CV Project Storyline - Group 13

Monocular Spacecraft Pose Estimation with Vision Transformers

Monocular Spacecraft Pose Estimation is the problem of finding the relative position and orientation of a target spacecraft, with respect to the camera reference mounted on a chaser (or servicer) spacecraft

- **Why Interesting?**
  - Many tasks like crew transfers and resupply missions require a chaser spacecraft to align and physically join another target spacecraft, an operation called **docking**. An accurate real-time estimation of the target's pose ensures that the docking ports of the chaser and target are correctly aligned.
  - To be **real-time**, the target's current pose should be estimated by the current camera frame only, avoiding the computational burden of processing long image sequences.
  - The chaser needs to estimate the target's pose using **onboard** sensors and computers only, since connection with the Earth is often either delayed, noisy, or unavailable.

- **How done now?**
  - **Model-based** approaches: rely on pre-built models of target vehicles to estimate their pose. Have to perform manual feature extraction/matching, **difficult to adapt** to new scenarios.
  - **Hybrid modular** approaches: many steps involving object detection/localization, keypoint regression, and pose computation, of which the first two are usually DL-based, and the last is usually a classical algorithm. Can be **complex to implement** and optimize.
  - **End-to-end** approaches: These methods directly learn to estimate the pose using a single, end-to-end neural model, often CNN-based. Usually simpler than hybrid approaches, but can be **data-hungry**.

- **What is missing?**
  - Many datasets are synthetic renders as gathering real satellite images is hard, but models trained on rendered data struggle to match the same performance when tested on real operational scenarios -> **Domain Shift**
  - CNNs usually achieve good in-distribution generalization, but may not generalize as well under domain shift, where test samples are **out-distribution**.
  - The CNN-based model by UniAdelaide which won the ESA pose-estimation challenge yields an approximately **40 times worse** performance on the real data compared to the synthetic one, according to the ESA accuracy metric.
  - Current state-of-the-art models are **too heavy** –hundreds of millions of parameters-- to run real-time inference on the small ARM processors typically used in the space domain.

- **Proposed Solution**
  - Reproducing the results of **Mobile-URSONet**, an end-to-end model based on the MobileNet CNN backbone with only 7.4M parameters, small enough to run real-time inference on existing ARM processors used in the space domain (e.g. Qualcomm ARM A-72). The reproduction is performed on the synthetic data of the SPEED dataset, and later its ability to generalize is tested on realistic lab data from SPEED+ .
  - Modifying the network by changing the CNN backbone with a **ViT** pretrained tiny model, which allows us to test the generalization power of ViTs while maintaining a comparable size of the network (6.8M parameters).
  - Experimenting with additional versions of the model based on **RepViT** and a combination of ViT and MobileNet backbones.
  - Training the modified networks on SPEED for the same amount of epochs of Mobile-URSONet, and comparing their ability to generalize on the real data from SPEED+.

- **Experimental Questions**
  - To what degree do Vision Transformers properties improve generalization in handling domain shift when trained on synthetic data and tested on real-world samples compared to CNNs?
  - Can this enhanced generalization be achieved without increasing the number of parameters through the proposed model modifications?