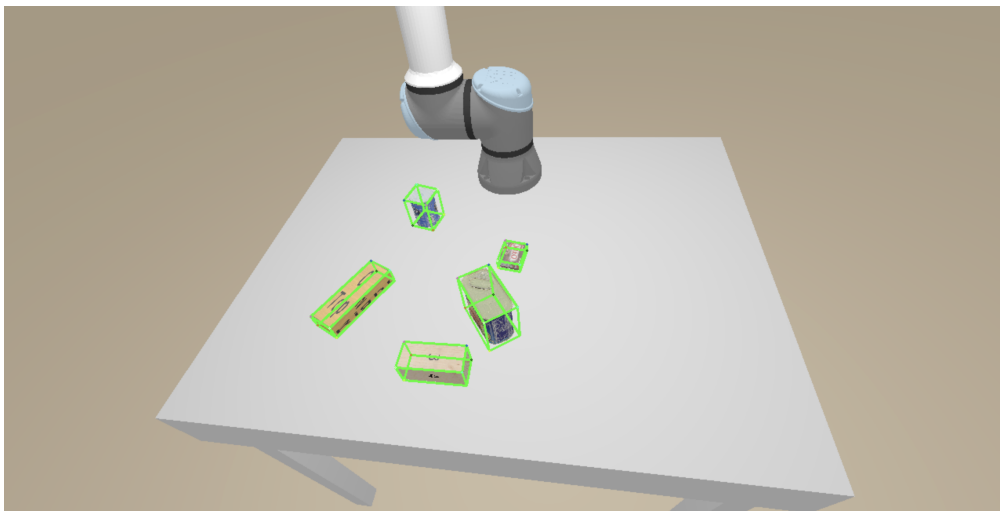

Figure 2: Example of the dataset for pose estimation

2. Here are some notations you might need to know before reading the rest.
   - NUM_OBJECTS: total number of objects we have. It is 79 in this assignment.
   - object_id: each object class is assigned with a unique id, decided by its order in objects.csv (explained next).
3. Data description. In training data, there are 2 folders: v2.2 and splits/v2, and there is a file object_v1.csv. The v2.2 folder contains the training data and splits/v2 is a pre-made train/validation split. You are allowed to create your own train/validate split and ignore splits/v2 completely. In splits/v2/train.txt and splits/v2/val.txt, each line is in the format "{level}-{scene}-{variant}". You can use it to find corresponding data files in the v2.2 data folder. Each variant with the same {level}-{scene} uses the same set of object classes with different poses. Here we give a detailed description of all files.
   - **objects.csv**. It provides metadata for the object meshes used to generate the scenes. The important fields are:
     - location: described where to find the meshes in models.zip.
     - geometric_symmetry: describe what symmetrical properties this object has. For example "z2|x2" means this object has a 2 fold symmetry around z axis and a 2 fold symmetry around x axis (which also implies a 2 fold symmetry around y axis). You can see Wikipedia entry Rotational_symmetry for details. "no" means this object has no symmetry; "zinf" means this object has an infinite-fold symmetry around z (e.g. cylinder). The symmetry properties are considered in our evaluation metric (e.g., for a cube, there are 24 rotations that will result in 0 error in evaluation).
     - visual_symmetry: similar to geometric symmetry, but considers object texture. Our evaluation metric will NOT consider visual symmetry, but it is an interesting research topic.
   - **{level}-{scene}-{variant}_color_kinect.png**: an RGB image (1280×720) captured from a camera containing the target objects.
   - **{level}-{scene}-{variant}_depth_kinect.png**: a depth image (1280 × 720) captured from a camera containing the target objects. The depth is in the unit of millimeter. You need to convert it into meter.

- **{level}-{scene}-{variant}_label_kinect.png**: a segmentation image (1280×720) captured from a camera containing the target objects. The segmentation ids for objects are from 0 to 78. Other ids (79, 80, 81) stand for the table, ground and robot.
- **{level}-{scene}-{variant}_meta.pkl**: meta information like camera parameters, object names, gorund-truth poses, etc.
  - poses_world (list): The length is NUM_OBJECTS; a pose is a 4x4 transformation matrix (rotation and translation) for each object in the world frame, or None for non-existing objects.
  - extents (list): The length is NUM_OBJECTS; an extent is a (3,) array, representing the size of each object in its canonical frame (without scaling), or None for non-existing objects. The order is xyz.
  - scales (list): The length is NUM_OBJECTS; a scale is a (3,) array, representing the scale of each object, or None for non-existing objects.
  - object_ids (list): the object ids of interest.
  - object_names (list): the object names of interest.
  - extrinsic: 4x4 transformation matrix, world $\rightarrow$ viewer(opencv)
  - intrinsic: 3x3 matrix, viewer (opencv) $\rightarrow$ image
- **models.zip**: meshes (with textures) for all objects. The object pose in the dataset is the transformation matrix relative to the canonical pose defined in those meshes.

We provide a Jupyter notebook to show how to use the data and also some handy scripts for visualization. For testing data, the only difference from training data is that there are no splits and the ground truth poses are not provided in the metadata.

4. Goal. Your goal is to predict the poses of objects in the testing data. The segmentation masks for objects are given so you do not need to solve the detection problem. You can simply segment out the point cloud for each object and use ICP to solve the problem.

5. Evaluation metrics. Your assignment is graded based on the pose accuracy score: success if rotation error < 5 degree and translation error < 1 cm.

   Your score is computed as the higher of absolute and relative score.

   Absolute score

   - 5 points: You method achieves higher than 80% accuracy on the overall scoreboard.
   - 4 points: You method achieves higher than 60% accuracy on the overall scoreboard.
   - 3 points: You method achieves higher than 40% accuracy on the overall scoreboard.
   - 2 points: You method achieves higher than 20% accuracy on the overall scoreboard.
   - 1 points: You make a submission.

   Relative score

   - 5 points: rank 1-5.
   - 4 points: rank 6-15.
   - 3 points: rank 16-20.
   - 2 points: rank 21-25.

- 1 points: the rest.

6. Submission. The submission has 3 parts

   - A short report describing your method and results (submitted to Gradesope). It should at least include two sections about "method" and "experiments". For example, the method section can include: 1) how the initial pose is obtained for ICP; 2) the hyperparameters of ICP. The experiments section can include: 1) quantitative results (pose accuracy) on a validation set and testing data; 2) qualitative results (visualization) for some success and failure cases. 1 page should be enough, but you may write more if you have interesting findings. This report should be combined with the solution of problem 1 into a single PDF. You can use a template from academic conferences like CVPR.
   - A .zip archive containing your code (submitted to Gradesope).
   - You need to submit your results on testing data (a JSON file) to our internal benchmark.

     The evaluation of a submission will take about 60s. Please be patient and avoid trying to submit multiple times.

     The format of submission is as follows:

     – The json is a dict whose keys are {level}-{scene}-{variant} for each level. The length of this dict must be the same as the number of test cases.
     – The value of each dict is another dict with a single key "poses_world". The value for "poses_world" is a list of 79 elements. Each element is either a None or a 2D list of size (4×4), representing the pose matrix (transformation from object frame to world frame) of the object. For example, if the data contains objects with ID 4 and 5, then the list should have 77 None values and 2 matrices at position 4 and 5.

     A trivial baseline "baseline.json" is provided. You can submit it to have fun. We also provide "benchmark_utils" to evaluate results on a validation set locally. You are free to choose your user name, however, you cannot use another name after the initial submission.

7. Hints:

   - First of all, please get familiar with data. It might be more convenient if you preprocess data, e.g., cropping and saving point clouds of objects from images.
   - The output poses should be in the world frame. However, it might be easier to estimate poses in the camera frame first and then convert them to the world frame, since you might crop point clouds in the camera frame during pre-processing.
   - ICP will align the source point cloud to the target one. Note that it is not commutable for source and target. The point clouds to align are complete point clouds sampled from meshes and partial point clouds observed from cameras.
   - The initial pose estimation can make large difference for ICP.

Problem 4 (Course Feedback) [1 pt]

1. How many hours did you spend on this homework?
2. How many hours did you spend on the course each week?
3. Do you have any course related feedback?