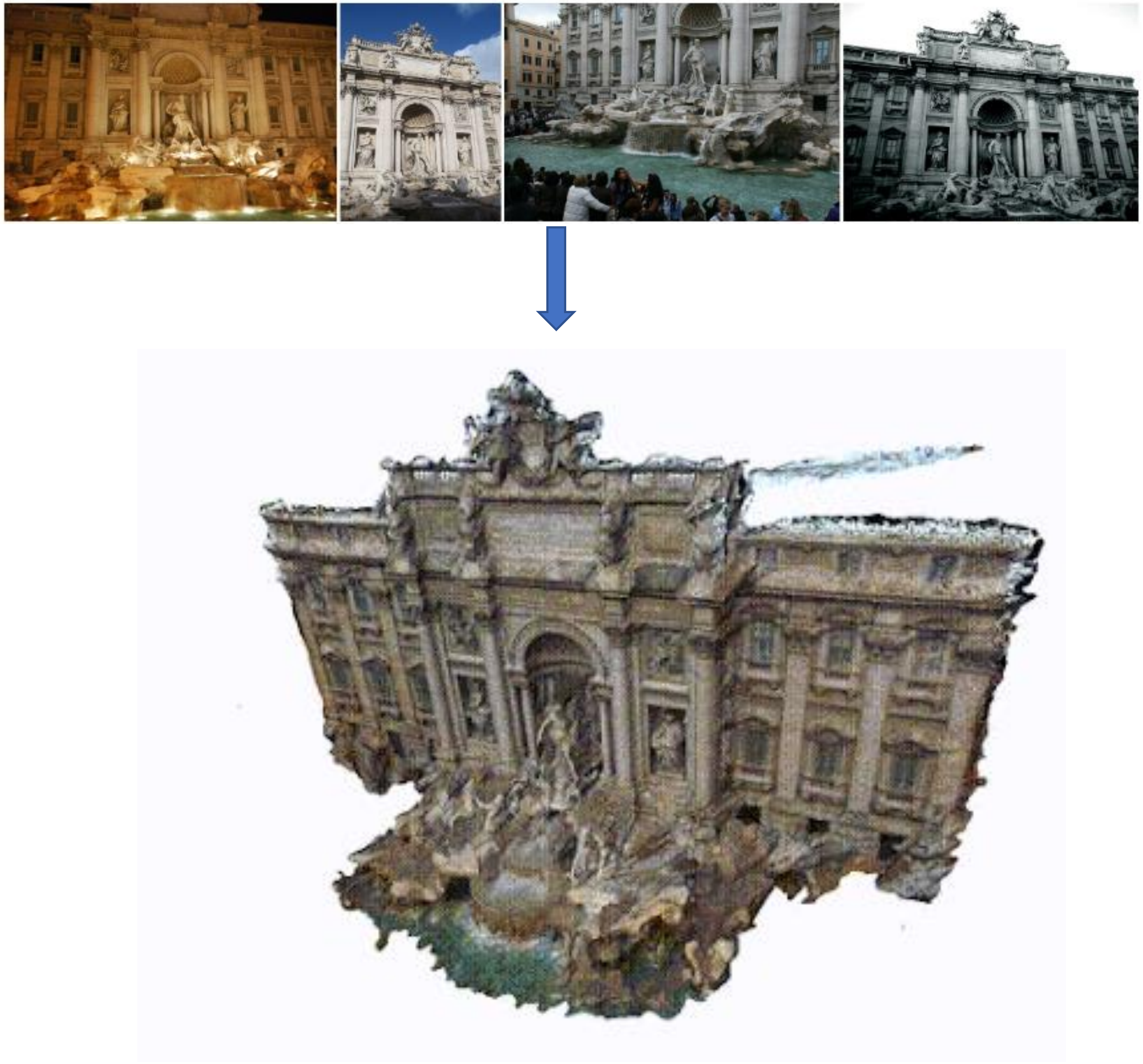
Image Matching Challenge

A data augmentation and ensembling approach

Background and Introduction



Question:
- What if machine learning could help better capture the richness of the world using the vast amounts of unstructured collections of images freely available on the internet?

Given:
- two images
- intrinsic camera properties

Find:
- >= 8 pairs of matching points between the two images
- A fundamental matrix F describing the relative camera pose
- Ultimately: depth of points in the image (3D reconstruction)

Our innovation:
- We find that none of existing work tackles the problem of ensembling and data augmentation in this context. We propose a general framework for data augmentation as well as creating ensembles.

Problem Statement

Problem definition

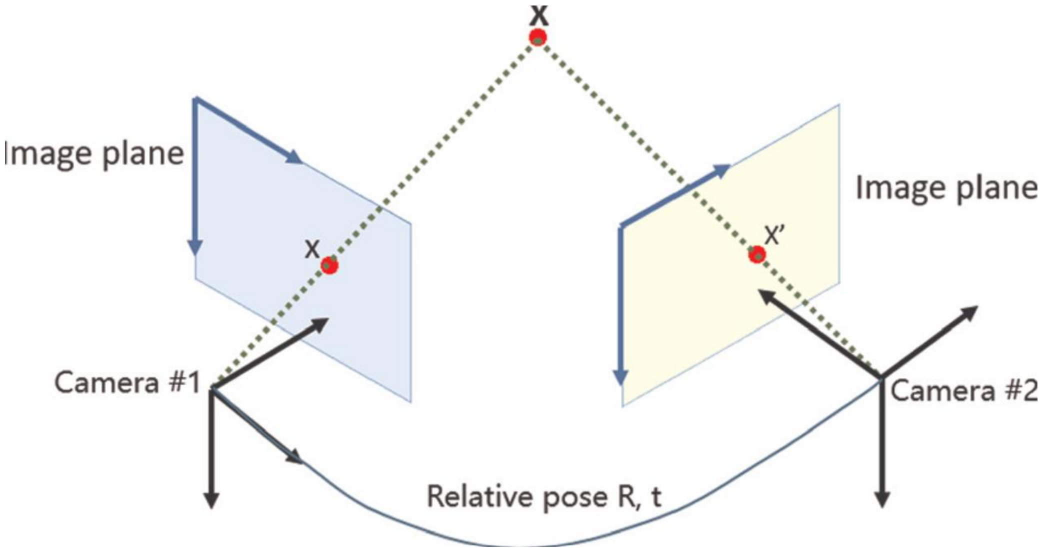
Given a pair of images that capture the same scene from two different cameras with unknown relative pose, we need to estimate the fundamental matrix, F.

The core task involved in the estimation of the fundamental matrix is image matching. The estimated F, coupled with knowledge of camera intrinsics, will be used to estimate relative pose between the cameras in a downstream task.

Evaluation Metric - mean Average Accuracy

Error between ground truth F and estimate $\hat{F}$ is quantified in terms of rotation and translation required to obtain F from $\hat{F}$. An estimate is accurate if these rotation and translation are within a specified threshold.

- Many choice of thresholds are specified
    - Rotation-threshold (in degrees) = np.linspace(1, 10, 10)
    - Translation-threshold (in metres) = np.geomspace(0.2, 5, 10)
- Average Accuracy on a scene is average, over threshold choices, of percentage of $\hat{F}$ that are accurate given the threshold choice.
- Mean Average accuracy of the model is the mean of average accuracies on the scenes.



Dataset & Data-augmentation

DATASET

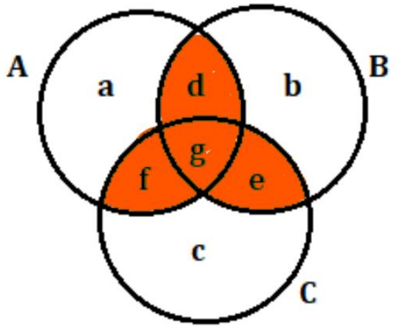
- **Train set:** 5720 images from across 16 landmarks.
  - For each image of a landmark, camera intrinsics, rotation matrix and the translation vector have been provided.
  - For each possible pair of images of the same landmark, an estimate of overlap between the images as well as the fundamental matrix have been provided.
- **Test set:** 10,000 images. It is hidden from the Kaggle competition contestants.

DATA AUGMENTATION

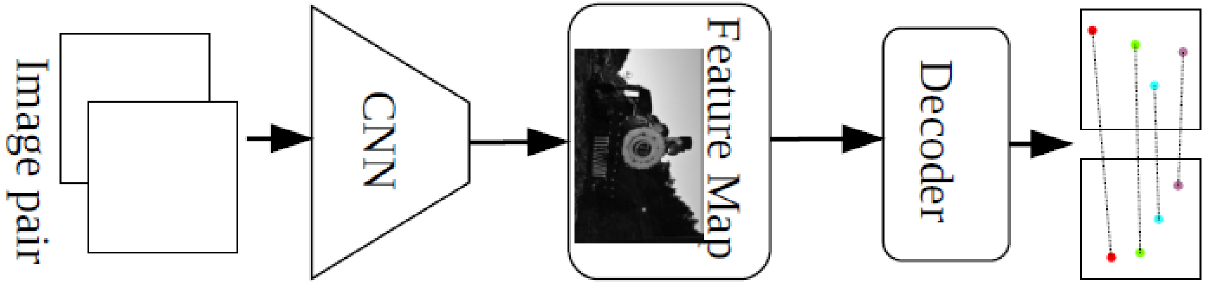
- **Consider:** a linear transformation A applied to 3D space (e.g. reflection across the yz-plane)
- **Find:** the new fundamental matrix. We prove it is $(A^{-1})' \, F \, A^{-1}$.
- **Empirical verification:** use this formula to generate new labeled examples. Check the performance of best-performing models on augmented data

Methods and Ensemble Framework

- Create an ensemble from F estimates.
  - Simple average
  - Dynamic Quality-of-estimate weighted average. We use “count of inliers that went into estimating F” as a measure of estimation quality (we are trying to mimic inverse variance weighting to reduce the variance of estimate).
- Create an ensemble of matching points
  - Pool all points
  - Pool overlapping points (overlap indicates higher quality-of-estimate)



- We used three pretrained models as inputs to ensemble.
  - All models uses CNN to extract a vector features for each pixel
  - LoFtr[1] uses interleaved self and cross attention to capture global context in the cost volume (CNNs have limit receptive fields and hence fail to capture global context well).
  - DKDGM[3] uses Geometric Process regression to capture global correlation.
  - ASLFeat[2] attempt to capture selective and accurate keypoints by adding geometric constraints to keypoint extraction (i.e. the CNN).



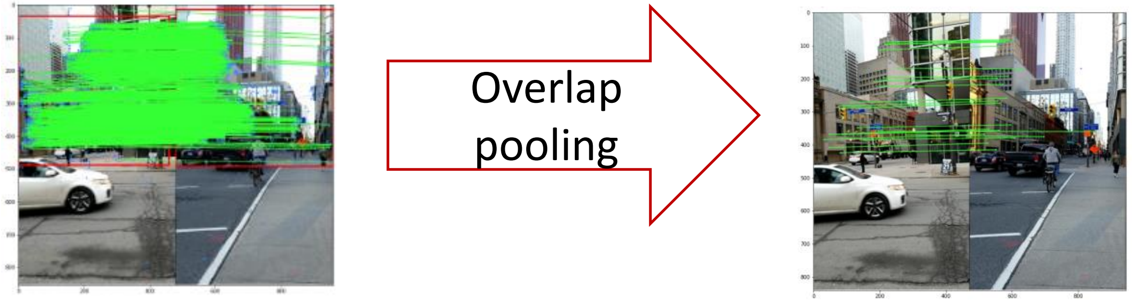
Experiment and Analysis

- Data Augmentation
  - Evaluation: we transform the image pairs and then inverse-transform our estimate of F. We compare this estimate with the F estimated on normal images. We look at the operator norm of the difference matrix.
  - Mean error: 0.0230 relative error
  - Maximum error: 0.130 relative error
- Finetuned LoFtr
- Ensembling

Model	mAA (Test)
KeyAffHardNet (hand-crafted features) pretrained	0.523
LoFtr pretrained	0.726
ASLFeat pretrained	0.673
DKDGM pretrained	0.668
LoFtr finetuned	0.721
F Simple-avg	0.682
F Performance wted	0.655
Pool matched points	0.787
Overlapping matched points	0.468

Conclusion and Future work

- LoFtr provides the best mAA for an individual model.
  - But it detects fewer matched pairs than DKDGM.
  - ALSFeat detects fewest matched pairs but has a performance that beats DKDGM.
  - This means that quality of keypoints detected matters quite a bit and may explain why hand-crafted features do well.
- Creating an ensemble of matched points shows promise.
  - Pooling matched points identified by different models leads to an improvement in mAA.
  - Overlapped pooling leads to worsening of performance. Looking for overlaps greatly reduces count of matched points.
- Aggregating F estimates doesn’t improve performance much.
- Future work:
  - Explore other measures of estimate-quality to create the “inverse variance weighted” ensembles of F and matched points.
  - A larger list of input models, with simple pooling of matched points, may help push the state-of-the-art on this task.



References

1. **Loftr** - Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. CoRR, abs/2104.00680, 2021
2. **ASLFeat** - Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and lo- calization. CoRR, abs/2003.10071, 2020
3. **DKDGM** - Johan Edstedt, M˚arten Wadenb¨ack, and Michael Fels- berg. Deep kernelized dense geometric matching. CoRR, abs/2202.00667, 2022.

.