2021 ICCV OCTOBER 11-17 VIRTUAL

ETH zürich

Microsoft

Pixel-Perfect Structure-from-Motion with Featuremetric Refinement

Philipp Lindenberger*

Paul-Edouard Sarlin*

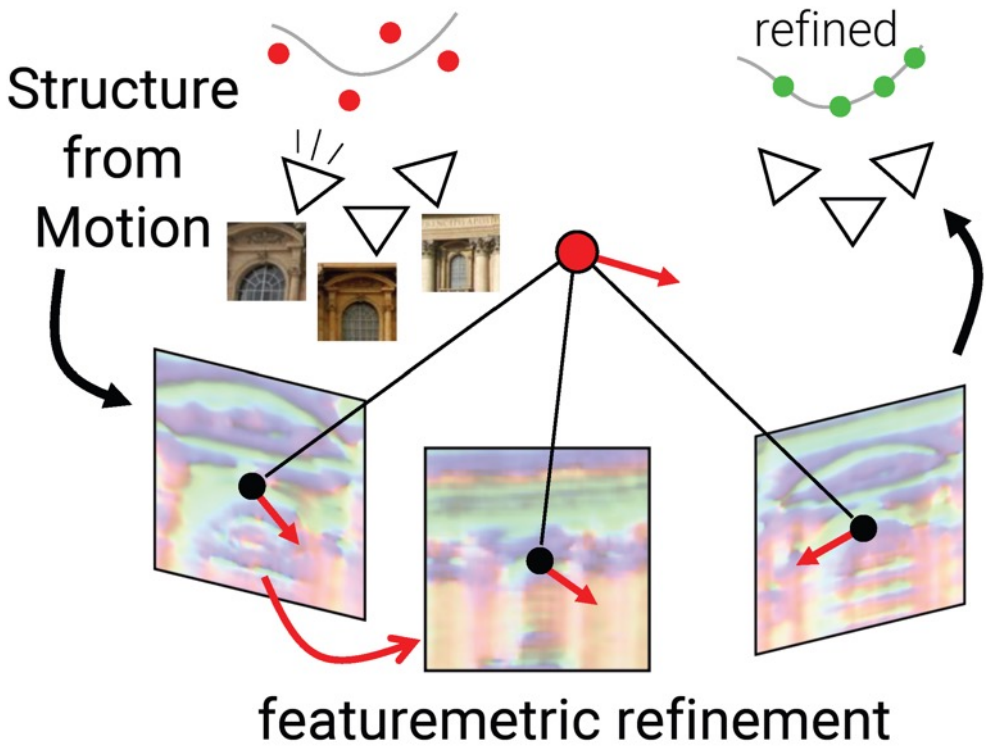
Viktor Larsson

Marc Pollefeys

ICCV 2021 Oral presentation

*equal contributions

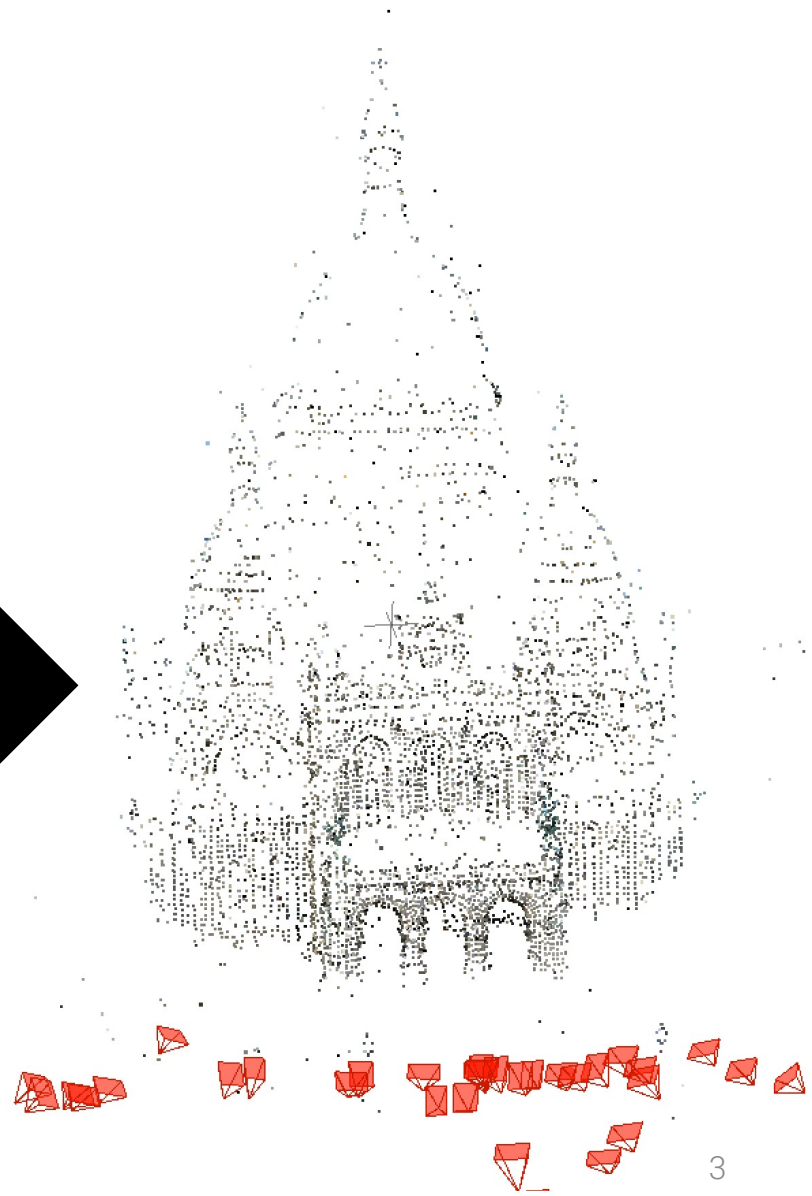
Pixel-Perfect Structure-from-Motion with Featuremetric Refinement



- Refine **sparse SfM**
  = 3D points + poses
- Direct **alignment of deep features**
- Makes visual **localization @ 1mm**
  3-10x more accurate
- **Dense but local** → scales to 1000s of images with only 20% overhead
- Robust to real-world challenges

Structure-from-Motion

SfM

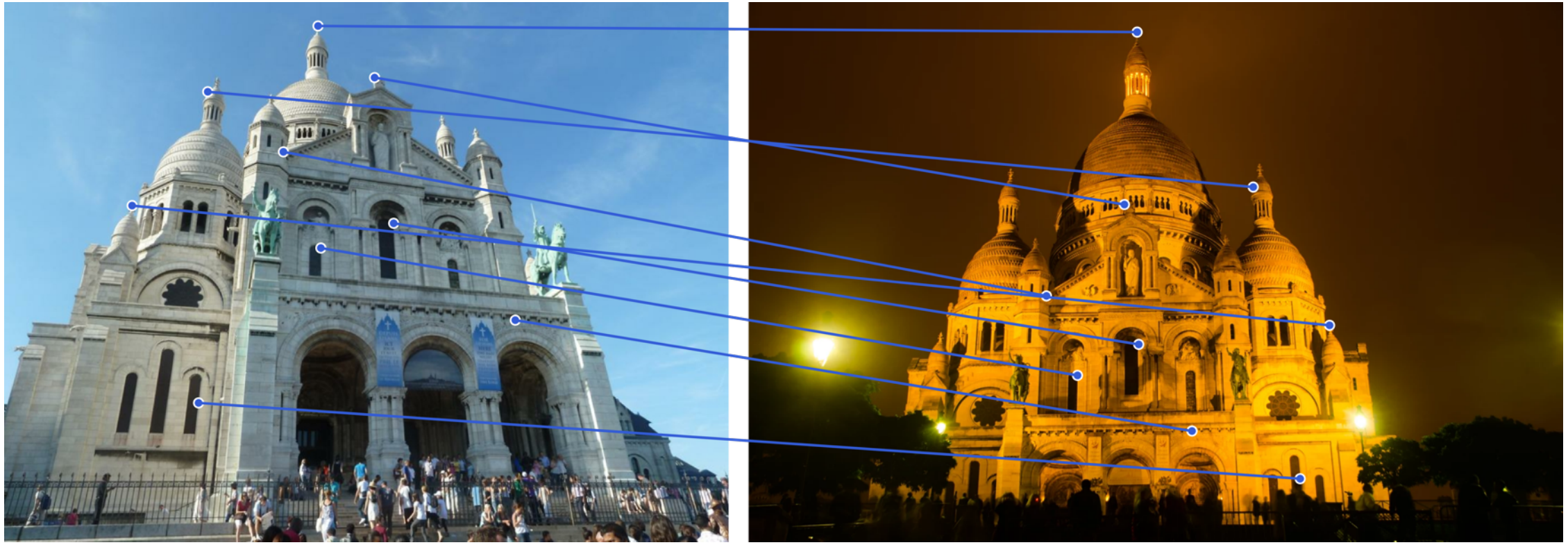


1. Feature Detection and Description

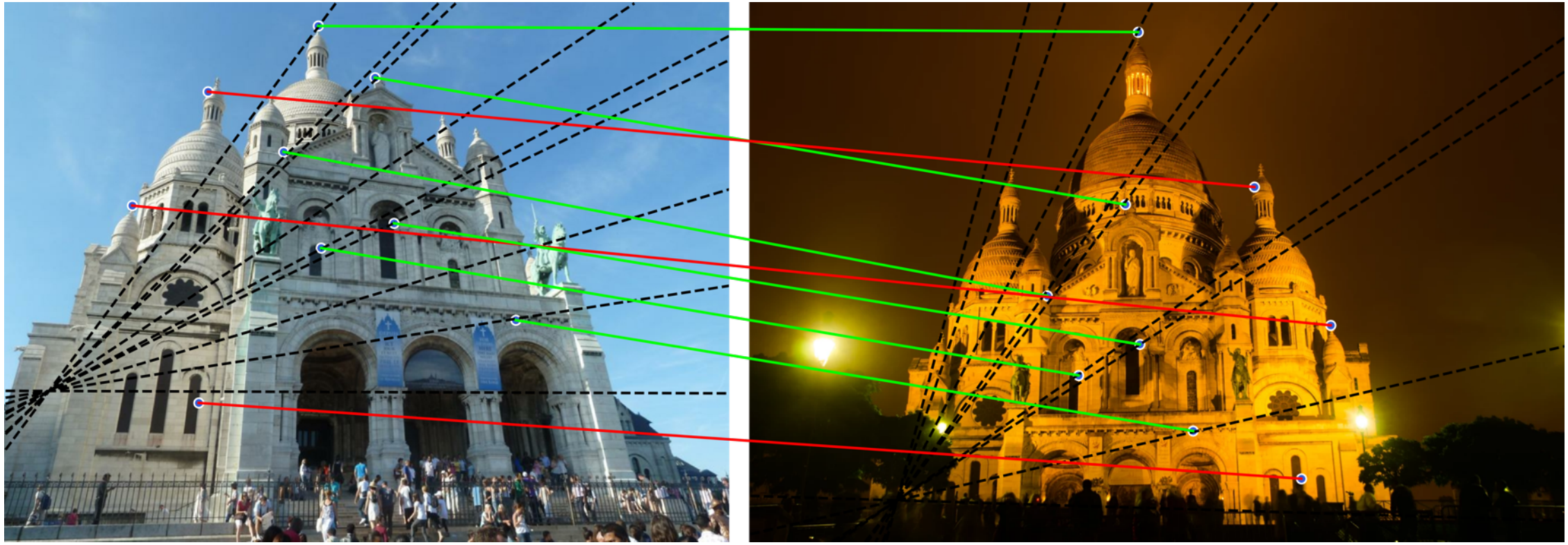
1. Feature Detection and Description

1. Feature Detection and Description

2. Feature Matching

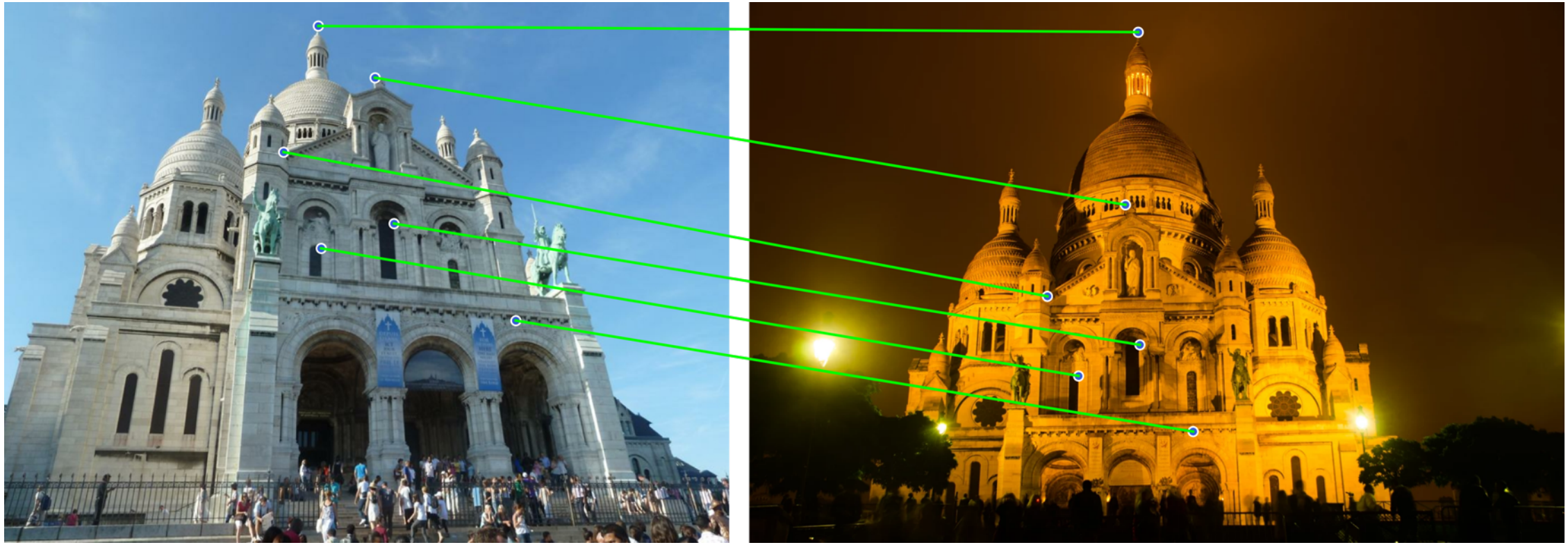


3. Geometric Verification

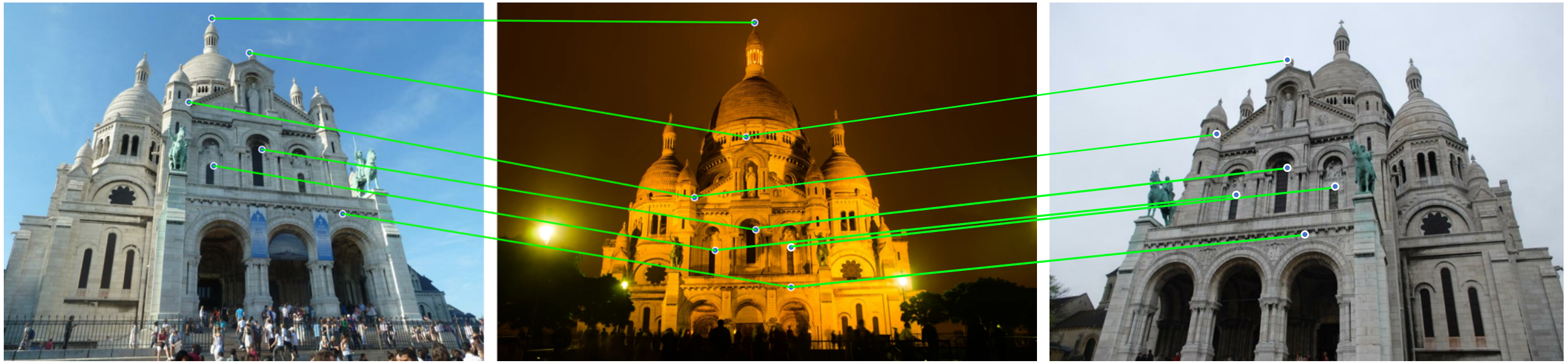


Use epipolar constraints to remove incorrect matches

3. Geometric Verification



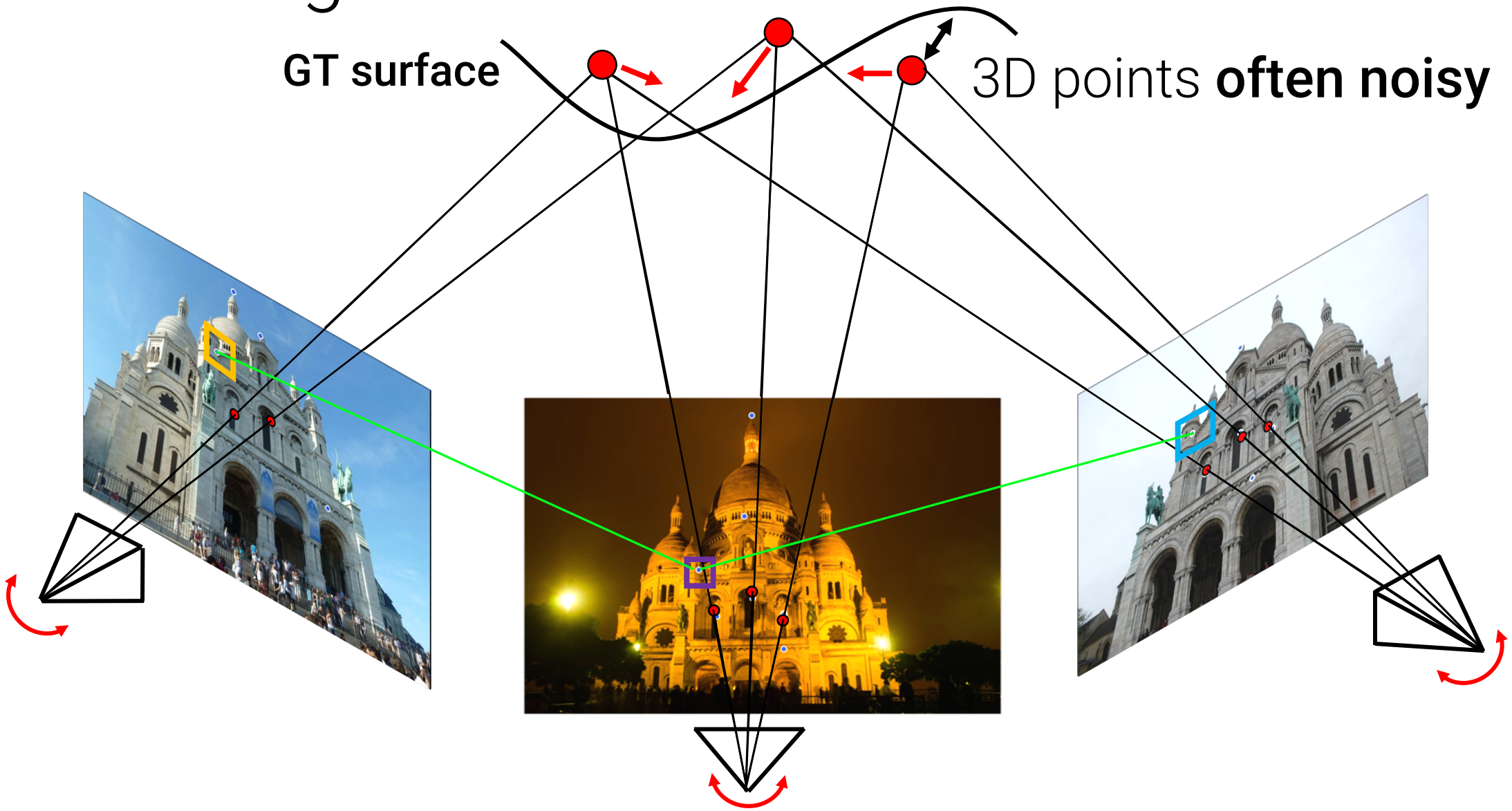
4. Triangulation



4. Triangulation

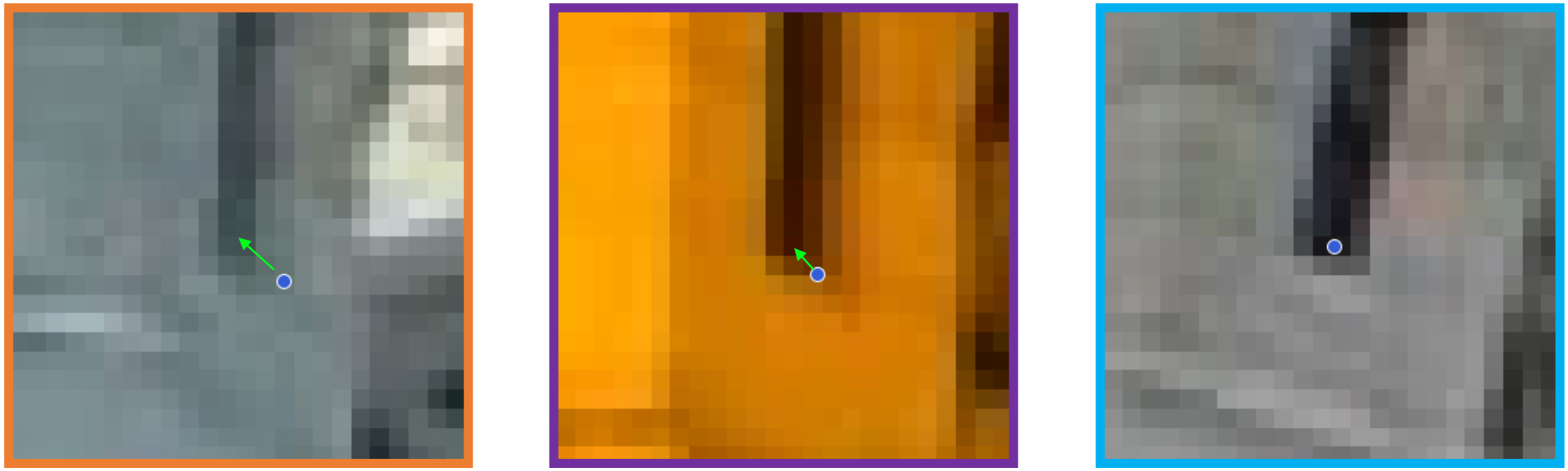


4. Triangulation



Problem: keypoints lack accuracy

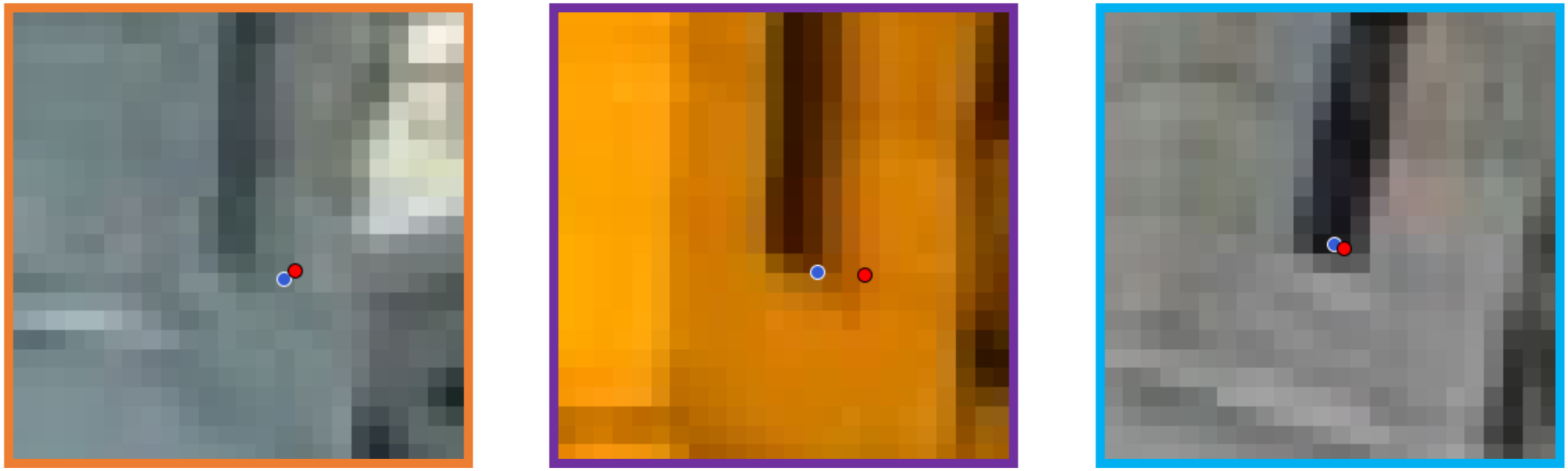
detections •



- **Single-view** keypoint detection: **inherently hard & noisy**
- Deep learned keypoint detectors:
  - ✅ Global context but ❌ retain little local information

Bundle Adjustment

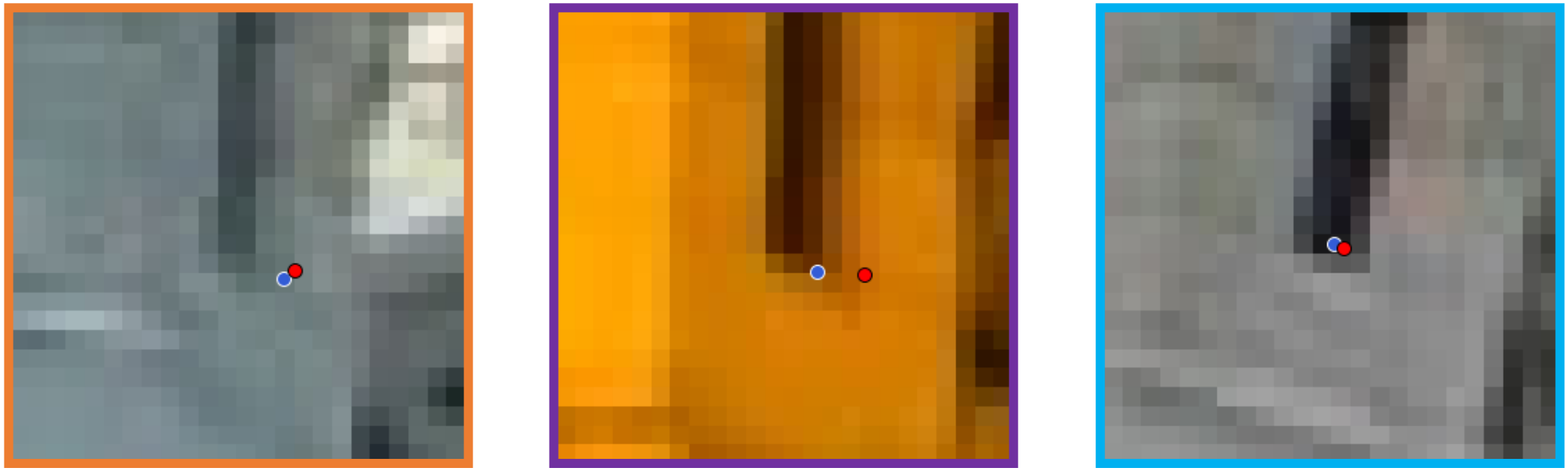
reprojections
detections



$$E_{\mathrm{BA}} = \sum_{j} \sum_{(i,u)\in\mathcal{T}(j)} \left\| \Pi\left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i\right) - \mathbf{p}_u \right\|_\gamma$$

Bundle Adjustment

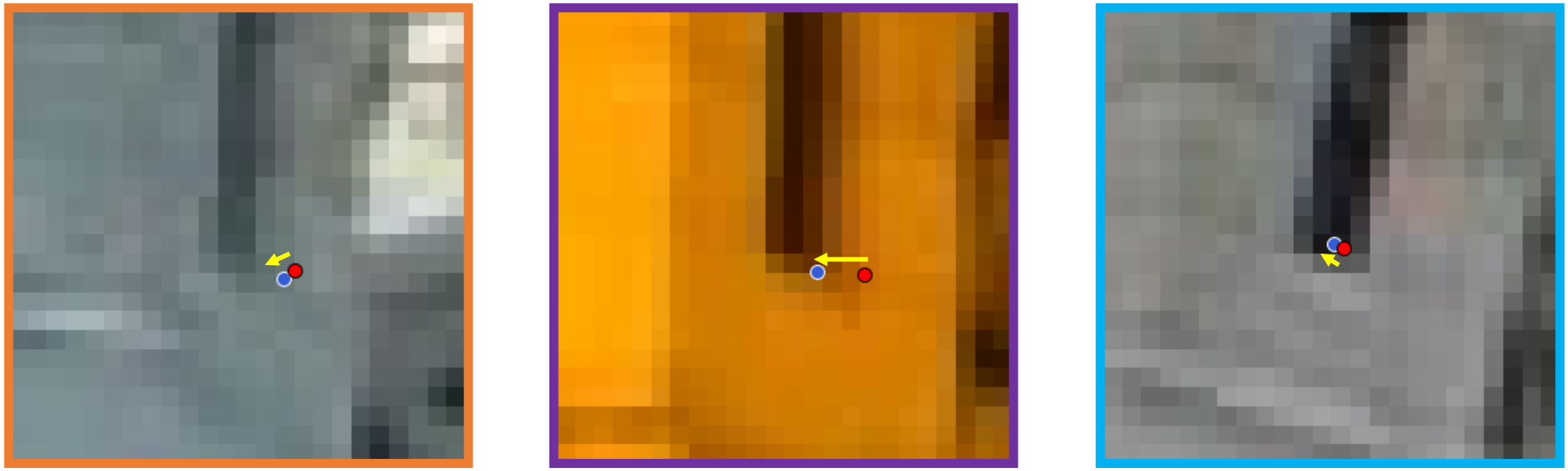
reprojections
detections



$$E_{\mathrm{BA}} = \sum_{j} \sum_{(i,u)\in\mathcal{T}(j)} \left\| \Pi\left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i\right) - \mathbf{p}_u \right\|_\gamma$$

Bundle Adjustment

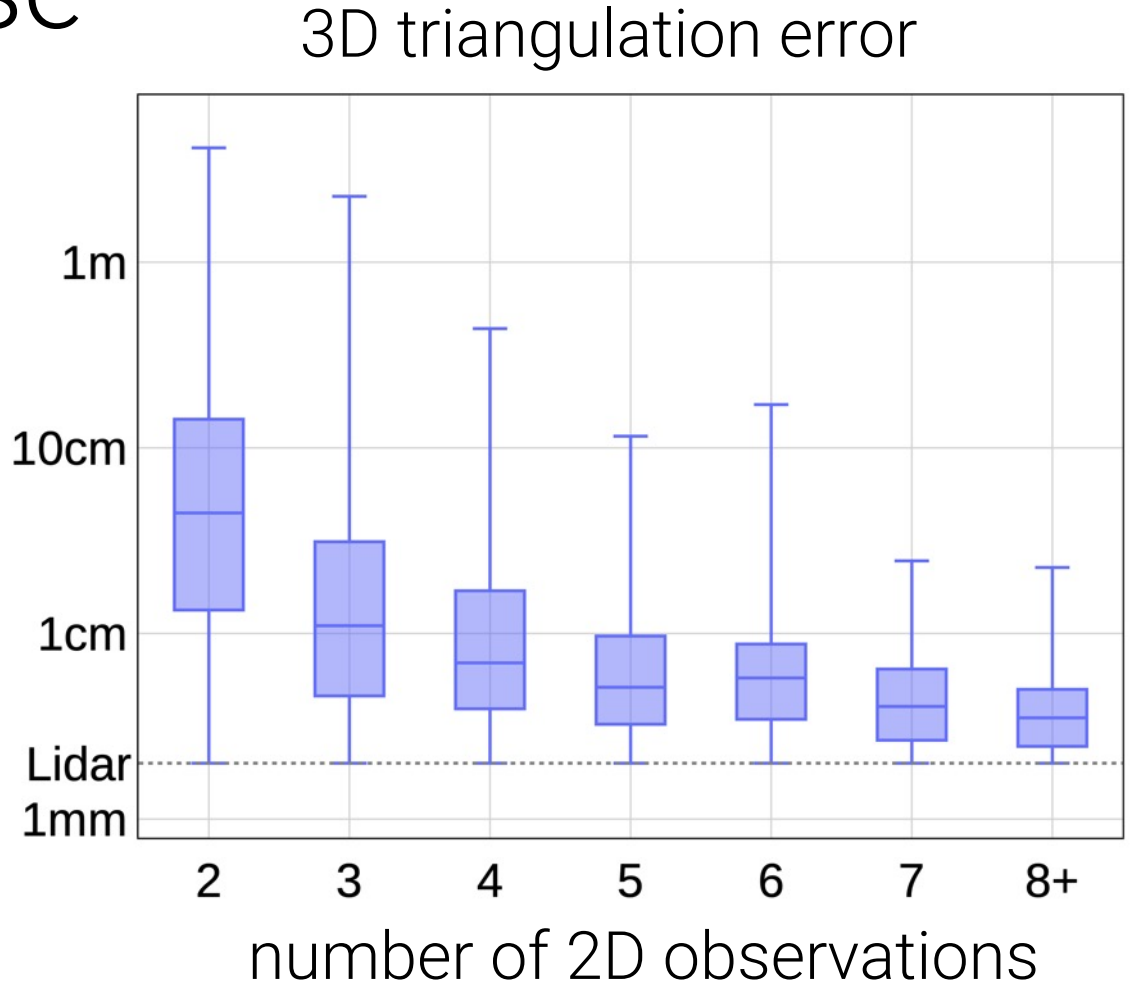
reprojections
detections



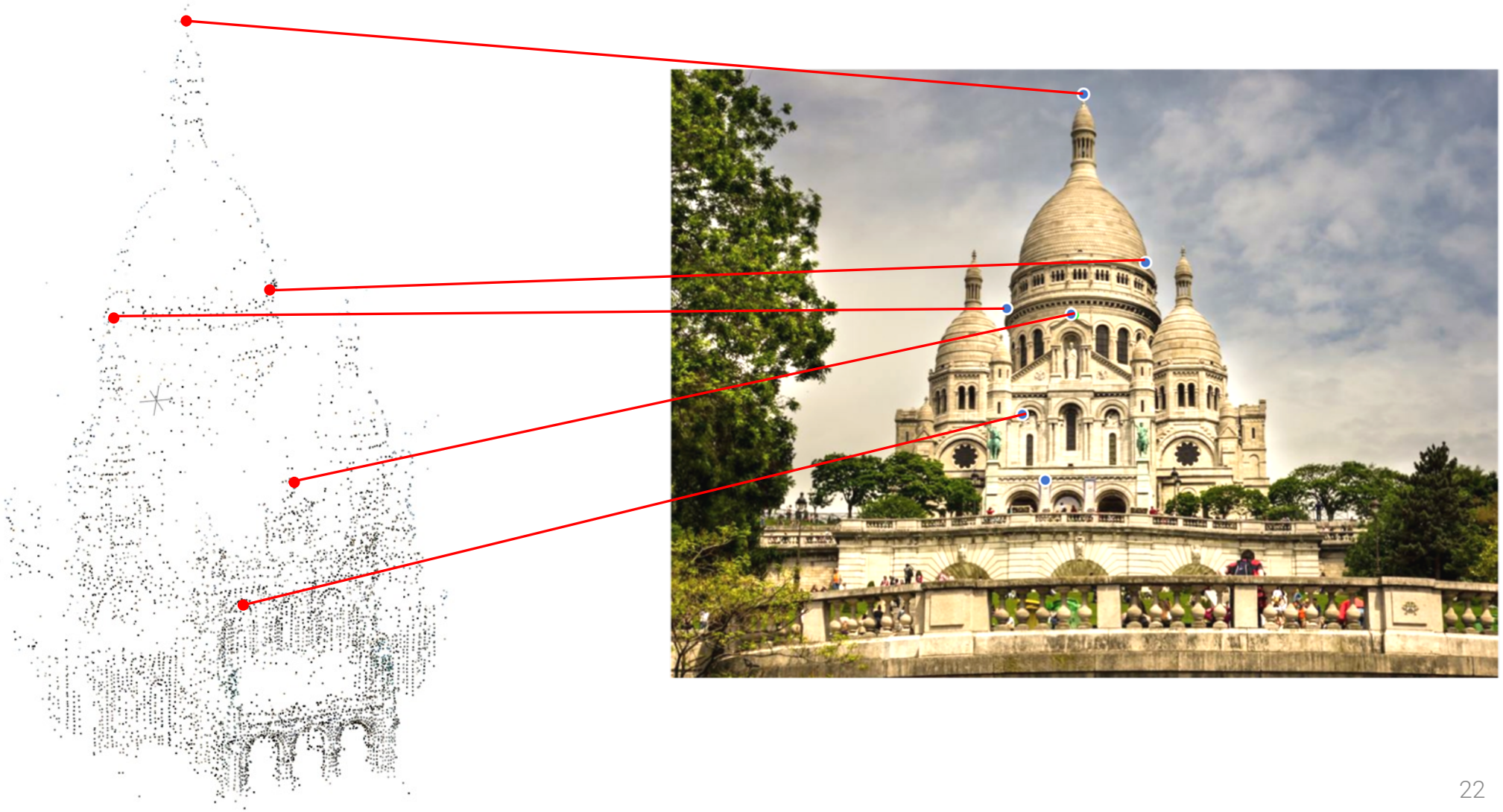
$$E_{\mathrm{BA}} = \sum_{j} \sum_{(i,u)\in\mathcal{T}(j)} \left\| \Pi\left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i\right) - \mathbf{p}_u \right\|_\gamma$$

Impact of keypoint noise

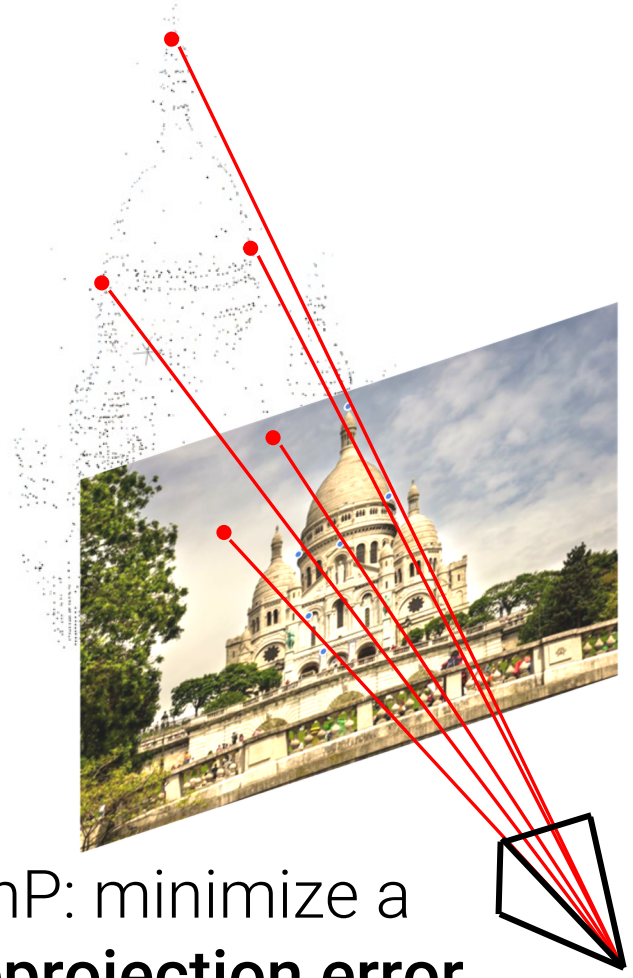
- Few observations
  → **cannot average out the noise**
  → limited accuracy
- **Large-scale** SfM requires
  fewer images with lower ovelap!



Camera Re-Localization

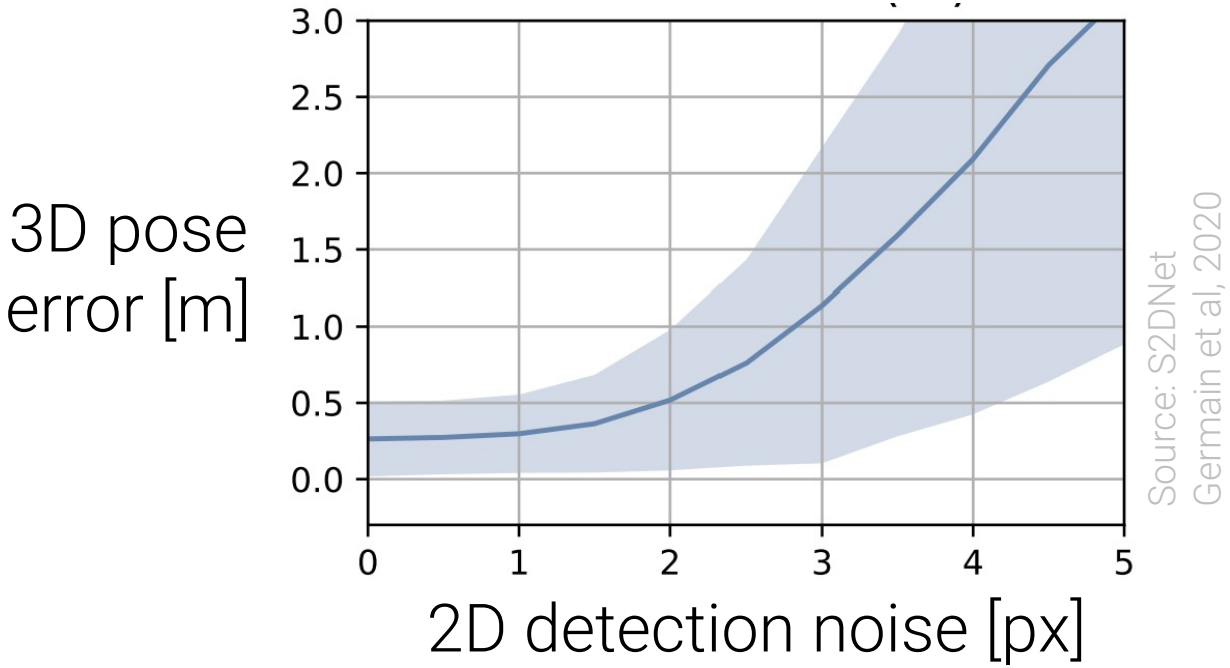


Camera Re-Localization



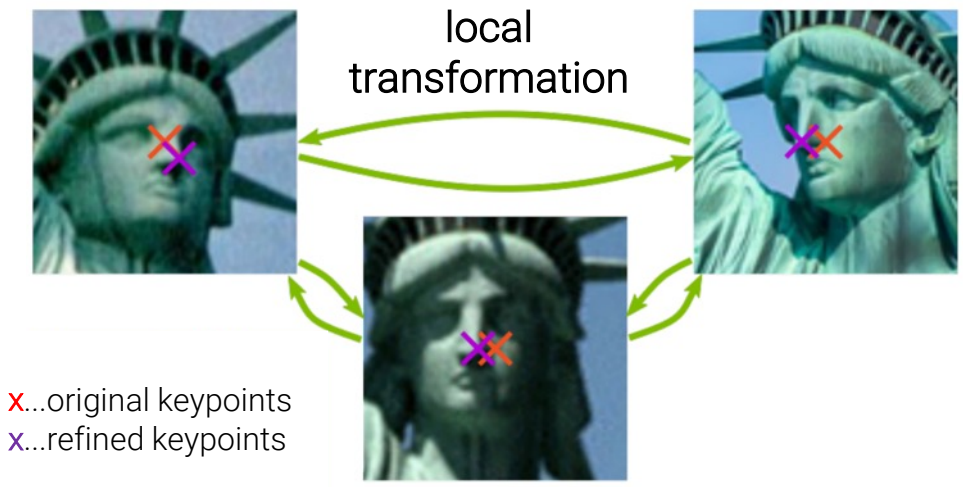
PnP: minimize a
reprojection error

- Small error in 2D detection
  → in **large 3D pose error!**
- Impacts both 3D mapping and localization



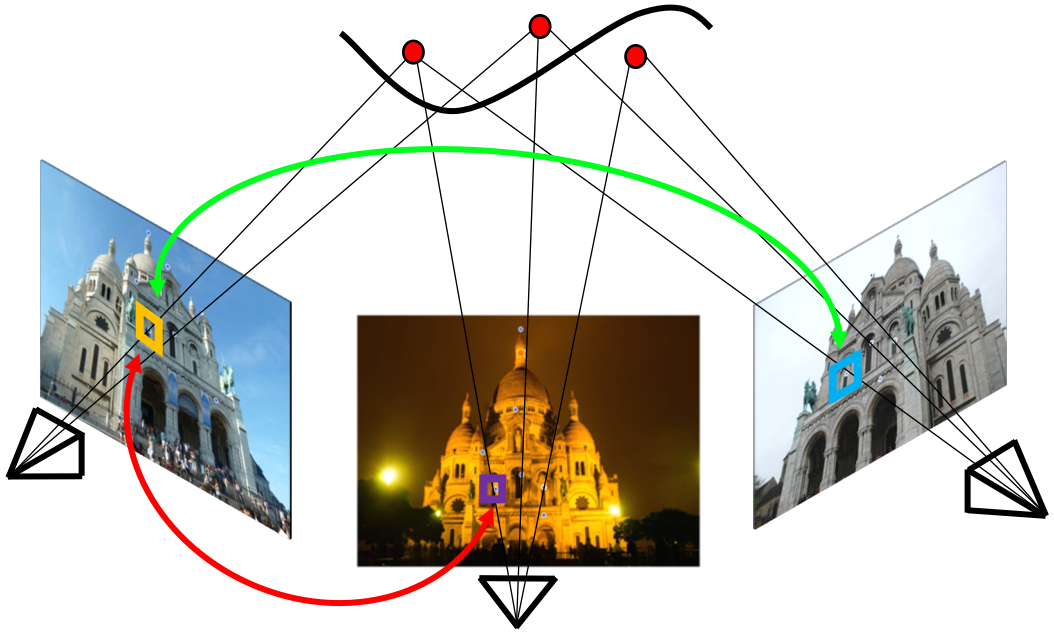
How to improve reconstructions?

Patch Flow, Dusmanu et al, 2020
Refine keypoints **prior to SfM**



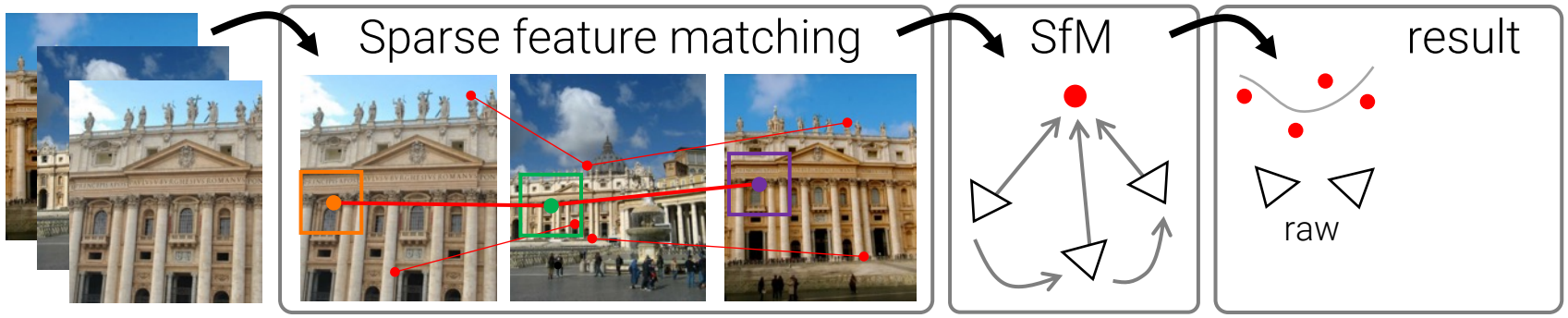
Flow prediction for each **image pair**

Photometric BA
refine 3D reconstruction **after SfM**

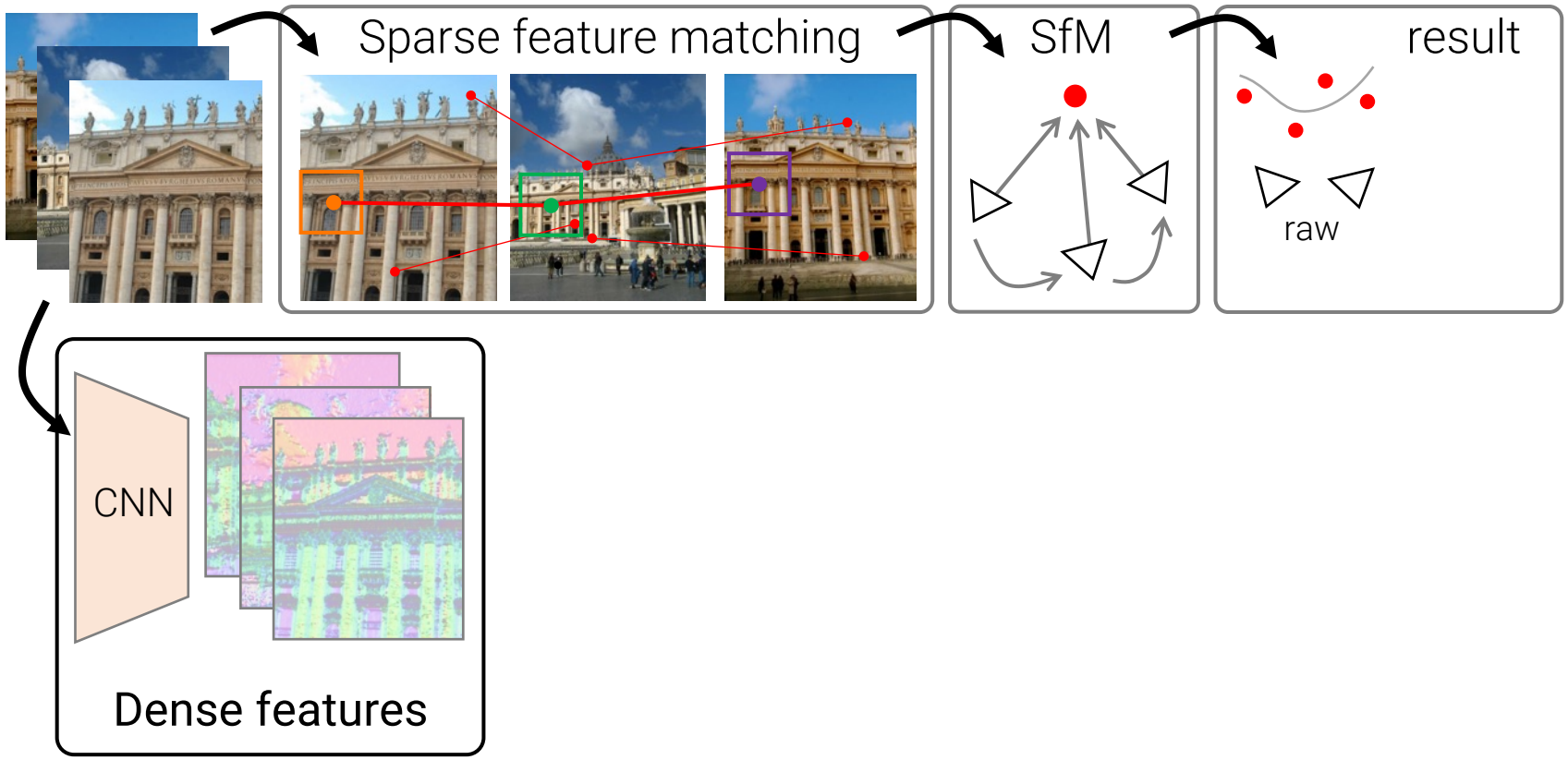


Fails under **large illumination changes**

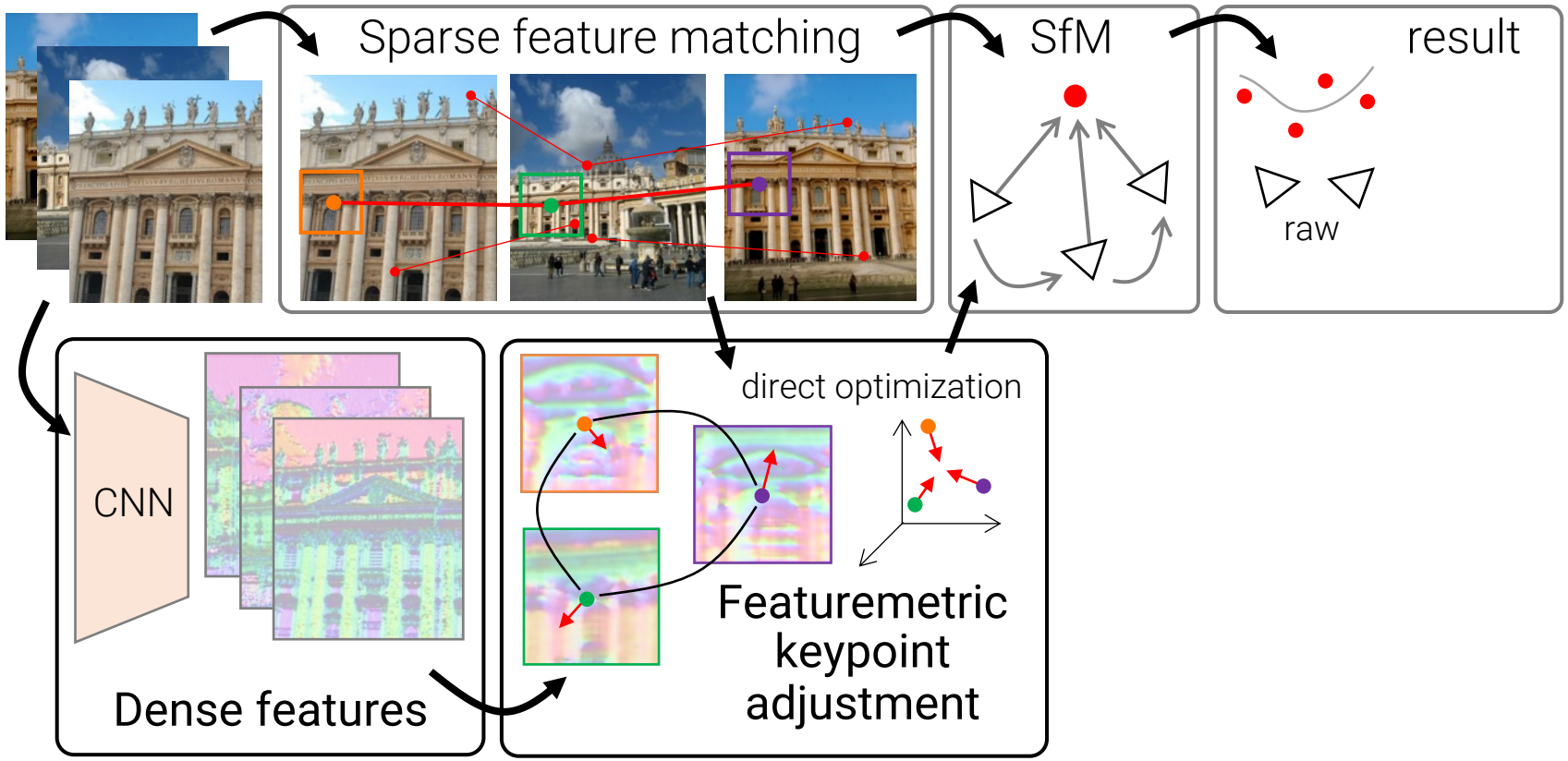
Pixel-Perfect Structure-from-Motion



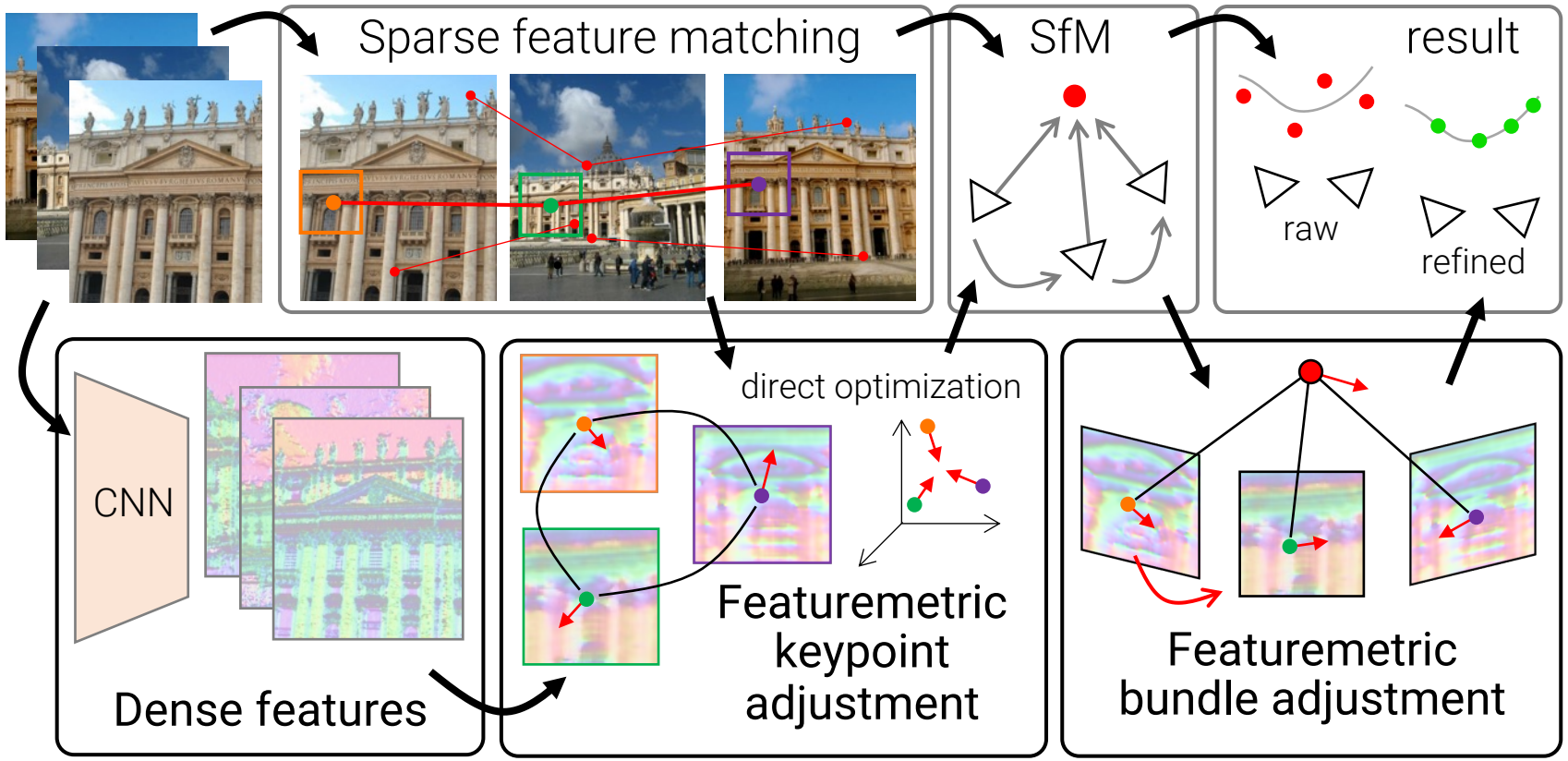
Pixel-Perfect Structure-from-Motion



Pixel-Perfect Structure-from-Motion

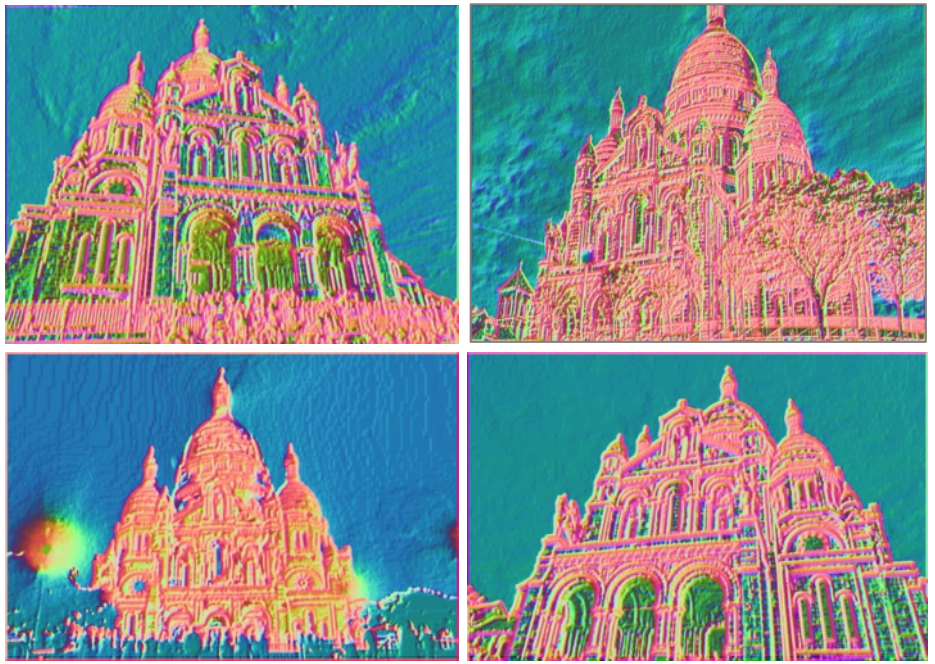


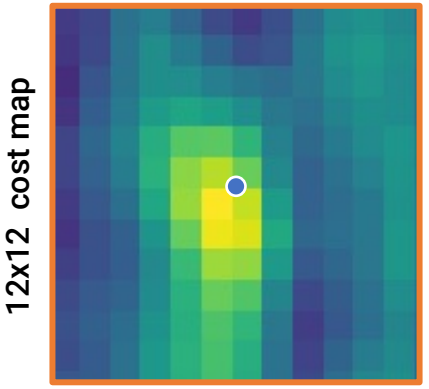
Pixel-Perfect Structure-from-Motion

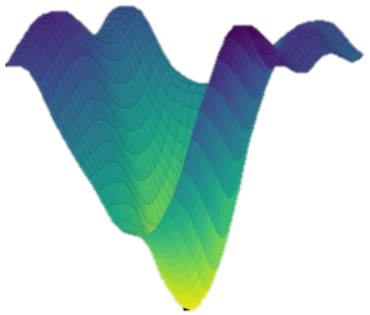


Deep features

CNN







✅ **robust** to **illumination changes**
✅ encode **local information**
✅ **High resolution**

Keypoint Adjustment

Patch Flow, Dusmanu et al, 2020:

$$E^j_{\mathrm{KA}} = \sum_{(u,v)\in\mathcal{M}(j)} \underbrace{\|\mathbf{p}_v + \mathbf{T}_{v\to u}[\mathbf{p}_v]}_{\text{flow-adjusted keypoint}} - \overbrace{\mathbf{p}_u}^{\text{detected keypoint}}\|_\gamma$$

Minimizes **2D point distance** in flow field $\mathbf{T}$

Requires forward-pass for each **image pair**

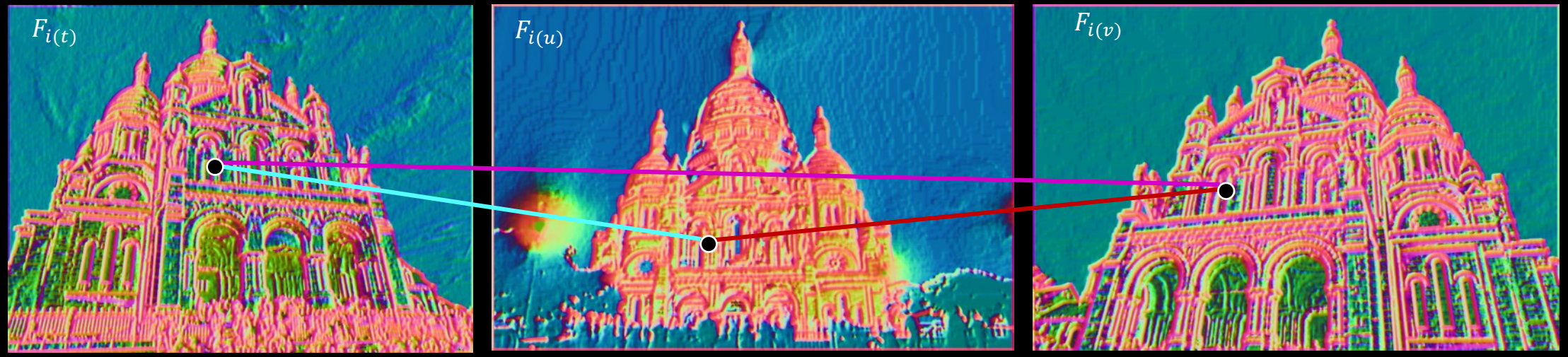
Feature-metric Keypoint Adjustment:

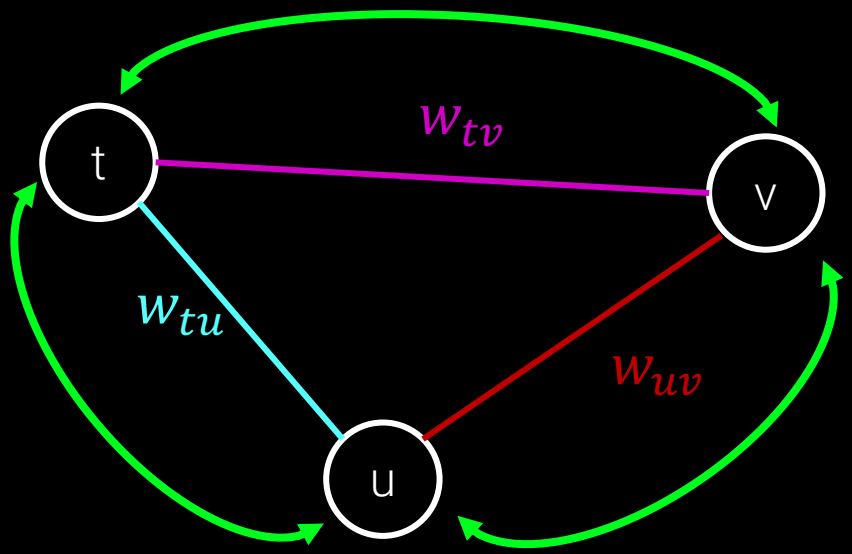
$$E^j_{\mathrm{FKA}} = \sum_{(u,v)\in\mathcal{M}(j)} w_{uv} \left\|\underbrace{\mathbf{F}_{i(u)}[\mathbf{p}_u]}_{\text{interpolated feature descriptor}} - \overbrace{\mathbf{F}_{i(v)}[\mathbf{p}_v]}^{\text{interpolated feature descriptor}}\right\|_\gamma$$

Minimizes **featuremetric error**

requires forward-pass for each **image**

Featuremetric Keypoint Adjustment



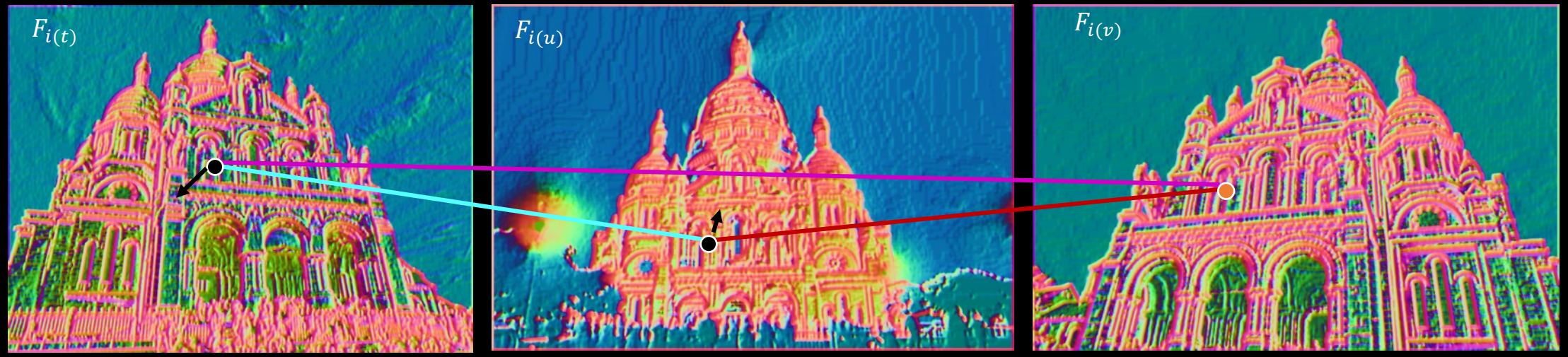


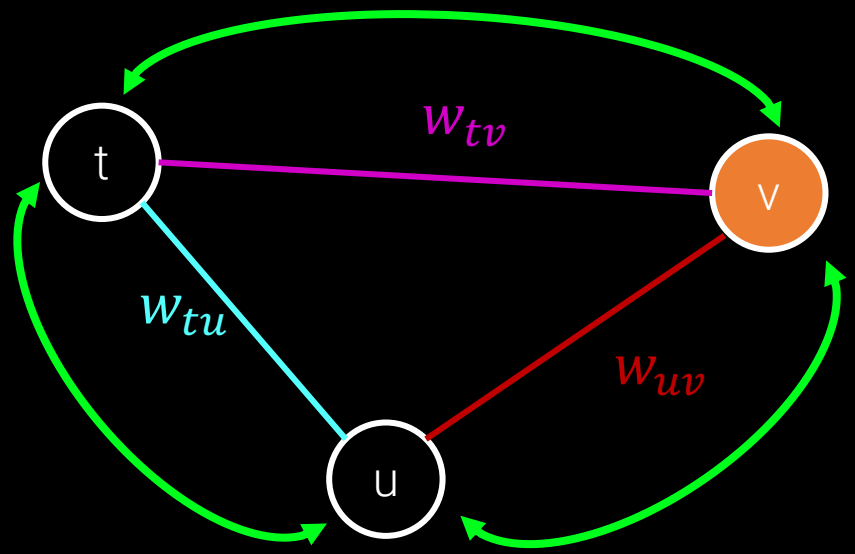
Minimize **featuremetric error** at keypoint locations

$$E^j_{\mathrm{FKA}} = \sum_{(u,v)\in\mathcal{M}(j)} w_{uv} \left\| \mathbf{F}_{i(u)}[\mathbf{p}_u] - \mathbf{F}_{i(v)}[\mathbf{p}_v] \right\|_\gamma$$

matching score

Featuremetric Keypoint Adjustment





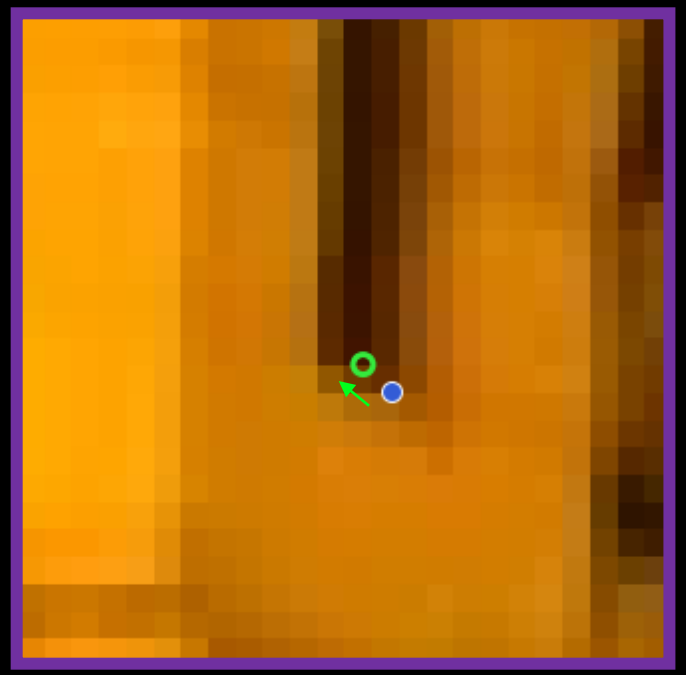
Minimize **featuremetric error** at keypoint locations

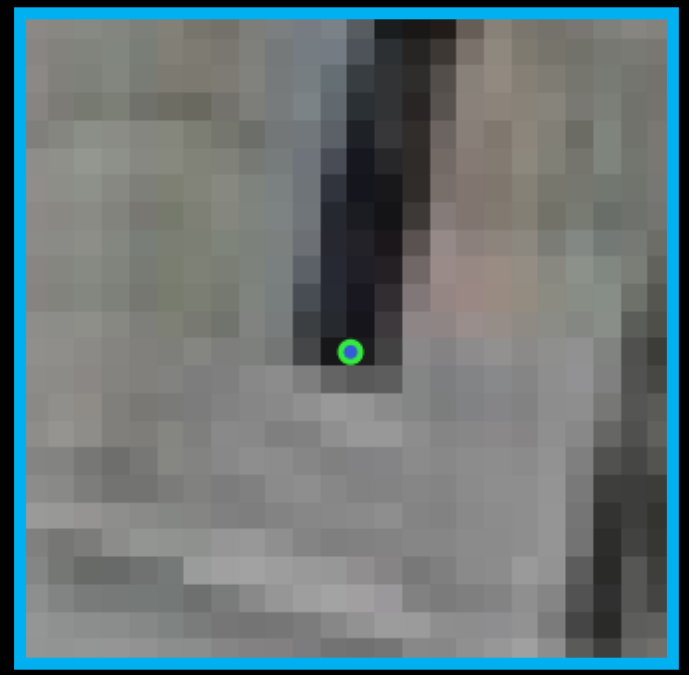
$$E^j_{\text{FKA}} = \sum_{(u,v)\in\mathcal{M}(j)} w_{uv} \left\| \mathbf{F}_{i(u)}[\mathbf{p}_u] - \mathbf{F}_{i(v)}[\mathbf{p}_v] \right\|_\gamma$$

matching score

Better localized keypoints

raw
refined

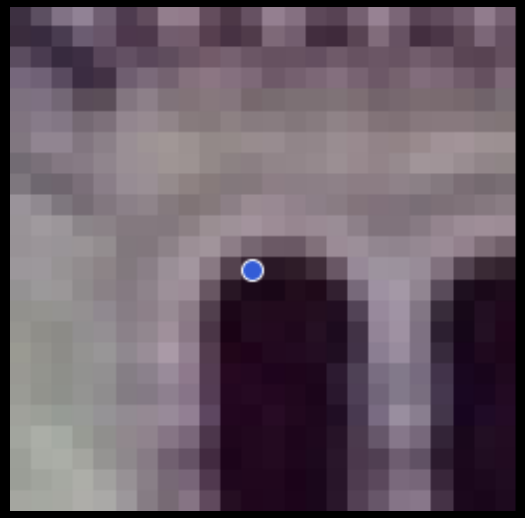
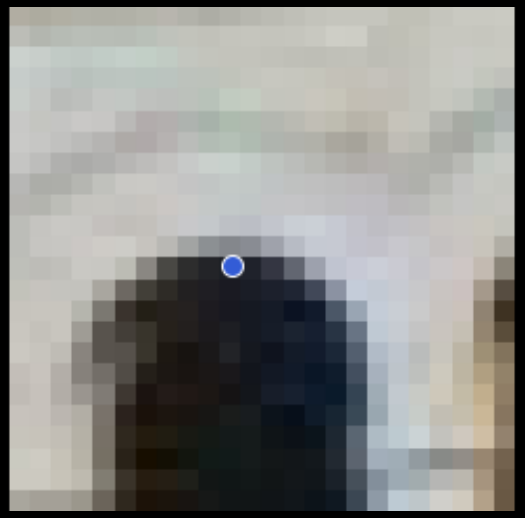
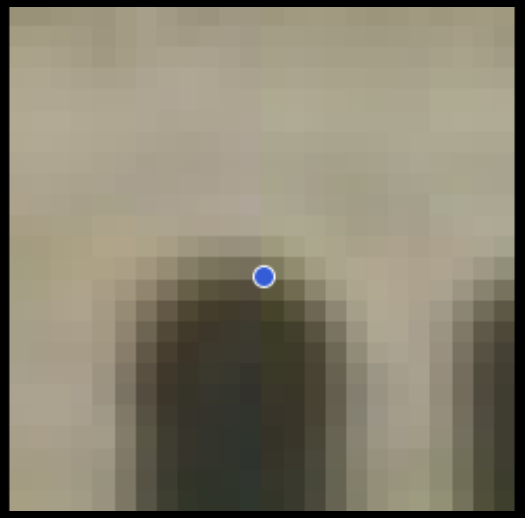
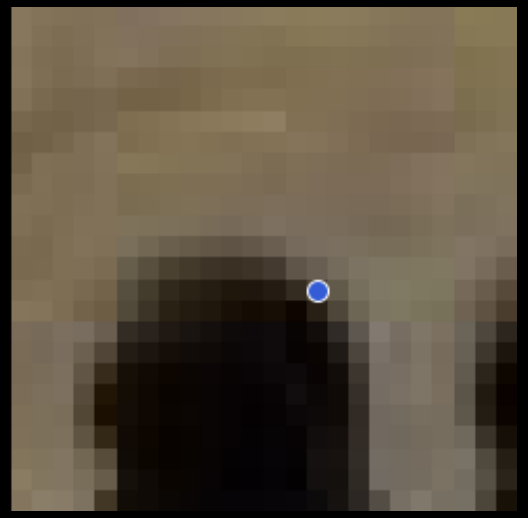


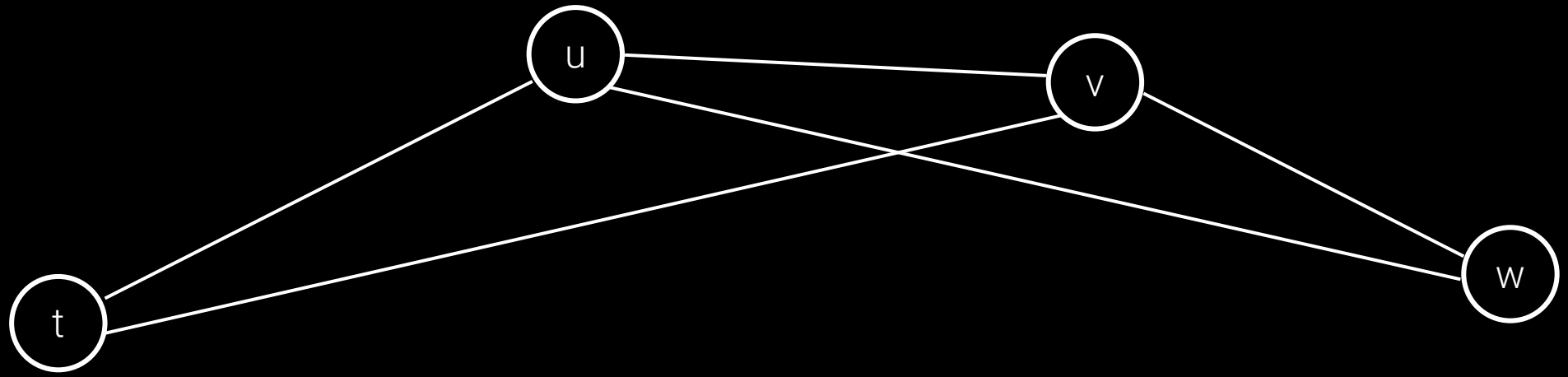


→ more inliers after geometric verification

Merging points

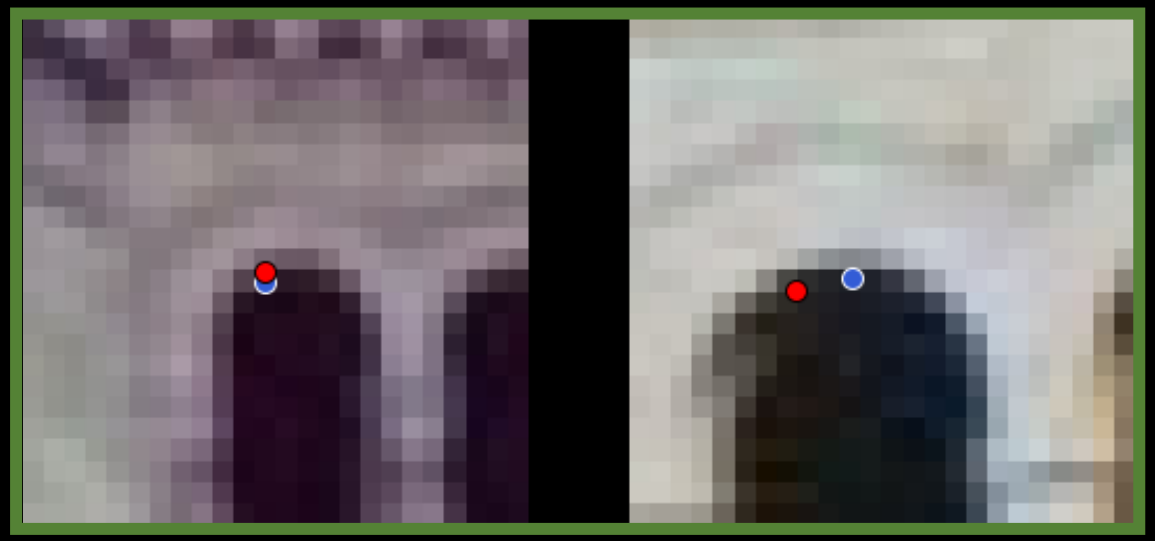
- raw detection
- raw projection
- refined detection
- refined projection

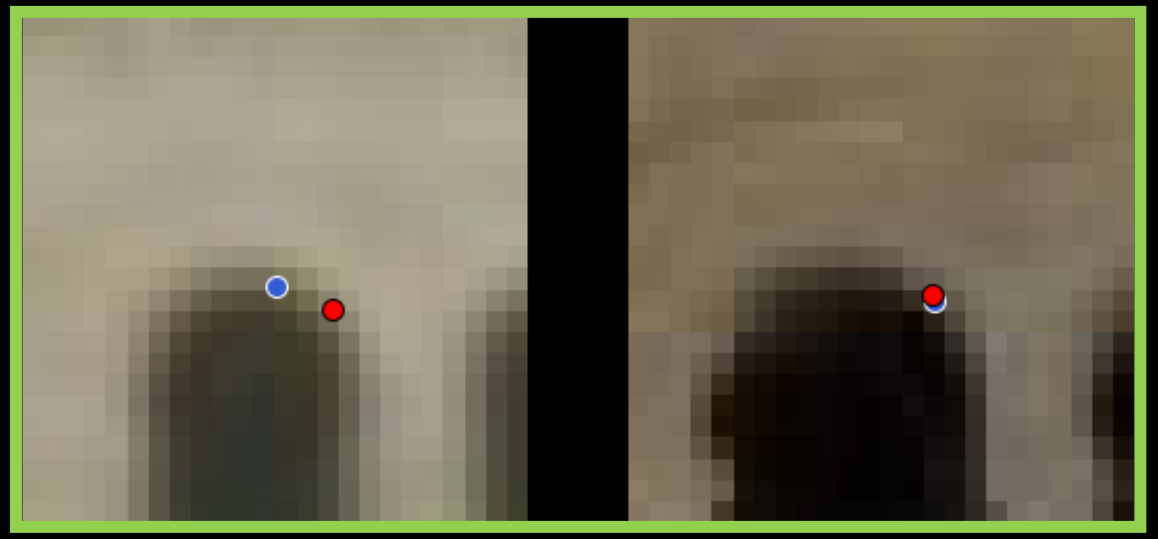







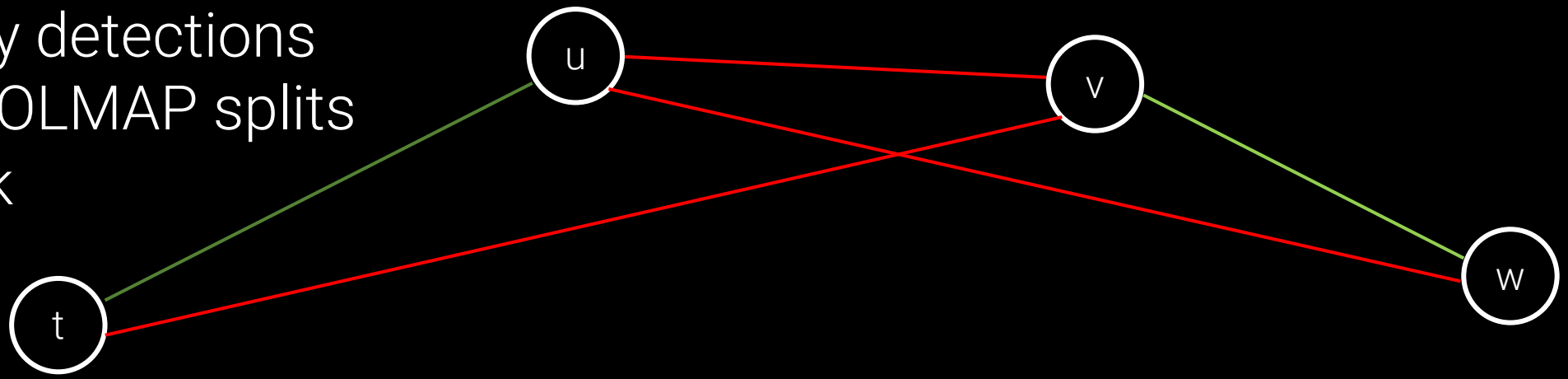
Merging points

raw detection
raw projection
refined detection
refined projection



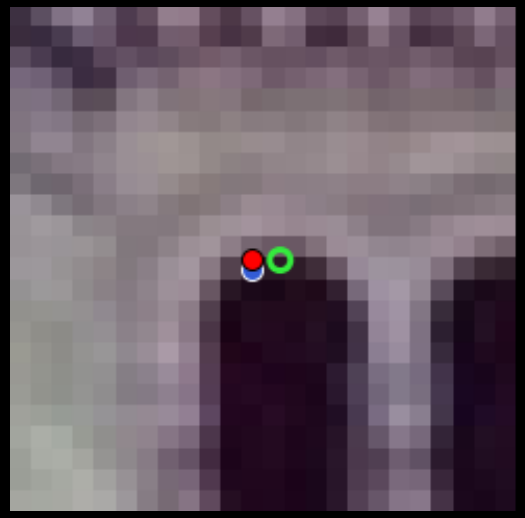


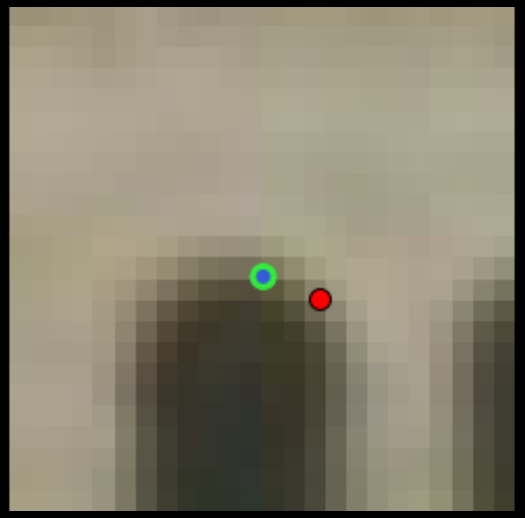
noisy detections
→ COLMAP splits
track

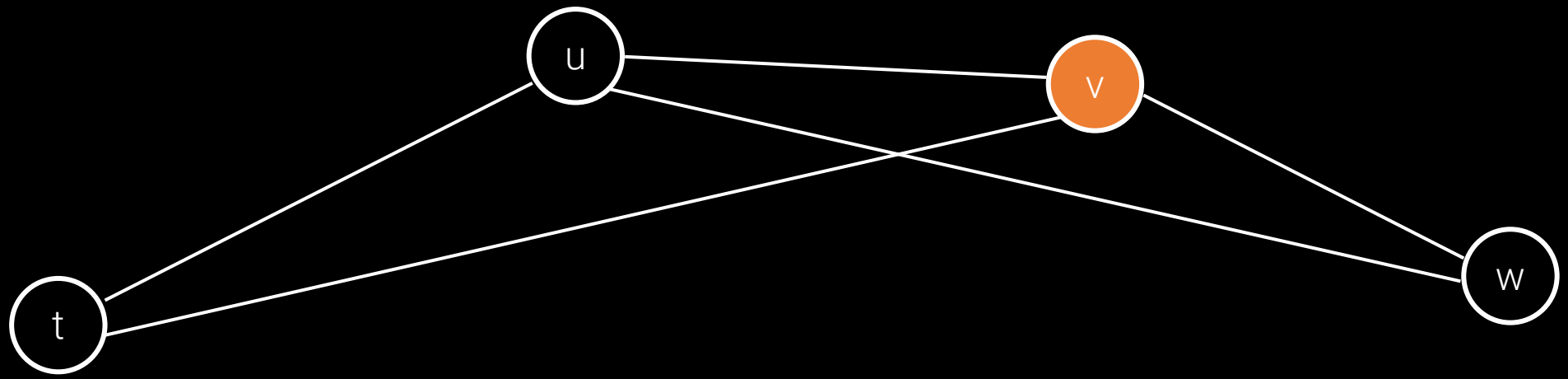


Merging points

- raw detection
- raw projection
- refined detection
- refined projection



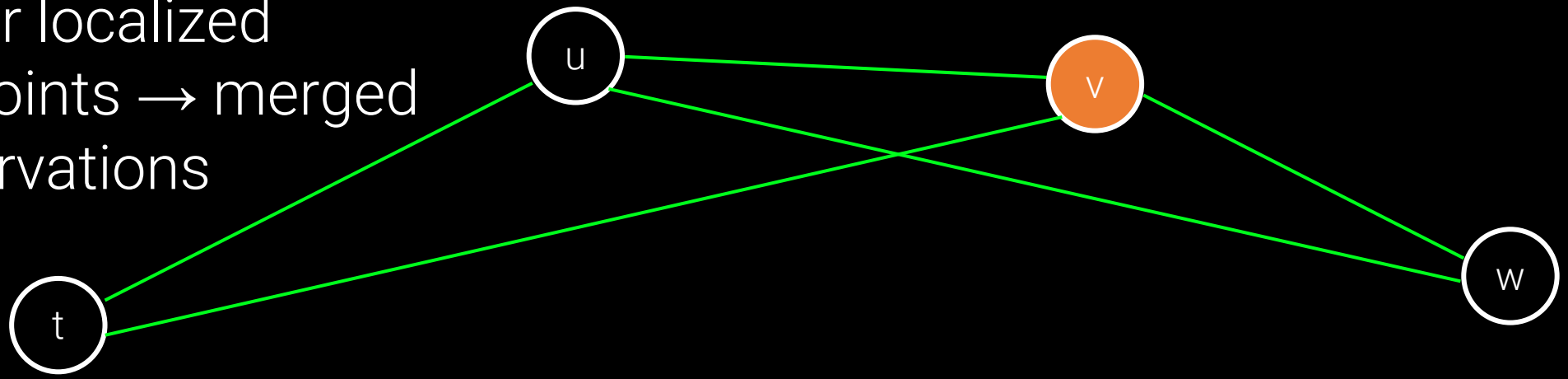


Merging points

raw detection
raw projection
refined detection
refined projection

better localized keypoints → merged observations



Bundle Adjustment

Geometric Bundle Adjustment:

$$E_{\mathrm{BA}} = \sum_{j} \sum_{(i,u)\in\mathcal{T}(j)} \left\| \Pi\left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i\right) - \mathbf{p}_u \right\|_\gamma$$

detected keypoint

projection of 3D point in image

Minimizes a **2D point error**

→ **ignores** the detection **uncertainties**

Featuremetric Bundle Adjustment:

$$E_{\mathrm{FBA}} = \sum_{j} \sum_{(i,u)\in\mathcal{T}(j)} \left\| \mathbf{F}_i\left[\Pi\left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i\right)\right] - \mathbf{f}^j \right\|_\gamma$$

reference descriptor

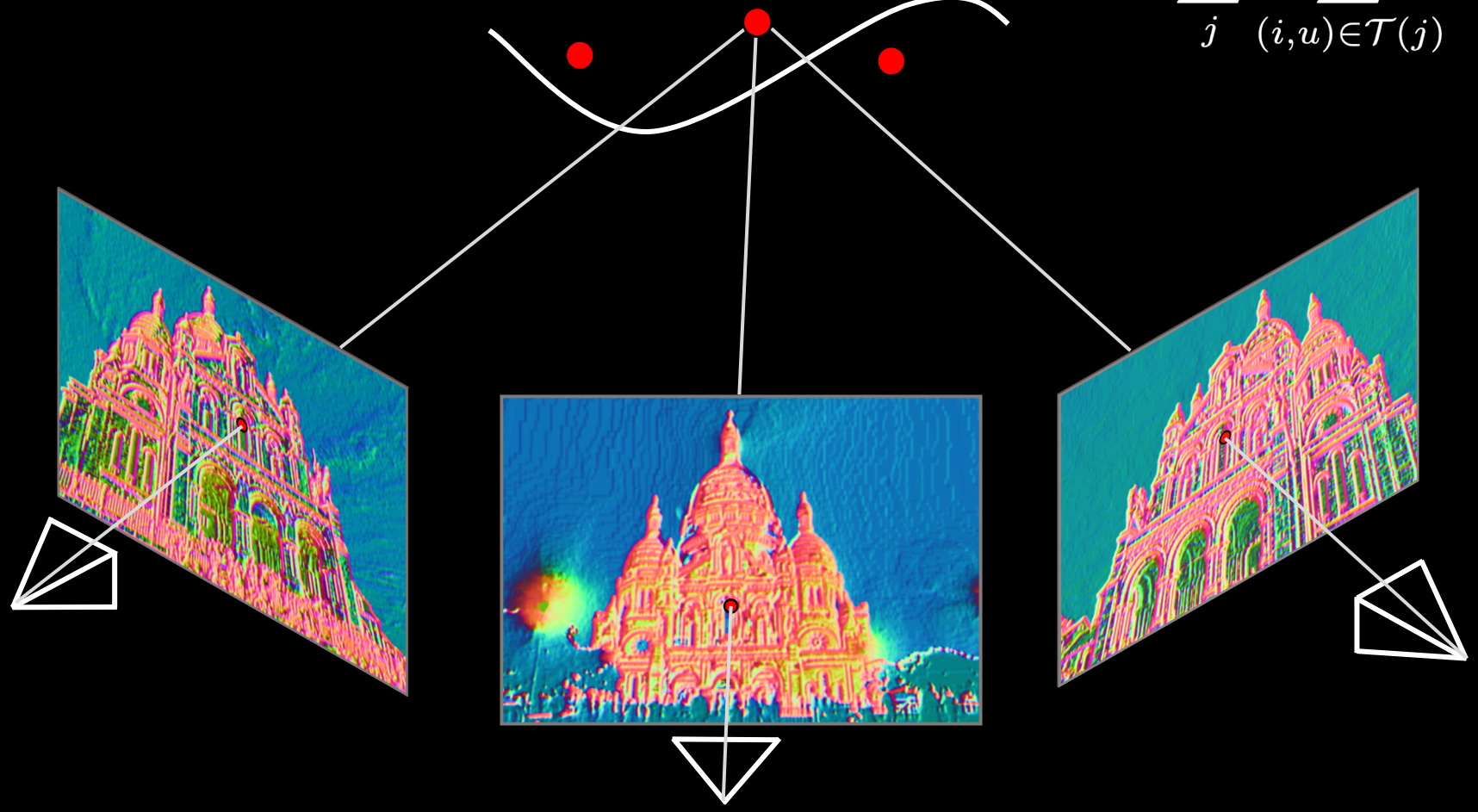
feature map interpolation at projected location

Minimizes **featuremetric error**

→ uses **robust appearance information**

Featuremetric Bundle Adjustment

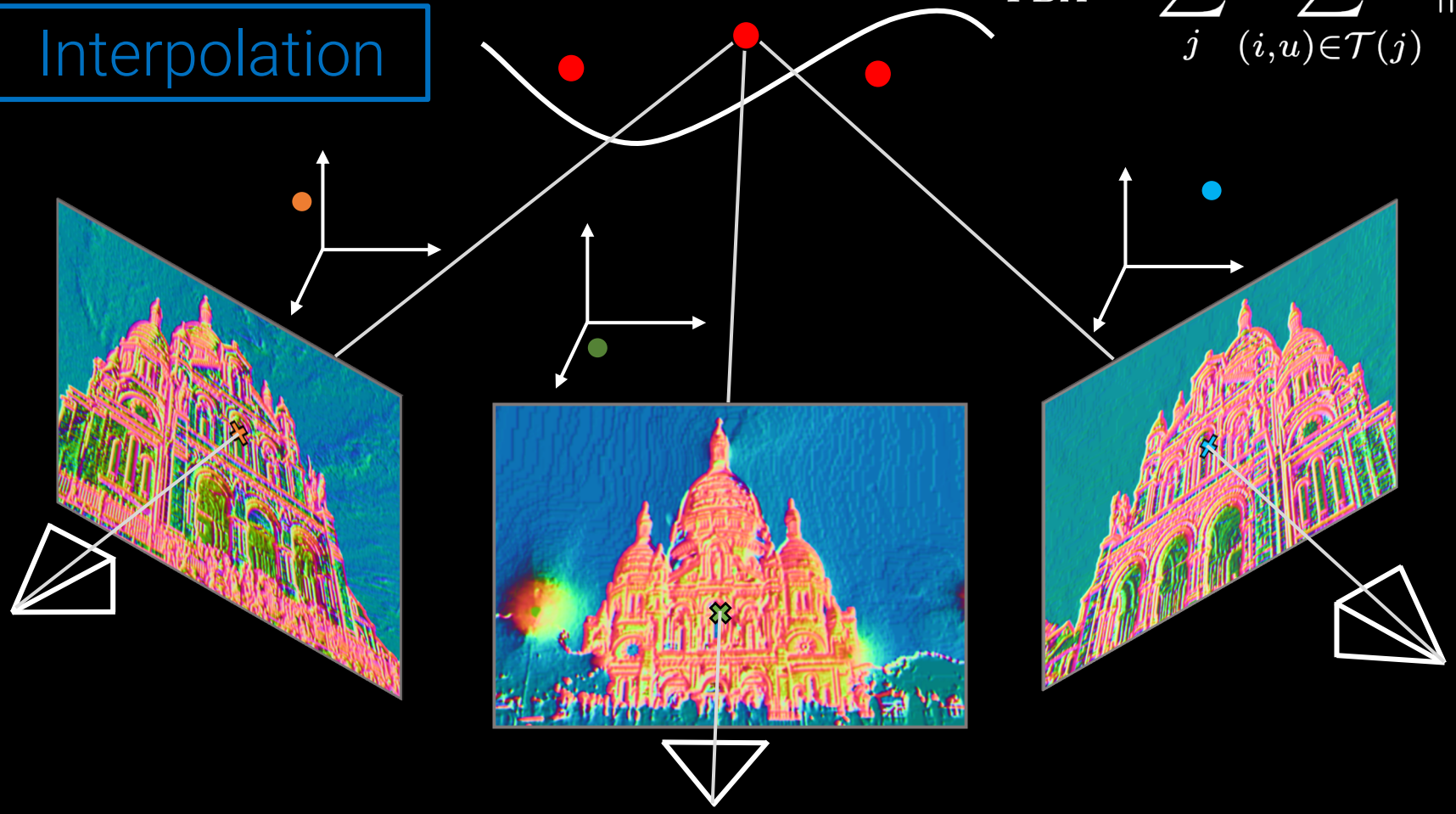
$$E_{\mathrm{FBA}} = \sum_{j} \sum_{(i,u)\in\mathcal{T}(j)} \left\| \mathbf{F}_i \left[\Pi \left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i \right) \right] - \mathbf{f}^j \right\|_\gamma$$



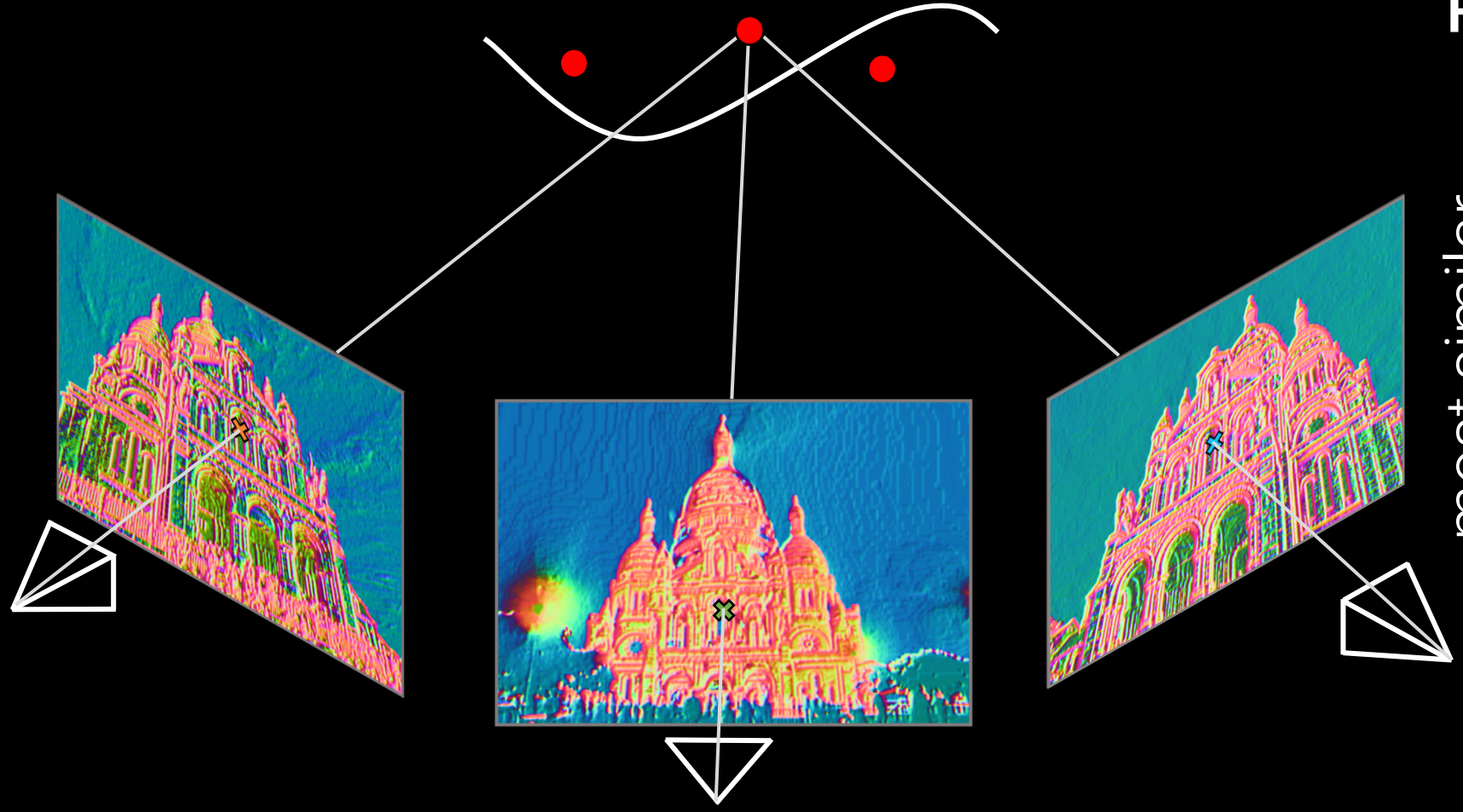
Featuremetric Bundle Adjustment

$$E_{\mathrm{FBA}} = \sum_{j} \sum_{(i,u)\in\mathcal{T}(j)} \left\| \mathbf{F}_i \left[\Pi \left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i \right) \right] - \mathbf{f}^j \right\|_\gamma$$

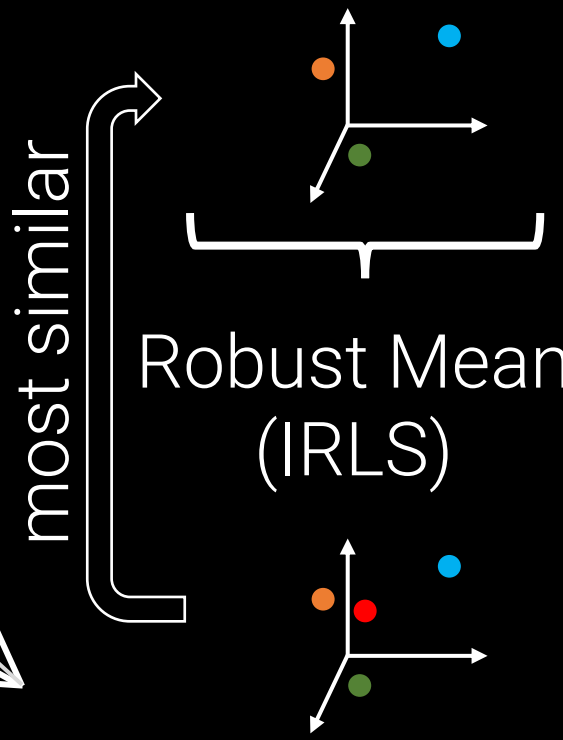
Interpolation



Featuremetric Bundle Adjustment



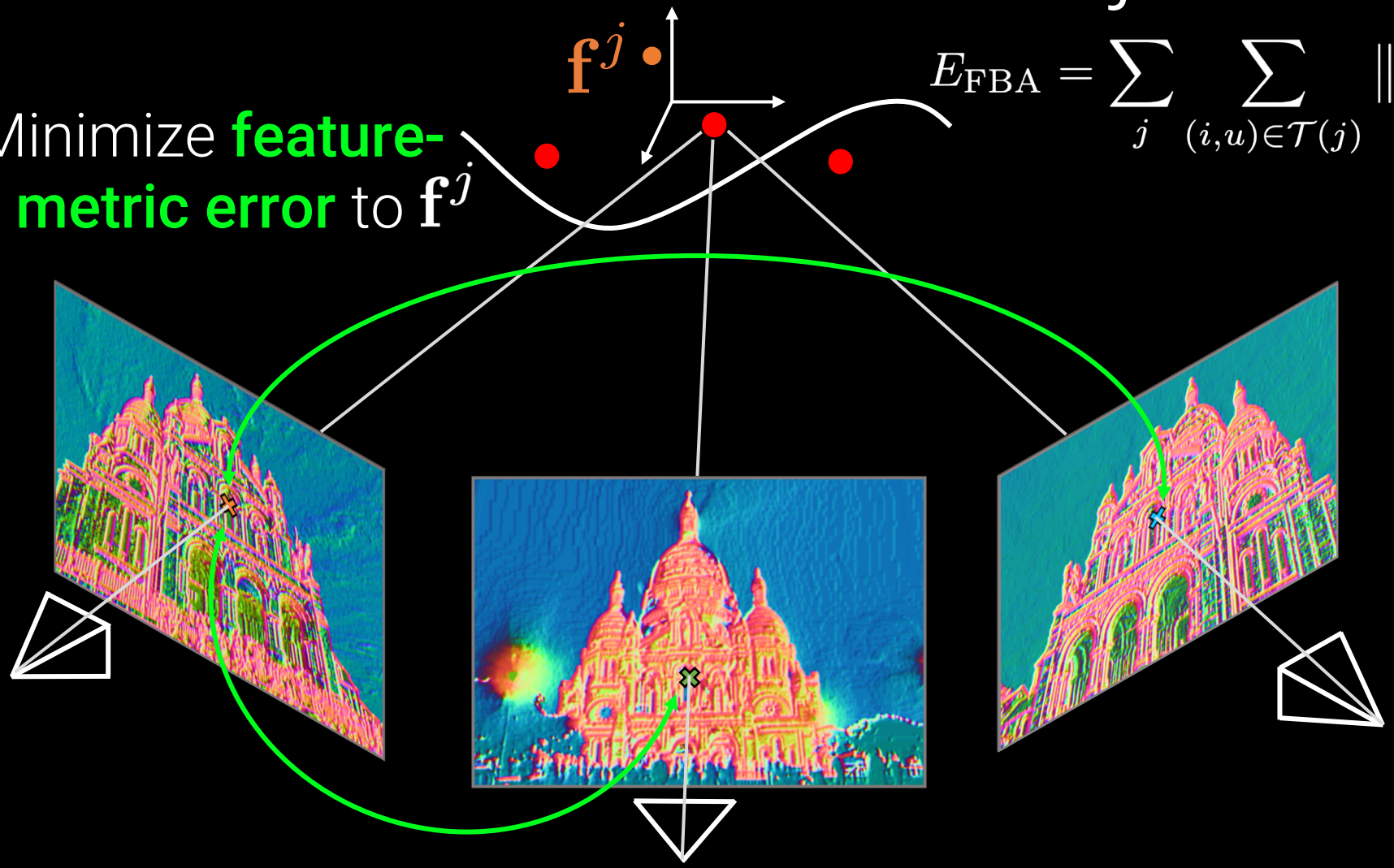
Reference Selection:



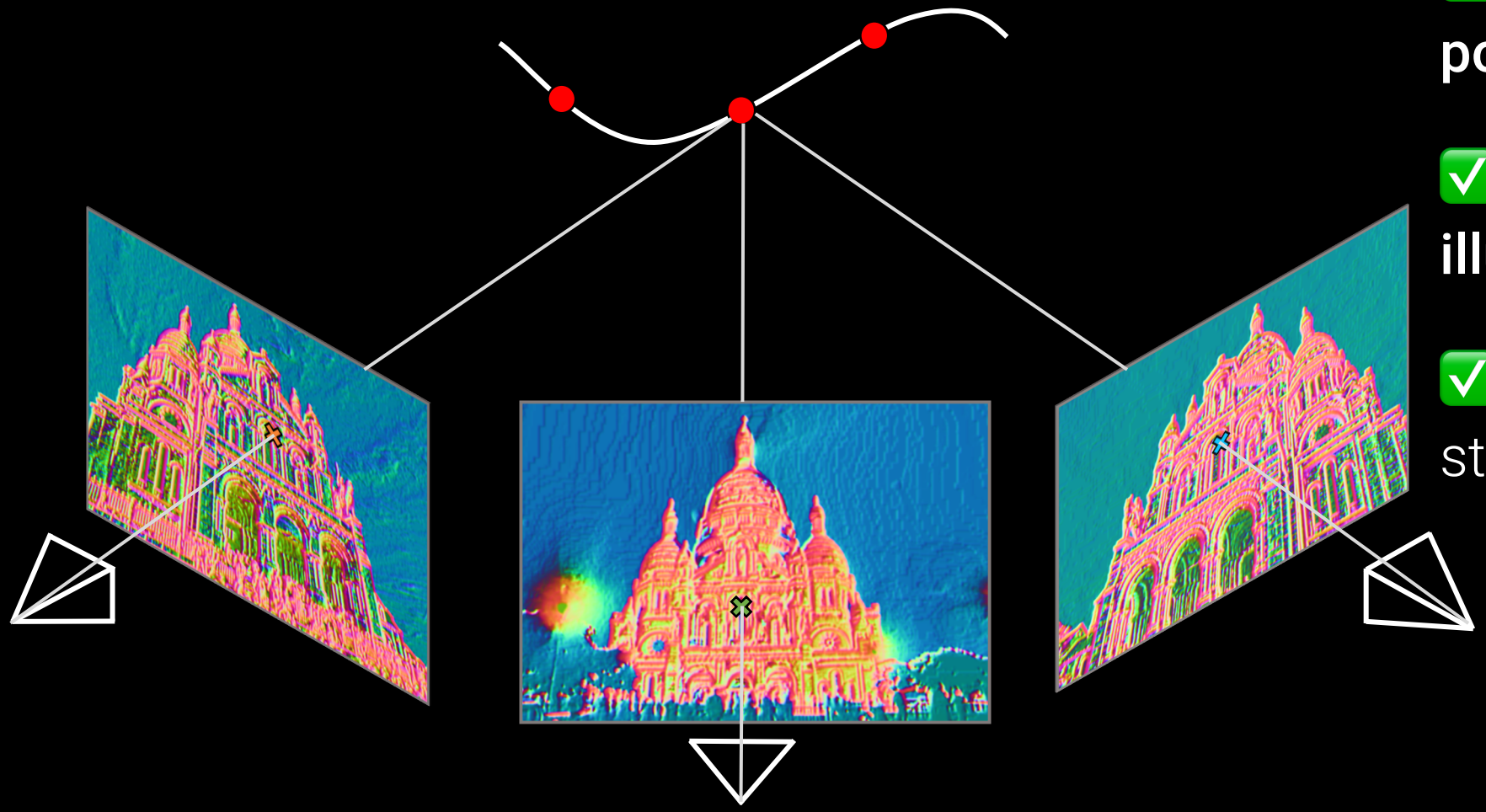
Featuremetric Bundle Adjustment

Minimize **feature-metric error** to $\mathbf{f}^j$

$$E_{\text{FBA}} = \sum_j \sum_{(i,u)\in\mathcal{T}(j)} \left\| \mathbf{F}_i \left[\Pi \left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i \right) \right] - \mathbf{f}^j \right\|_\gamma$$



Featuremetric Bundle Adjustment



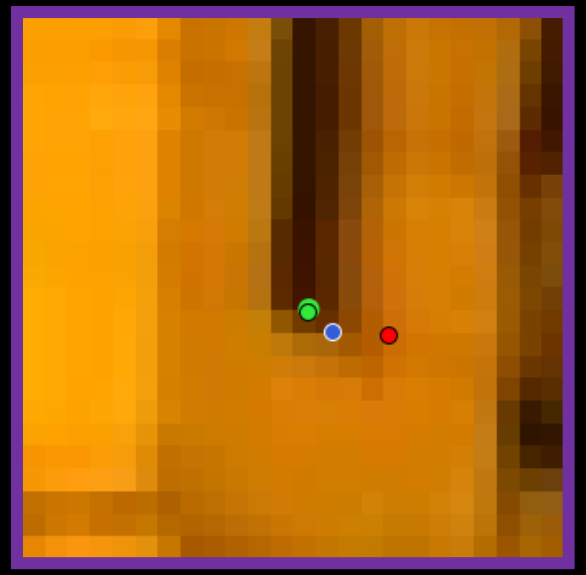
✅ **subpixel** accurate **points & poses**

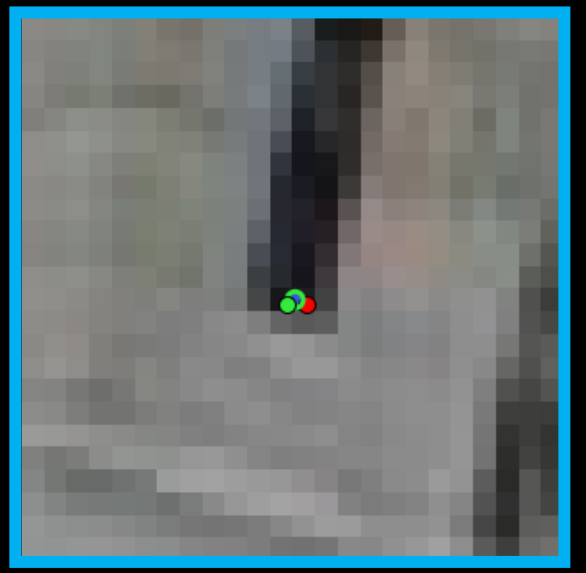
✅ Robust to **illumination changes**

✅ **Similar jacobian** structure as **geom. BA**

Gradients

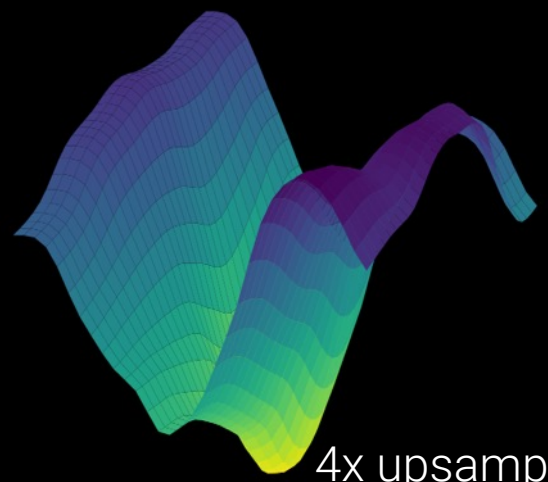
Imaeg Patch



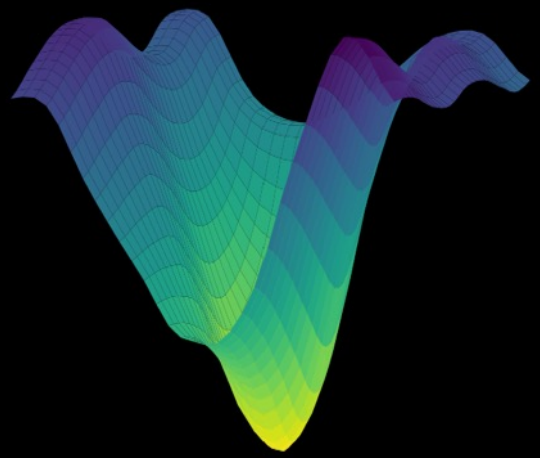


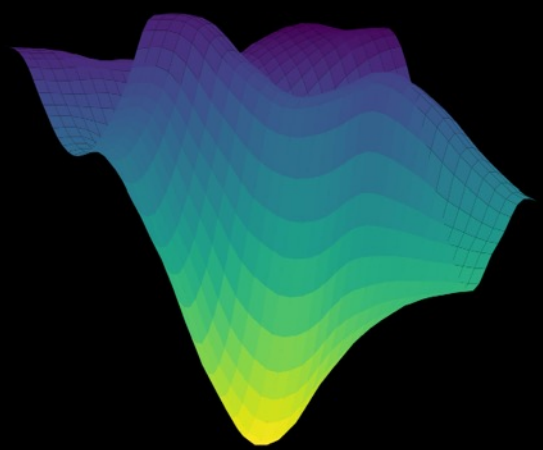
- raw detection
- raw projection
- refined detection
- refined projection

Local cost

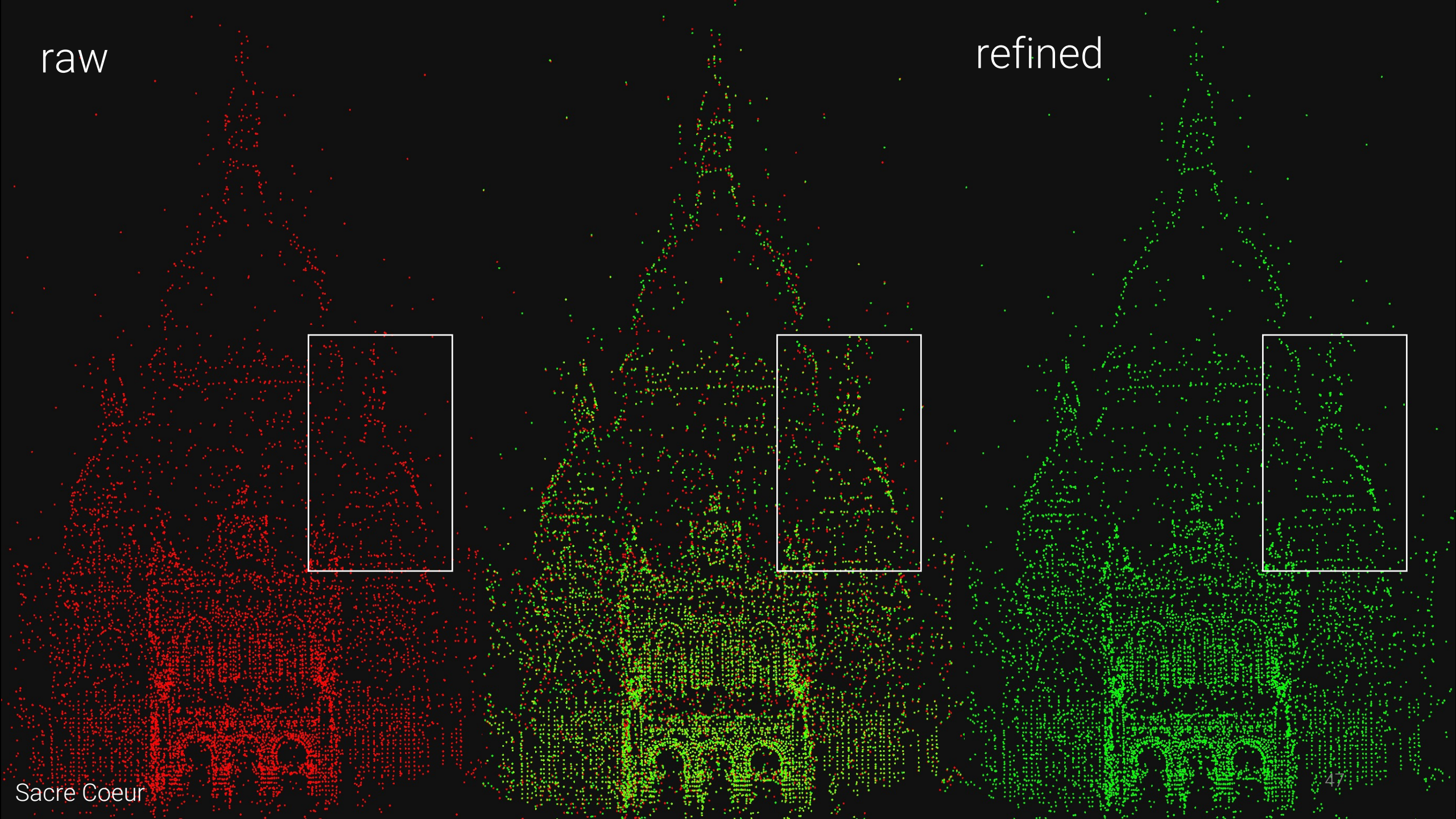


4x upsampled

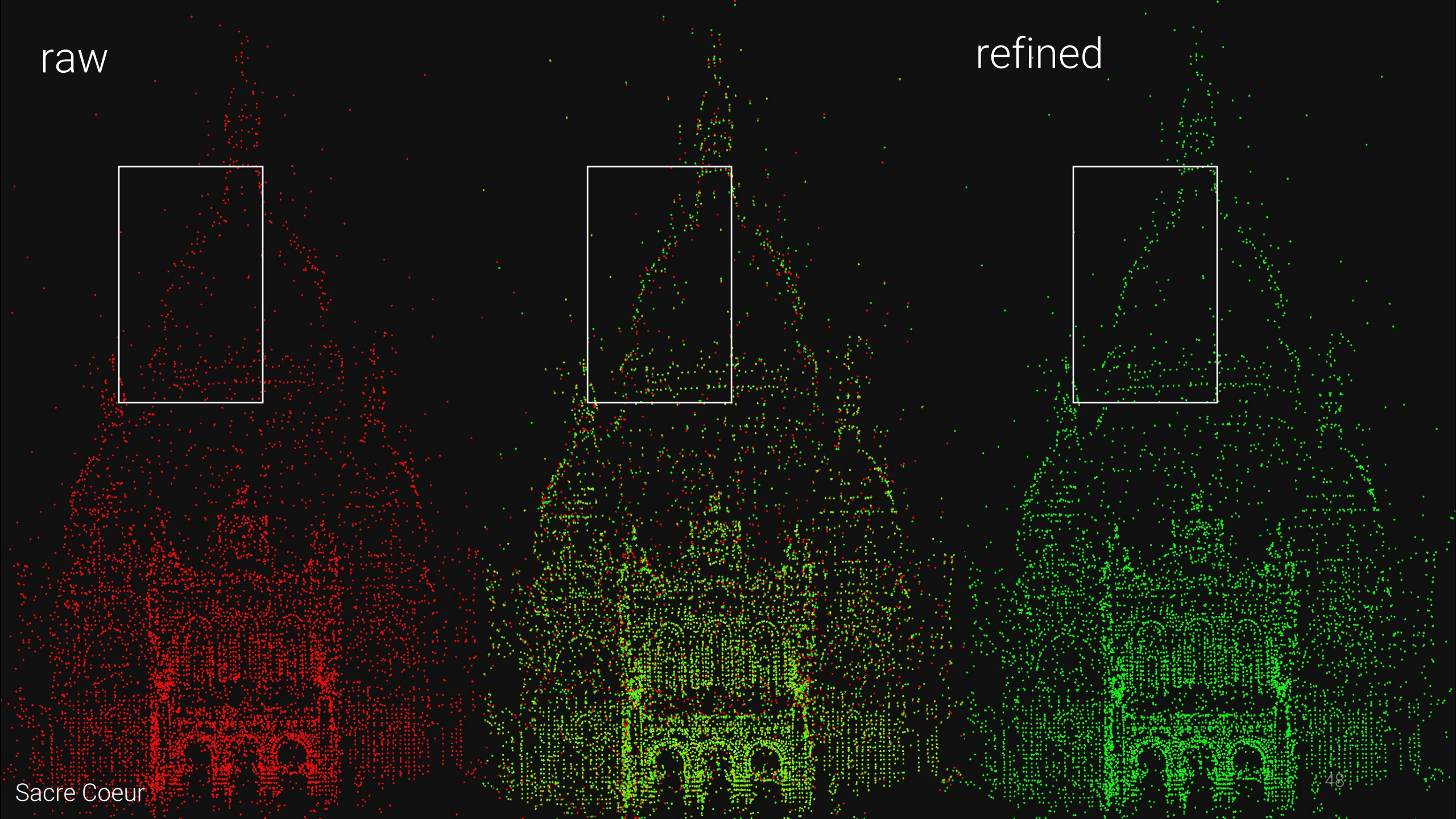




smooth gradients → **large convergence**



Sacre Coeur

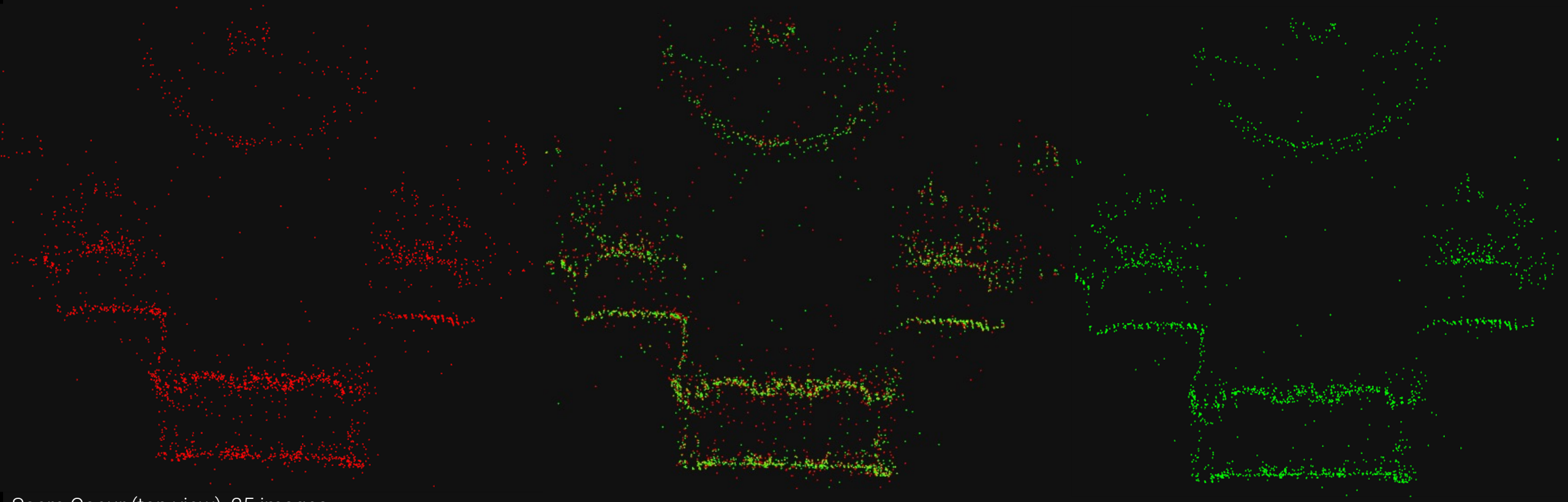


Sacre Coeur

Sacre Coeur (front view), 100 images

raw

refined



Sacre Coeur (top view), 25 images

Large improvements on **planar structures**, e.g. walls

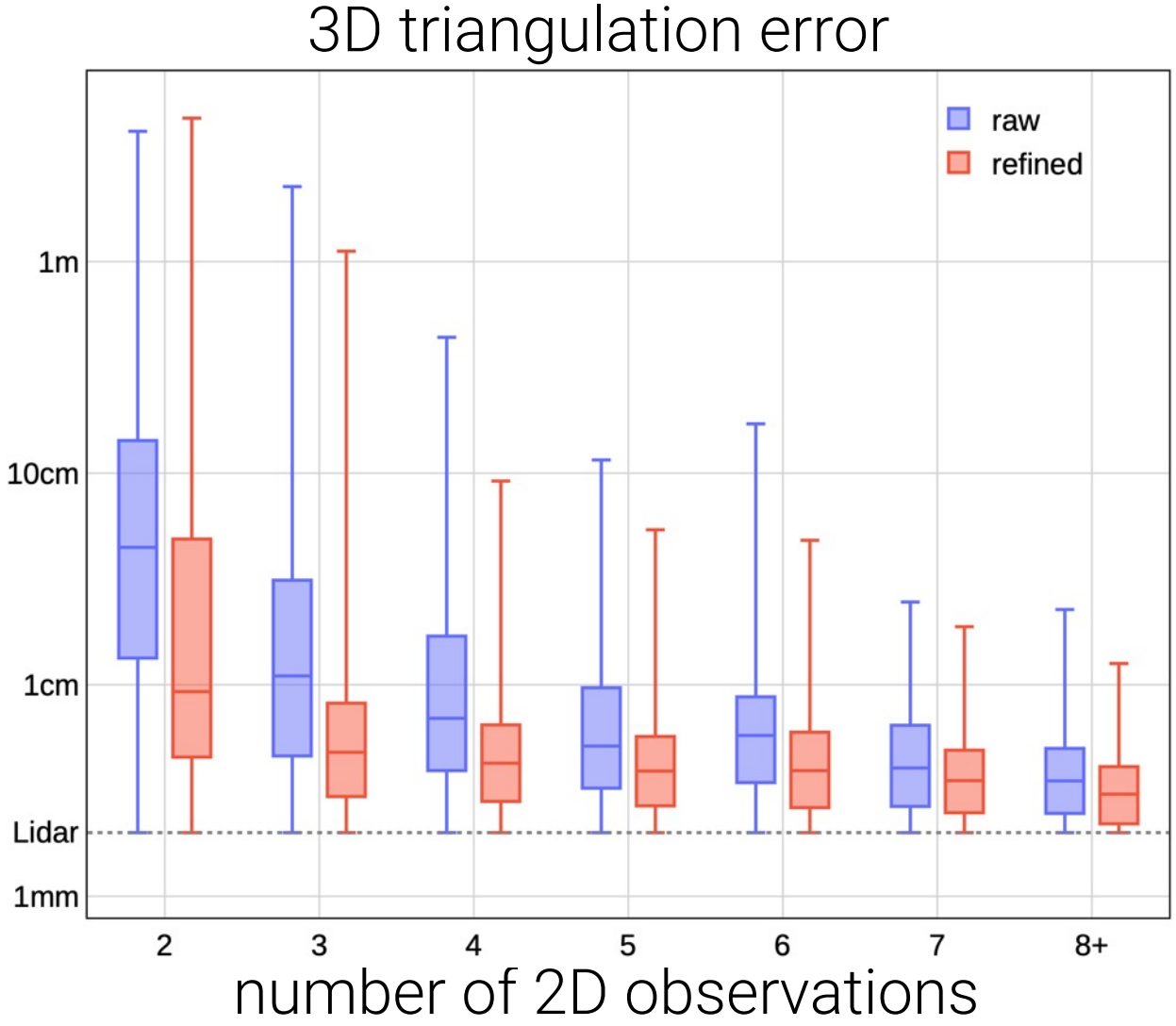
raw

refined

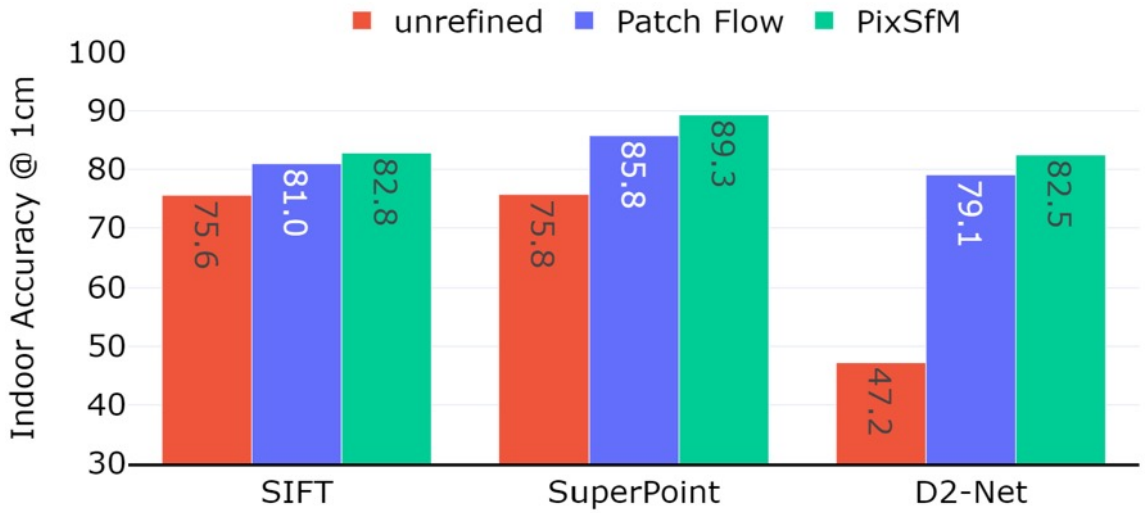
US Capitol (front- and top view), 50 images

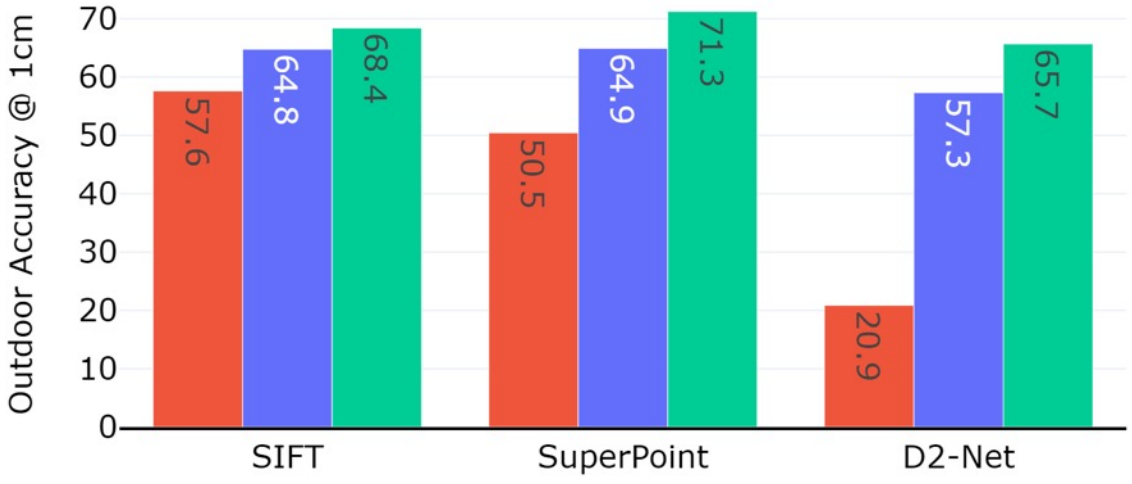
Accurate large-scale SfM

- few observations
  → **big improvements**
- **subpixel-accurate**
  large-scale SfM



Results on ETH3D Triangulation

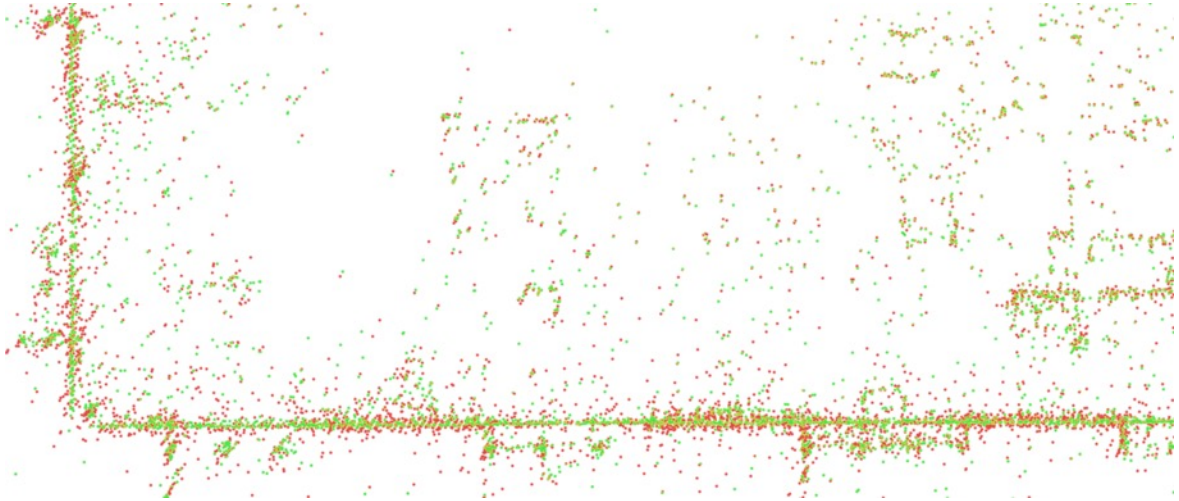




✅ **10-45% higher accuracy**

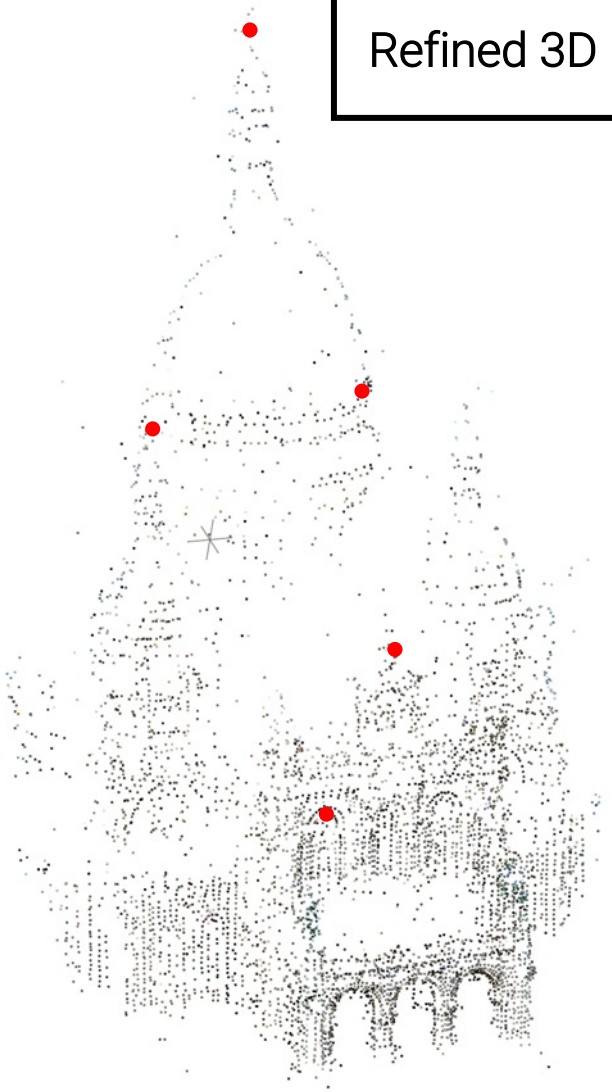
✅ Improves over **PatchFlow** by **>5%**

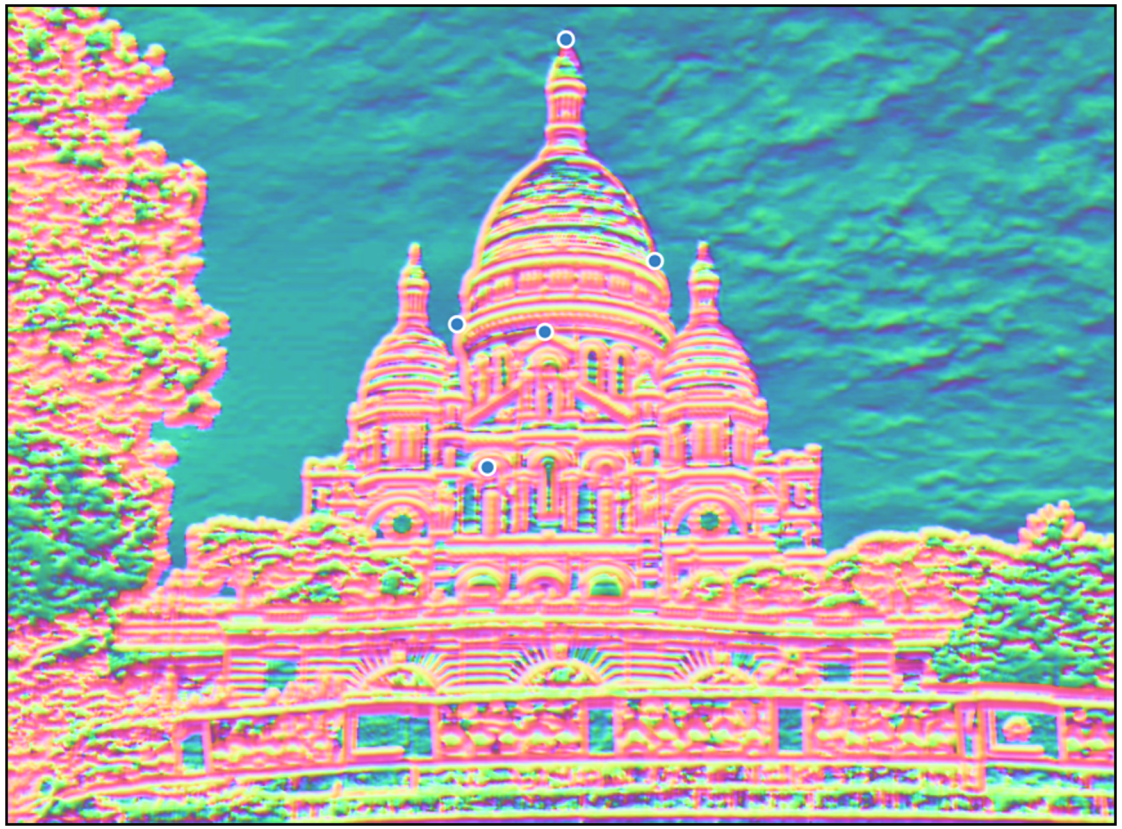
✅ improves **indoor** and **outdoor SfM**



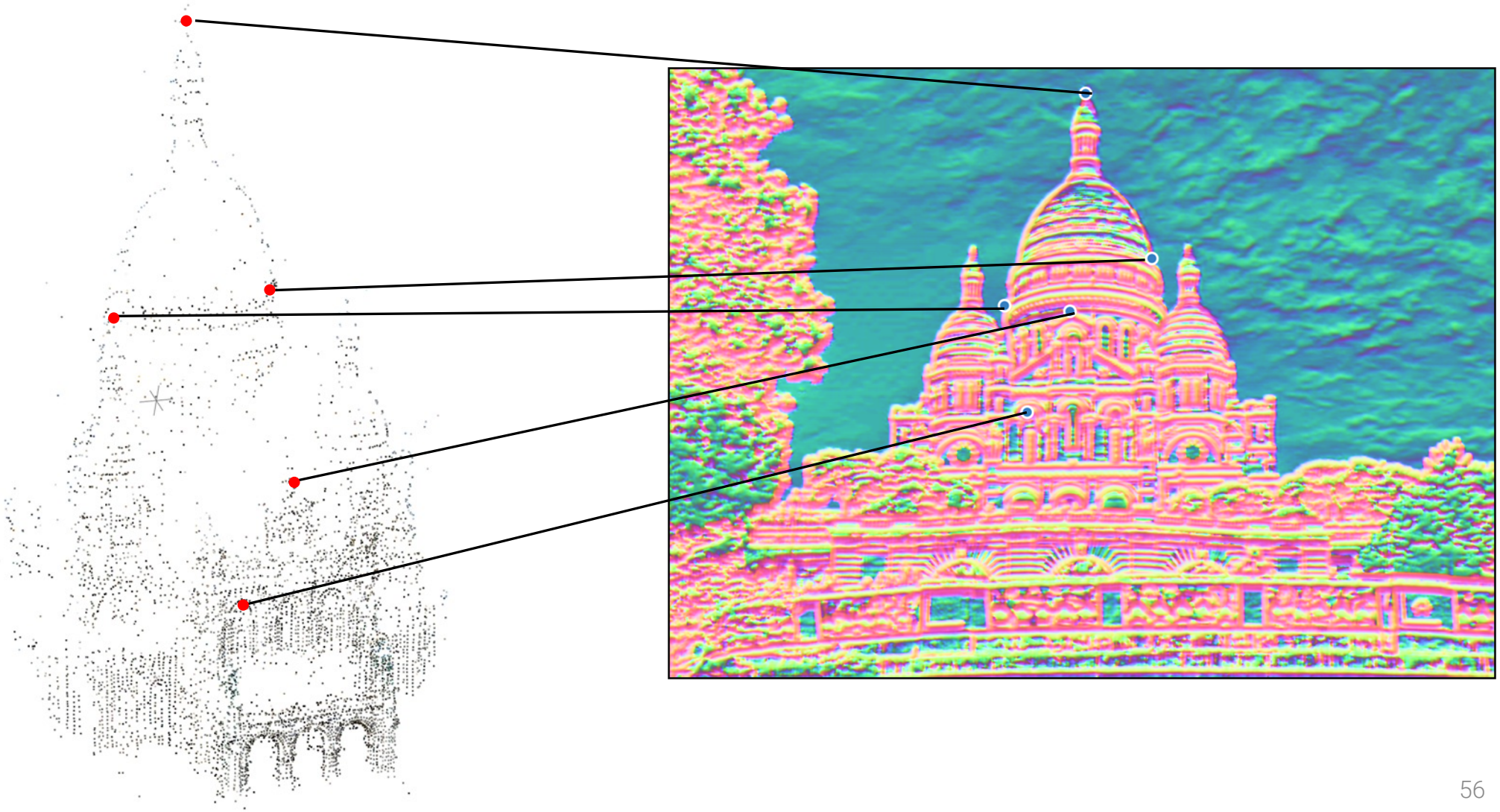
Camera Re-Localization - refinement

Refined 3D Model

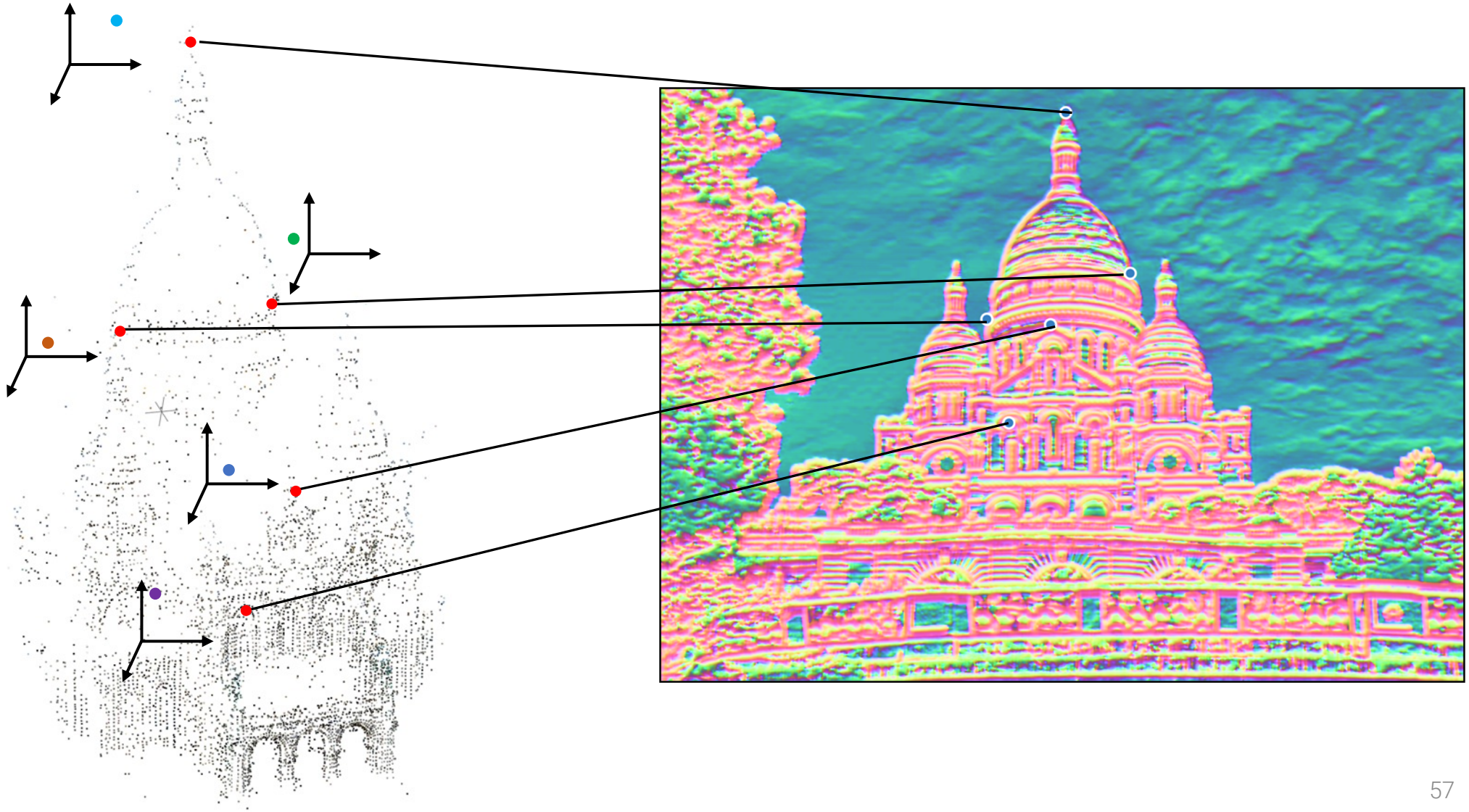




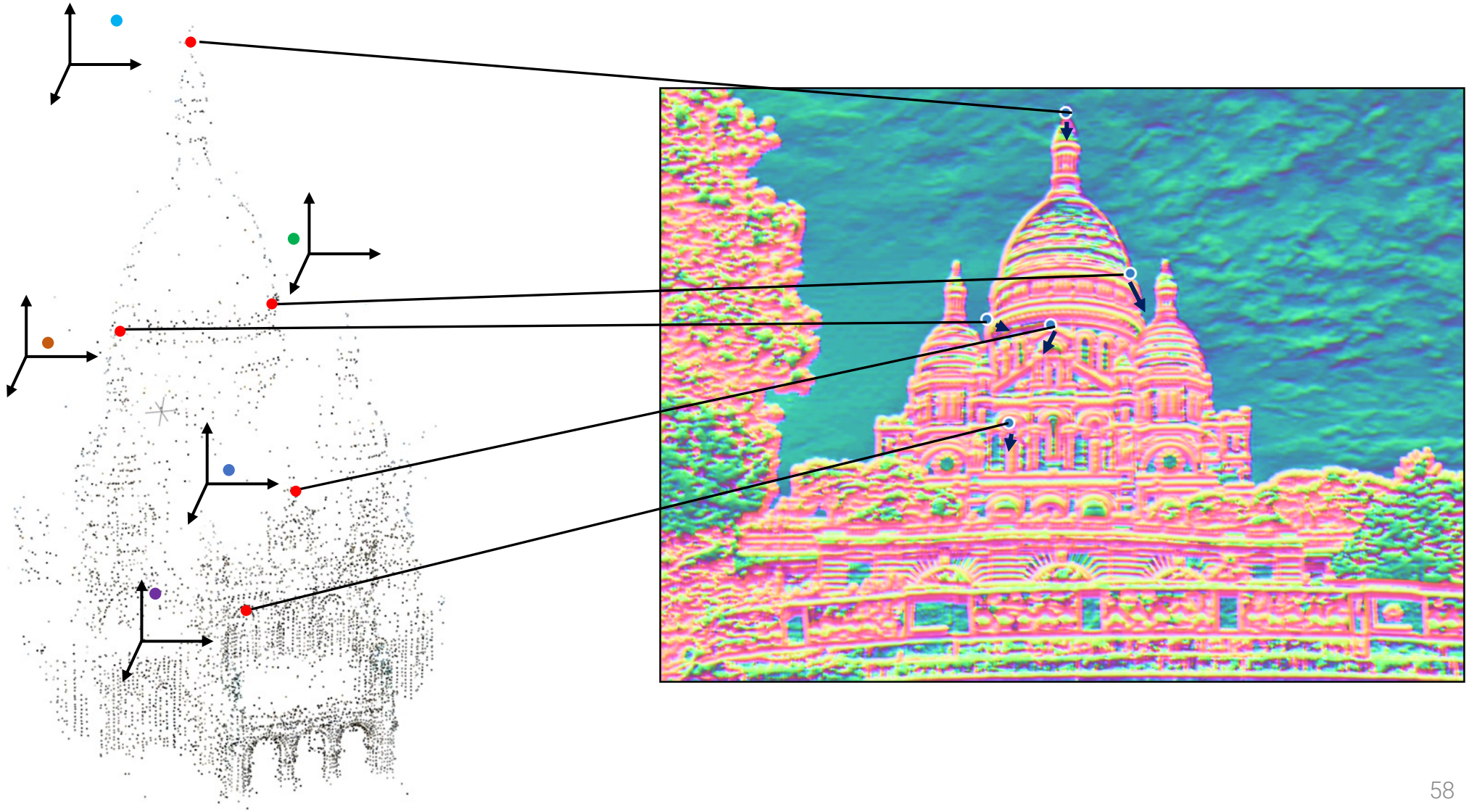
Camera Re-Localization - refinement



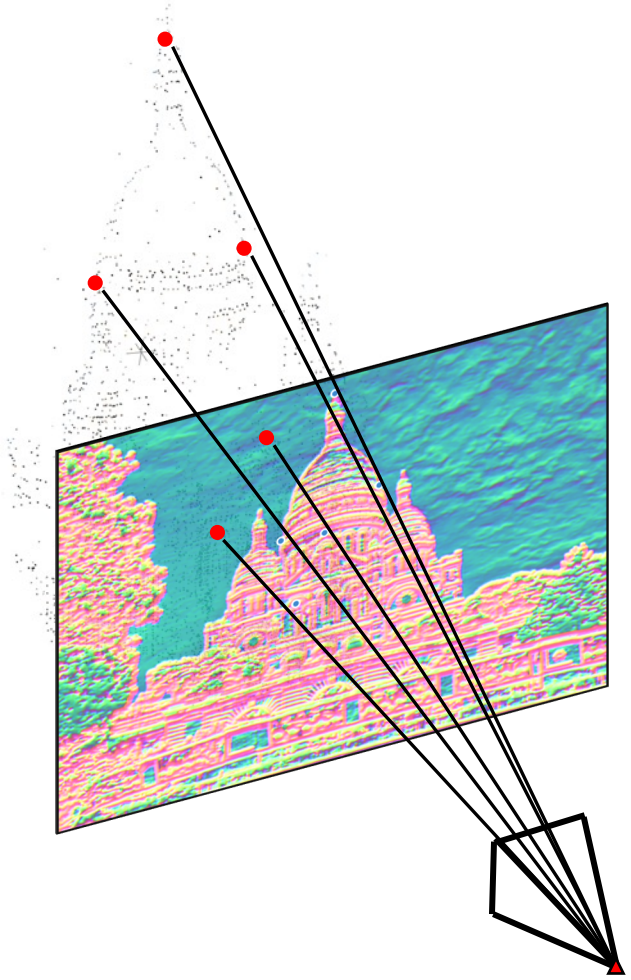
Camera Re-Localization - refinement



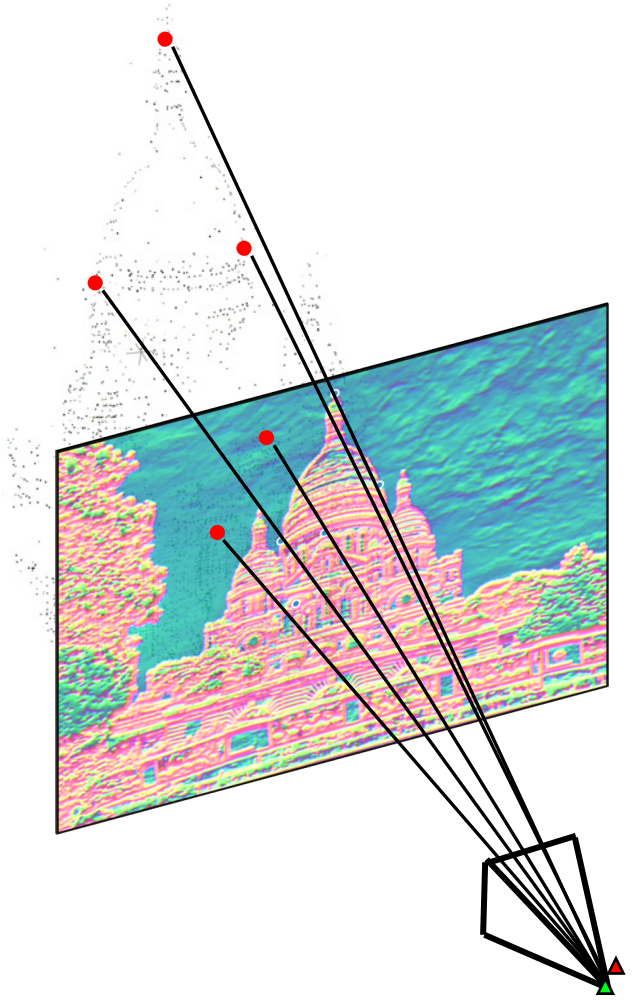
Camera Re-Localization - refinement



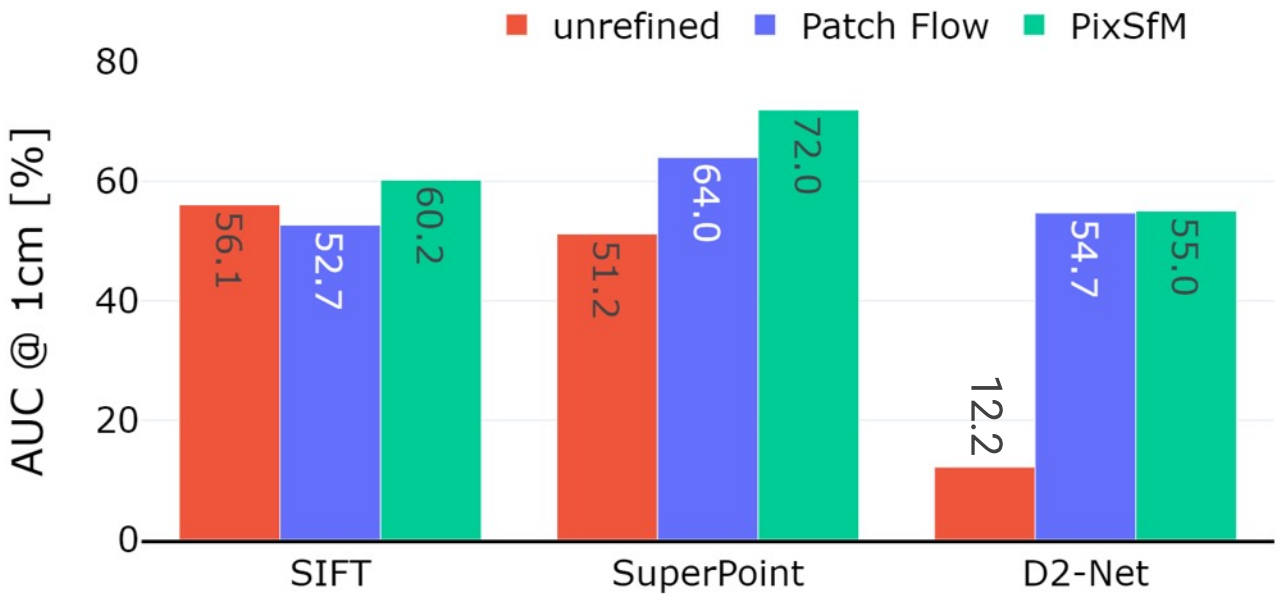
Camera Re-Localization - refinement



Camera Re-Localization - refinement



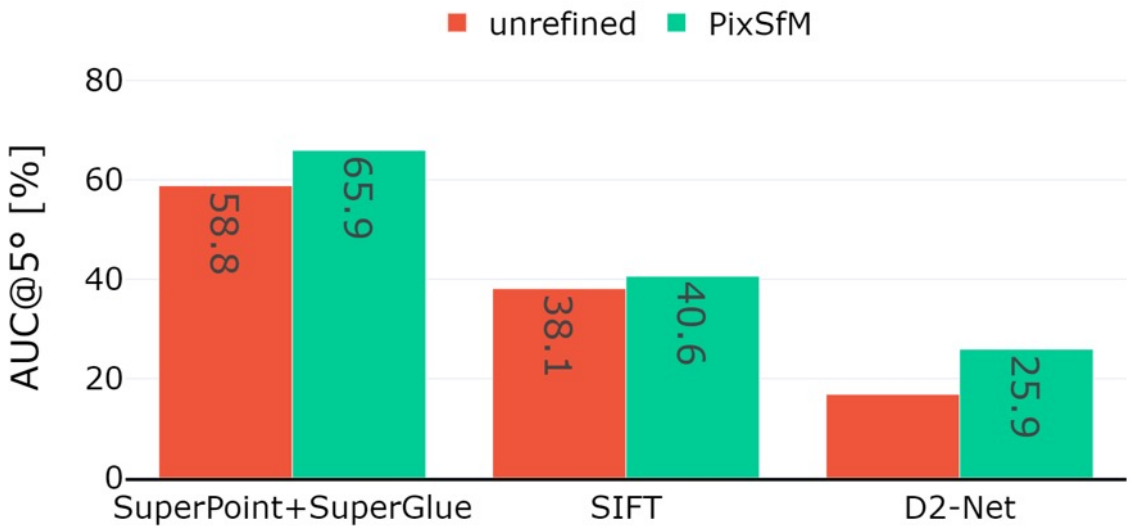
Results on **ETH3D Localization Benchmark**:



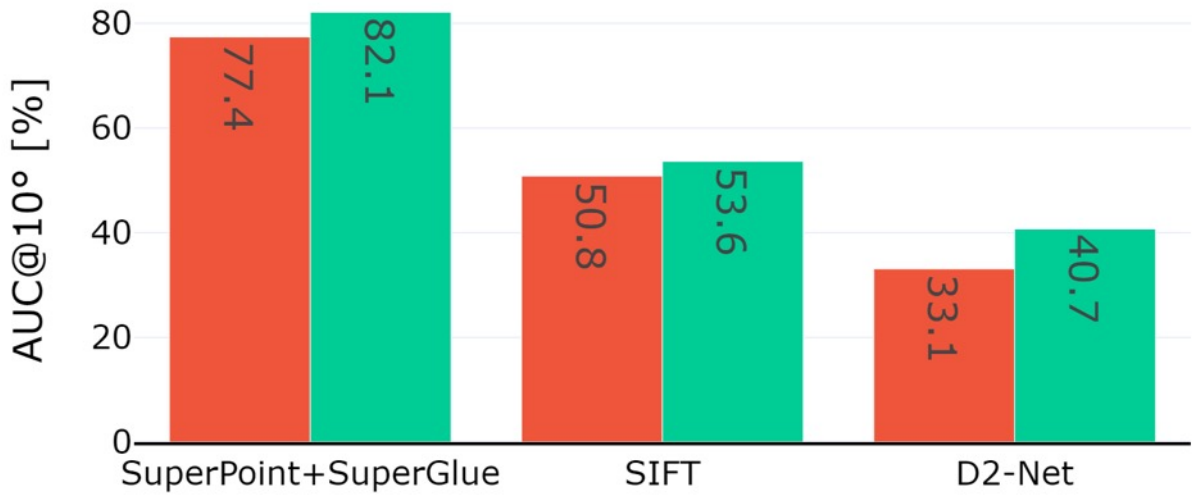
→ **Improvements** on **all features**

Phototourism – end-to-end SfM

Stereo:



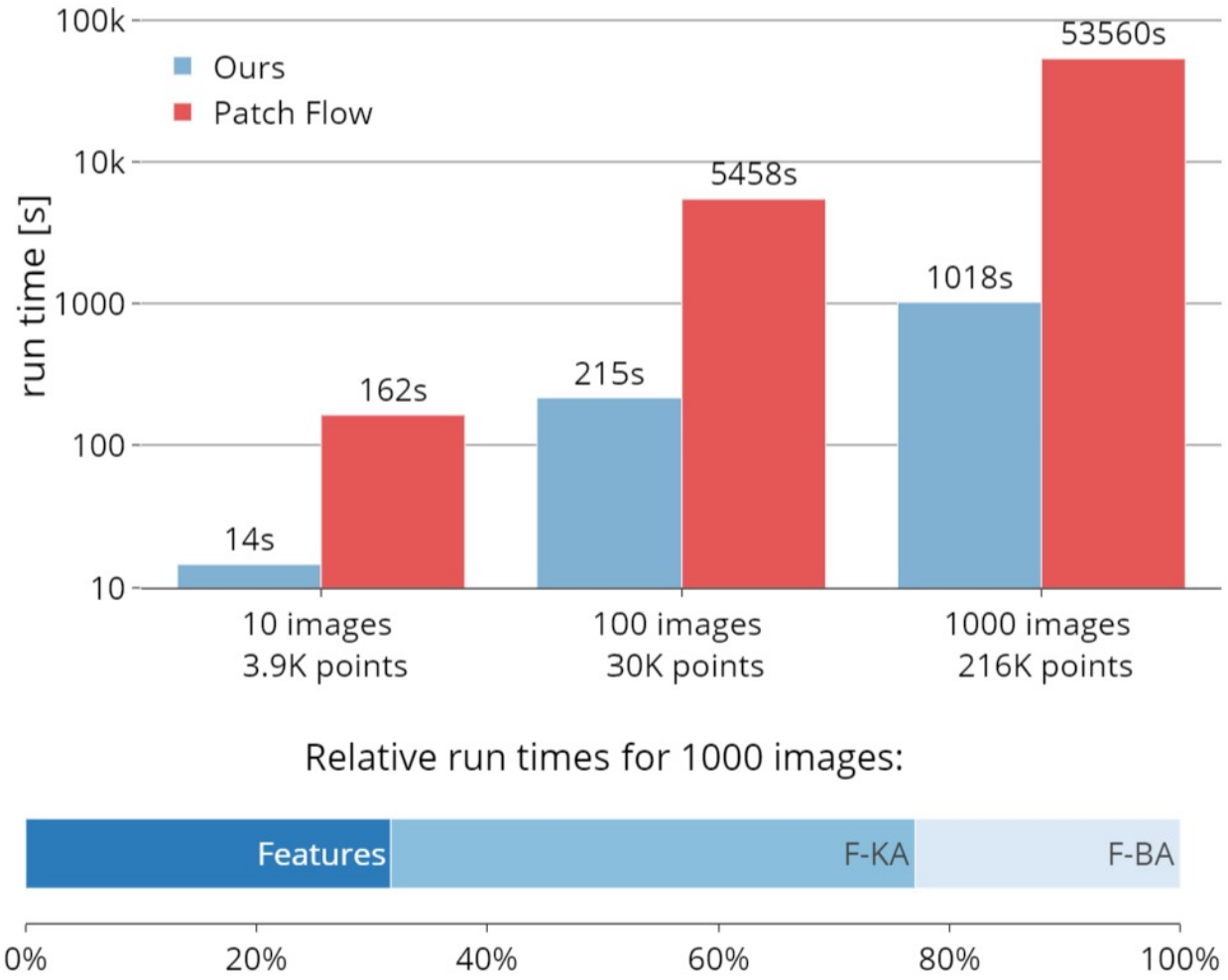
Multi-View:



Improves over SotA matching **SuperPoint+SuperGlue**

more **accurate small scenes** (5-25 images)

Runtime and Memory



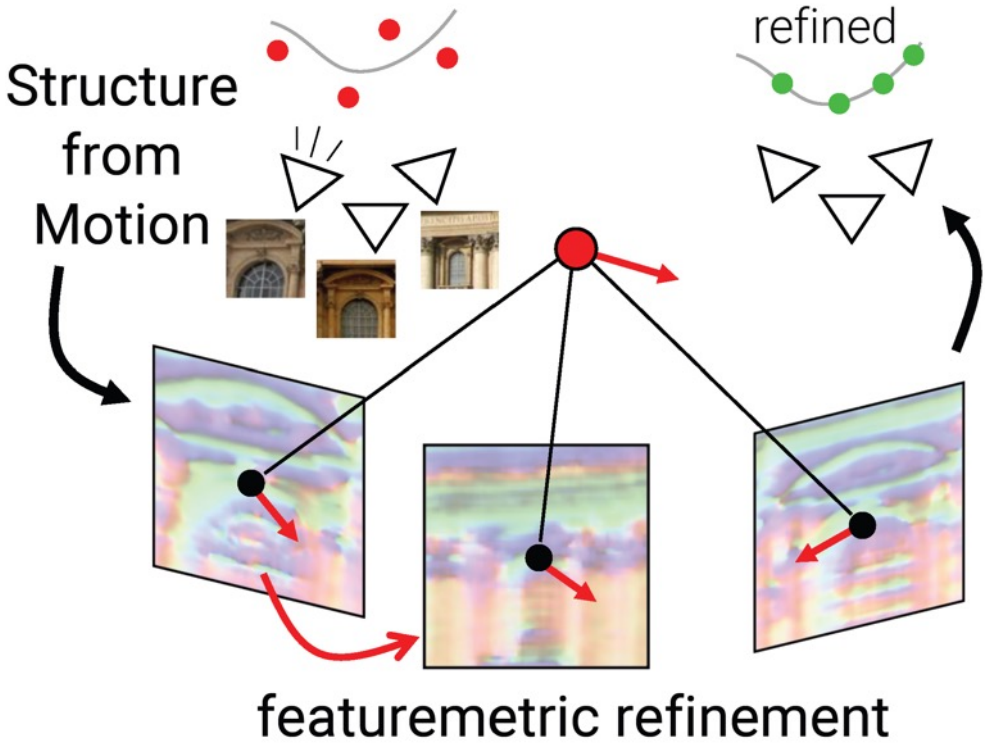
✅ Up to 50x **faster** than Patch Flow

✅ <**20% run time overhead** over COLMAP SfM

✅ Refine **thousands of images** on a standard PC

Pixel-Perfect Structure-from-Motion with Featuremetric Refinement

psarlin.com/pixsfm
github.com/cvg/pixel-perfect-sfm



Refine **keypoints + sparse SfM**
in a **few minutes!**