计算机视觉导论大作业（23-24 学年）

整体说明

大作业**占总评 30 分**。学生需要以小组形式（**最多两人一组**），自行选择完成以下五题中的**一题**。在**冬学期第 7、8 周**进行现场汇报展示。汇报的内容包括对方案的阐述，对问题的创新探索和思考，自己拍摄的 demo 展示，并进行问答等。在**冬学期第 8 周**需要上交一份简短项目报告。在**冬学期第 8 周实验课**，我们会对代码工程进行检查，确保不出现抄袭等违反学术诚信的现象。

评分标准

1. **工作量（总分 14 分）：**能实现基本要求得 10 分，完成进阶要求最多得 14 分。
2. **创新（总分 6 分）：**调研了一些新的方法，提出了一些新的思路占 2 分，实现了一些新的方法，让作业的整体效果更好占 2 分，和其他人的项目选题或实现方式有显著区别占 2 分。
3. **展示（总分 10 分）：**口头汇报质量与回答问题正确性占 6 分，自己采集数据完成的 DEMO 及其效果占 4 分。

FAQ

1. 完成基础方案时，可以和每题对应的助教讨论。
2. 如果需要显卡等计算资源，可以咨询林浩通助教。建议尽早开始作业，实验室显卡资源有限，去年最后一到两周里，因为过度拥挤，学生几乎无法使用实验室显卡资源。
3. 展示顺序将在冬学期第六周随机抽签。
4. 请及时关注课程群。

选题（五选一）

一、光心不一致时的全景图拼接

问题定义:

该问题探索在多张图片的光心（即相机中心）不一致的情况下，如何有效进行全景图的拼接。当光心不一致时，图片间的几何关系会变得更为复杂，导致基于单应性变换的模型误差增大，在拼接处的效果不佳。

基础解决方案:

1. 首先自行采集图片，探索在光心不一致的情况下，传统的基于课程学习的图像拼接方案会遇到的问题。
2. 利用课程学习的 Structure from Motion 和 Multi View Stereo 技术来获取图片的相对姿态、场景的稠密点云和图片的深度信息。
3. 建立具有相同光心视角的目标图画布，将重建的稠密点云投影到该视角，得到目标画布每个像素的颜色，这个技术叫做 depth-warping，可以参考 kornia 的函数（见参考）。
4. 可以利用“单应性变换模型得到的全景图”补全空洞部分，也可以利用图像补全的方法做补全。最终能够得到一个光线、几何一致的全景图。

额外提升方案:

- 结合 Structure from Motion 和神经辐射场（Neural Radiance Fields, NeRF）技术，根据多视角图片重建场景并渲染出全景图。这一过程可以利用像 Instant-NGP 这样的快速 NeRF 实现。
- 进一步探讨如何更高效地结合 SfM 和 NeRF，以提高场景重建的准确性和渲染质量。
- 研究如何提升渲染速度，实现在多目视频序列上快速拼接全景图（可参考 ENerf 的方法）。

讨论助教：沈泽弘

最新参考链接和论文:

[1] “COLMAP: A General-Purpose Structure-from-Motion and Multi-View Stereo”：这是一个关于 SfM 的基础工具和算法的论文，帮助理解和实现图片姿态估计。

[2] https://kornia.readthedocs.io/en/v0.1.4/geometry.warp.html：Kornia 的 Depth-warpping 函数实现。

[3] “https://github.com/NVlabs/instant-ngp”：InstantNGP 的实现方案。
[4] “Efficient Neural Radiance Fields for Interactive Free-viewpoint Video”：这篇论文提出了一种高效的 NeRF 实现，可以作为提高渲染速度的参考。
[5] http://vision.ia.ac.cn/zh/teaching/CV2015/Course5.pdf 全景图拼接技术讲解。

二、重建校园某区域，基于图像对自身做定位

问题定义：

该问题包含两个部分：重建与定位。重建指的是基于一组多视角图像，恢复出场景的三维结构（点云与相机位姿）。定位则是基于场景的重建，估计同一场景新采集图像的六自由度位姿。这里新采集图像可以是不同视角，以及不同时间拍摄的。重建与定位在 VR 导航等场景有着广泛的应用。

基础解决方案：

1. 重建数据采集：在校园的某区域 a，用手机采集多视角图像（最好是白天）。
2. 三维重建：利用课程学习的 Structure from Motion (SfM)方法恢复多视角图像位姿态与场景点云，构建场景地图
3. 定位数据采集：在区域 a，采集多张不同视角图像用于定位（最好是傍晚或者夜晚）
4. 视觉定位：将每张待定位图像与多张重建图像匹配，进而通过重建图像与场景点云之间已知道的 2D-3D 关系，得到待定位图像与场景点云之间的对应匹配。最终利用 2D-3D 匹配求解 PnP，恢复待定位图像的位姿。
5. 可视化：将三维重建得到的场景 SfM 点云利用估计的测试图像位姿投影到测试图像上可视化，通过投影点云与实际图像的吻合程度来观察定位的准确性。

额外提升方案：

- 针对性能：1.1 利用多种不同的匹配算法进行重建与定位，对他们的优劣性有直观的理解。（参考文献 1 以及对应 repo）1.2 采用其他定位流程，例如针对场景训练一个特定网络直接回归测试图像对应 3D 匹配。效果不一定要比基础解决方案好，进行实际效果比较详细分析清楚即可（参考文献 2）1.3 采集较难的数据进行方法测试：重建图像白天采集， 定位数据夜晚采集，并且之间有大角度视角偏差，测试方法的鲁棒性。
- 针对效率：优化方法流程，提升每帧定位图像的定位速度。（参考文献 3，4）

- 可视化与实际应用：采集一段视频进行定位，然后基于定位结果制作一段 AR demo：例如在场景里放置一个物体 3D 模型，然后将物体用估计的位姿渲染到各帧合成视频。

讨论助教：贺星毅

最新参考链接与论文:

[1] From Coarse to Fine: Robust Hierarchical Localization at Large Scale https://github.com/cvg/Hierarchical-Localization

[2] NeuMap: Neural Coordinate Mapping by Auto-Transdecoder for Camera Localization

[3] OnePose: One-Shot Object Pose Estimation without CAD Models

[4] OnePose++: Keypoint-Free One-Shot Object Pose Estimation without CAD Models

三、从多视角视频还原车辆、行人轨迹

问题定义:

对校园一个区域拍摄一段时间的多视角视频，从多视角视频中检测行人和车辆，并恢复它们的三维轨迹，在三维空间中做可视化。

基础解决方案:

1. 首先用多个固定位置的手机采集多视角视频，需要包括运动的车辆和行人。
2. 标定相机的内参和外参。
3. 利用 2D 视频做车辆和行人的检测、分割、姿态估计。
4. 利用三角化方法，得到这些车辆、行人的三维姿态和运动轨迹。
5. 能够做可视化（可以参考 wis3d）。

额外提升方案:

- 提升基础解决方案的准确度和效率，比如可以得到更加平滑的三维可视化。
- 能在同一个场景中（已经重建好的），只使用一个相机做到上述效果。

讨论助教：沈泽弘

最新参考链接和论文:

[1] “https://github.com/zju3dv/Wis3D”：这是一个三维信息可视化的工具，类似 tensorboard。

[2] https://pjreddie.com/darknet/yolo/：二维物体、人体检测

[3] “COLMAP: A General-Purpose Structure-from-Motion and Multi-View Stereo”：这是一个关于 SfM 的基础工具和算法的论文，帮助理解和实现图片姿态估计。

四、基于多视角图片重建小动物/人

问题定义：

通过手机/相机采集一段关于某个人/或者某个小动物的视频，我们希望能够重建出拍摄对象的三维表面。

基础解决方案：

1. 首先需要对得到的视频进行图片的提取，例如我们可以每隔 0.5 秒提取一张图片。得到图片后，我们首先需要对图片进行相机标定，然后对相机进行标定。我们使用实验 5 中学到的 colmap 对图片进行标定即可。
2. 调研算法：调研传统的 MVS 算法以及最近基于神经表示的重建方案，尝试回答这两条路线的各自的技术路线是怎样的？各自有什么优缺点？目前的发展状况如何，例如这两条路线各自尝试解决什么样的问题？下面各列了一个具体方案，可以作为你的出发点。
   (1) 传统的 MVS 算法：这里指的是通过 PatchMatch 等算法重建每个视图的深度图，然后通过点云融合，以及泊松重建完成对 surface 的重建。colmap 已经完成了对上述算法的集成，我们直接调用即可。我们根据 colmap 的文档即可完成。https://colmap.github.io/tutorial.html#dense-reconstruction 分别对应以下三步：colmap patch_match_stereo、colmap stereo_fusion、colmap poisson_meshe
   (2) 基于神经表示的重建方案：NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction。我们参考官方 code 给的指南即可完成：https://github.com/Totoro97/NeuS

额外提升方案:

- 数据提升（1）：如果小动物一直在动，人发生了眨眼可以考虑重建一个动态场景，例如拍摄主体是小猫，小狗，他们往往会运动

  [1] Nerfies: Deformable Neural Radiance Fields

  [2] Unbiased 4D: Monocular 4D Reconstruction with a Neural Deformation Model

- 数据提升（2）：基于神经表示的重建方法是否可以重建一个超大场景，例如拍摄主体是蒙民伟楼，曹光彪楼，他们往往有着非常大的尺寸。
  [3] Neuralangelo: High-Fidelity Neural Surface Reconstruction.
- 方法提升（1）：NeuS 的重建速度很慢，如何提升？
  [4] nstant Neural Graphics Primitives
  [5] NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction
- 方法提升（2）：NeuS 的重建质量不高，尤其是反光/无纹理场景等。自己多尝试几组数据，尝试发现 NeuS 会出现明显 artifacts 的场景，尝试进行原因的分析，以及根据所分析的原因和对 NeuS 方法的理解，提出对 NeuS 方法的改进，实验验证你的改进。
- 方法提升（3）：改进无纹理问题
  [6] Neural 3D Scene Reconstruction with the Manhattan-world Assumption
  [7] MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction

讨论助教：林浩通

建议大家优先尝试简单，静止的物体（人脸/动物）。完成提高部分的任意一项(运动/超大场景/重建速度/重建质量)即可得到提升分。

拍摄物体视频时，可以在物体下方垫若干张作文纸，或者其他由很多有纹理的书籍报纸，这样有利于标定。拍摄物体视频时，应力求视角完整。

五、参考一段网络舞蹈视频，合成自己跳舞的视频

问题定义：

找到一段网络上单人跳舞的视频（5~10 秒），生成自己跳相同舞蹈的视频。

基础解决方案：

方案 1（基于 2D）：

1. 从网络上选择一个单人舞蹈视频作为参考。
2. 录制自己执行各种动作的视频，这些动作不能与参考视频中的舞蹈完全一致。
3. 使用姿态估计技术（如 OpenPose [1]）分析两个视频，提取每一帧的人体姿态。在自己的视频中找到与参考视频中舞蹈动作最相似的帧。
4. 将匹配到的帧按照参考舞蹈的顺序重新排列，生成一个新的视频。确保新视频中的动作流畅和自然。

方案 2（基于 3D）：

1. 1.从网络上选择一个单人舞蹈视频作为参考。
2. 使用一段从多个角度拍摄自己的视频，利用基于视频的三维人体重建[2]来从中重建出自己带贴图的三维模型（代码见[3]）。
3. 使用 3D 姿态估计技术（如 HMR [4]）分析网络舞蹈视频，提取出舞蹈动作的姿态。
4. 将提取出的舞蹈姿态应用到自己的三维人体模型上，这样模型就能模仿视频中的舞蹈动作。
5. 将动作驱动的三维模型合成到一个新的视频中，展示模型执行舞蹈动作。

额外提升方案：

- 基于 2D 的方法如何更好地保证时序一致性？可以尝试用自己的视频数据训练一个从人体姿态到视频的 2D GAN 来生成 [5]。
- 基于 2D 的方法能不能借助一些现有的生成模型来实现更好的效果？比如利用 stable diffusion [6, 7]。
- 基于 3D 的方法能不能渲染得到更真实的图片？比如用基于 NeRF 的表示 [8, 9]。

讨论助教：皮怀瑾

参考文献：

[1] OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

[2] Video Based Reconstruction of 3D People Models

[3] https://github.com/thmoa/videoavatars

[4] End-to-end Recovery of Human Shape and Pose

[5] Everybody Dance Now

[6] DISCO: Disentangled Control for Realistic Human Dance Generation

[7] DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion

[8] Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies

[9] HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video