# LocalAIze

Céline Gauchey
David Palumbo
Joel Salzman

# Project Objective

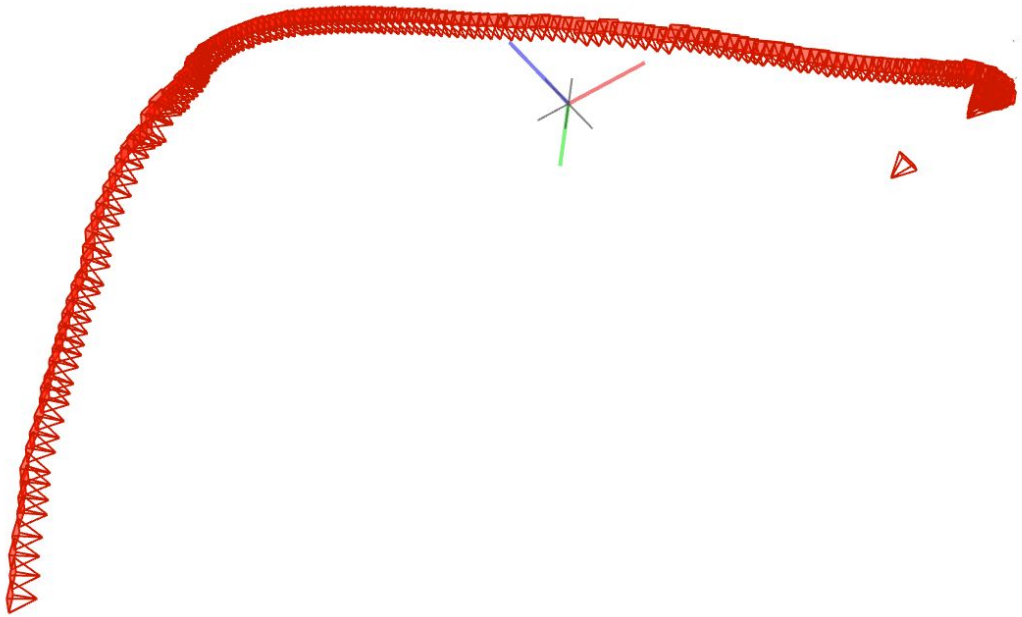
- Predict camera pose from random frames in a video
  - Input: Images
  - Output: Location
- Harness the power of **VSLAM** (Visual Simultaneous **Localization** and Mapping) to predict the location of an image taken within a specific room: Pupin 428

# Data Input

# Training Data

- Extracted 3927 frames from video
  - Cropped to square
  - Compressed to 512x512
  - Reduced RGB colorspace to Grayscale
- Structure from Motion to generate ground truth camera poses
  - COLMAP on GPU (2060 Super)
- Multiple separate models
  - We couldn’t register every frame in the same model
  - BUT doing it this way generalizes to a larger class of disjoint localization tasks (so... it was totally on purpose...)

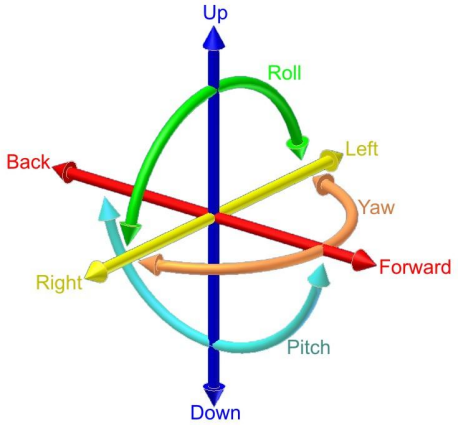
# SfM



- Each object is a camera
- Position is relative to centroid of segment
- Rotation is relative to first observation
- Scale is identical between segments since camera focal length is known

# Objective & Loss

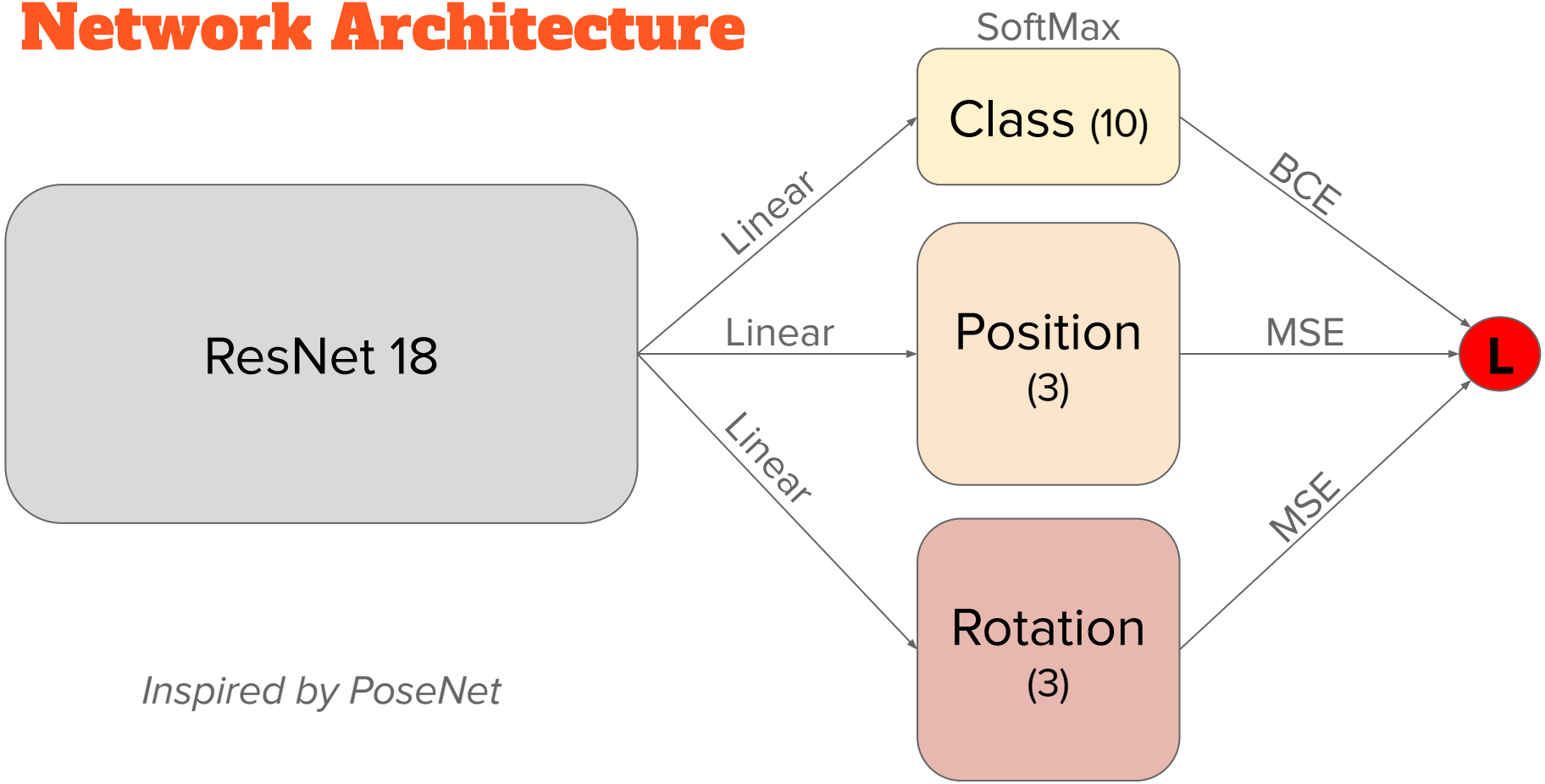
| | |
|---|---|
| c | Class (ballpark estimate of location; index of SfM model) |
| x | X coordinate of camera in local frame |
| y | Y coordinate of camera in local frame |
| z | Z coordinate of camera in local frame |
| p | Pitch (rotation about X axis) |
| r | Roll (rotation about Y axis) |
| y | Yaw (rotation about Z axis) |




https://en.wikipedia.org/wiki/Six_degrees_of_freedom

**Binary Cross-Entropy** Loss for class; **Mean Squared Error** Loss for 6DoF; ***α*** hyper

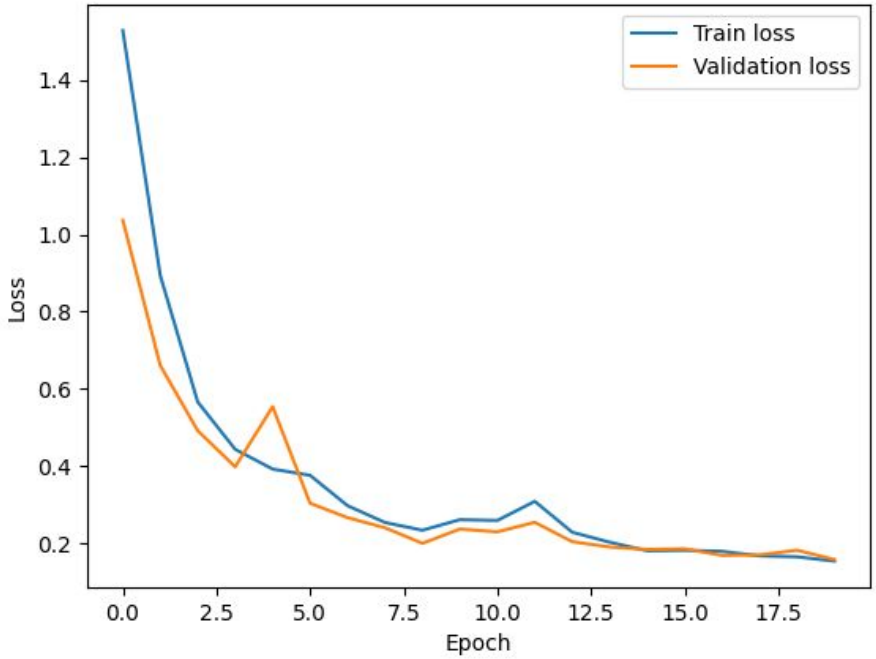
# Network Architecture




*Inspired by PoseNet*

# Results: Training

```
02:10:49    Epoch: 13 - train loss: 0.22813946593538334
02:10:49    Epoch: 13 - val loss:   0.20361114044984183

02:11:01    Epoch: 14 - train loss: 0.2023066628865982
02:11:01    Epoch: 14 - val loss:   0.1899414857228597

Updating best model, now found at epoch 14
02:11:12    Epoch: 15 - train loss: 0.18013706663249887
02:11:12    Epoch: 15 - val loss:   0.18354595369762844

Updating best model, now found at epoch 15
Saving best model weights at epoch 15
02:11:24    Epoch: 16 - train loss: 0.18124298948763107
02:11:24    Epoch: 16 - val loss:   0.1851540075408088

02:11:36    Epoch: 17 - train loss: 0.17863580928354172
02:11:36    Epoch: 17 - val loss:   0.1681344368391567

Updating best model, now found at epoch 17
02:11:47    Epoch: 18 - train loss: 0.1668879332531378
02:11:47    Epoch: 18 - val loss:   0.16944885585043165

02:11:59    Epoch: 19 - train loss: 0.16446445323336417
02:11:59    Epoch: 19 - val loss:   0.18153506186273363

02:12:10    Epoch: 20 - train loss: 0.15340091820646176
02:12:10    Epoch: 20 - val loss:   0.15732847981982762

Updating best model, now found at epoch 20
Saving best model weights at epoch 20
Training finished. Best validation loss of 0.15732847981982762 at epoch 20
```



# Results: Test Images

Predicted class: 1 - Actual class: 1

Predicted xyz ([-0.38865992 -0.5422972  -0.08911549])
Actual xyz ([-0.55068116 -0.4394174   0.01071588])

Predicted rpy ([ 1.4426252  -0.35953173 -1.3733114 ])
Actual rpy ([ 1.73502943 -1.26409687 -2.71521486])

Predicted class: 2 - Actual class: 2

Predicted xyz ([0.13467625 0.45333573 0.25124094])
Actual xyz ([-0.03780979  0.55761269  0.04036432])

Predicted rpy ([-0.7113406  0.7395303 -1.5051279])
Actual rpy ([-1.44847438 -0.37112511 -1.63266237])

Predicted class: 3 - Actual class: 3

Predicted xyz ([0.06528374 0.08230963 0.17236412])
Actual xyz ([0.1023216  0.12649211 0.00802225])

Predicted rpy ([-1.3210413  0.0551736 -0.555706 ])
Actual rpy ([-1.43965236  0.0797602  -0.65699823])

Predicted class: 0 - Actual class: 0

Predicted xyz ([-0.1836786  -0.29226592 -0.02850346])
Actual xyz ([-0.02705512 -0.07882347 -0.00632456])

Predicted rpy ([-1.2418957 -0.2940526 -2.1228151])
Actual rpy ([-1.26778938  0.1328009  -2.24772327])

Predicted class: 1 - Actual class: 0

Predicted xyz ([-0.26711962 -0.07887348  0.01788641])
Actual xyz ([-0.76756627 -0.32912194 -0.21573355])

Predicted rpy ([0.04543776 0.0460096  0.71685326])
Actual rpy ([-0.80339201  0.05432472 -6.37253114])

Predicted class: 7 - Actual class: 7

Predicted xyz ([-0.3713603   0.17623453  0.30363068])
Actual xyz ([-0.75274544 -0.19691871  0.15479946])

Predicted rpy ([ 3.565703   1.1255237 -2.8646848])
Actual rpy ([ 3.8695683   1.78146233 -3.66492545])

Predicted class: 1 - Actual class: 1

Predicted xyz ([-0.3971177  -0.08097175 -0.07701636])
Actual xyz ([-0.30702741  0.08269419 -0.13292921])

Predicted rpy ([ 0.64691657  0.02428344 -2.966156  ])
Actual rpy ([ 0.75237571 -0.19466621 -3.19146673])

Predicted class: 2 - Actual class: 2

Predicted xyz ([0.06829979 0.560788   0.13280231])
Actual xyz ([-0.11938004  0.22108825  0.02775515])

Predicted rpy ([-2.1284845  3.0878003 -2.0883293])
Actual rpy ([-1.94271456  2.32689928 -1.47385737])

Predicted class: 3 - Actual class: 3

Predicted xyz ([ 0.19646962 -0.17720933 -0.33185455])
Actual xyz ([ 0.32288196 -0.09715515 -0.11358771])

Predicted rpy ([ 1.613738   1.8200816  -0.20607328])
Actual rpy ([2.18363678 2.44343738 0.38250232])

Predicted class: 9 - Actual class: 9

Predicted xyz ([0.09582216 0.4965696  0.41501537])
Actual xyz ([-0.04420118  0.11017486 -0.09542109])

Predicted rpy ([ 0.55729395  1.562404   -1.7316667 ])
Actual rpy ([ 0.58593542  1.42400994 -0.13567736])

# Predictions - Visualized

Predicted class: 2 - Actual class: 2

Predicted xyz ([0.13467625 0.45333573 0.25124094])
Actual xyz ([-0.03780979  0.55761269  0.04036432])

Predicted rpy ([-0.7113406  0.7395303 -1.5051279])
Actual rpy ([-1.44847438 -0.37112511 -1.63266237])

# Parameter Tuning & Lessons

- Hyperparameters we played with:
  - Learning Rate
  - Batch Size
  - Patience
  - Epochs
- COLMAP is not as easy to use for large models as we thought
  - Global Bundle Adjustment requires CPU
- If we were to redo the project, we would have taken a video of a smaller section of the room, a smaller room, or a video of just one object

# Future Work (Possibly)

- Different ground truth
  - Single SfM model
  - Precise measurements
- More diverse coverage
  - Lighting
  - Movement
  - Other views (drone in the lecture hall??)
- Novel View Synthesis to validate pose prediction
- Transformers to leverage video structure

# Thank you!