Learning to find good correspondences

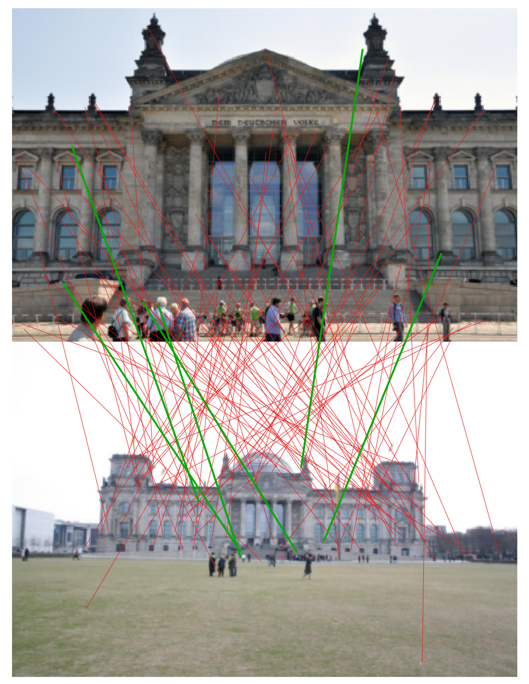
Kwang Moo Yi[1,*] Eduard Trulls[2,*], Yuki Ono[3], Vincent Lepetit[4], Mathieu Salzmann[2], Pascal Fua[2]

[1]Visual Computing Group, University of Victoria [2]Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne

[3]Sony Imaging Products & Solutions Inc. [4]Institute for Computer Graphics and Vision, Graz University of Technology (*: equal contribution)

CVPR 2018 SALT LAKE CITY • JUNE 18-22

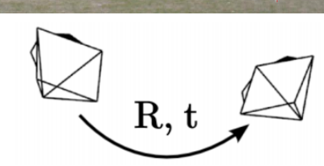
Contributions



We solve **sparse correspondence** with **deep networks.** Classical pipeline: (a) find putative matches (e.g. SIFT); (b) find inliers (e.g. RANSAC); (c) retrieve camera motion.

Our approach:

- **Input:** correspondences. **Output:** weights.
- Unordered data ⟶ **multi-layer perceptrons.**
- **Global context** from non-parametric units.
- **Hybrid loss:** joint classification & regression.
- **State of the art** results on indoors/outdoors.



Collecting the Ground truth

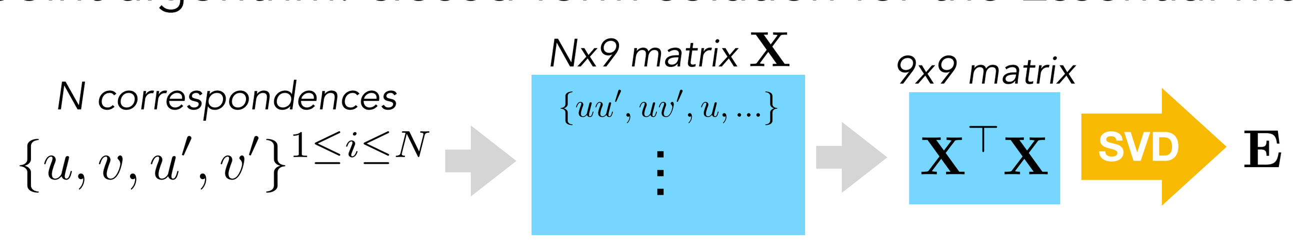
Indoors: RGB-D

Outdoors: SfM

Can't have pixel-to-pixel correspondences. We propose to use **only the pose as ground truth.** We can recover it with off-the-shelf SfM.

Learning with regression

- 8-point algorithm: closed-form solution for the Essential matrix:



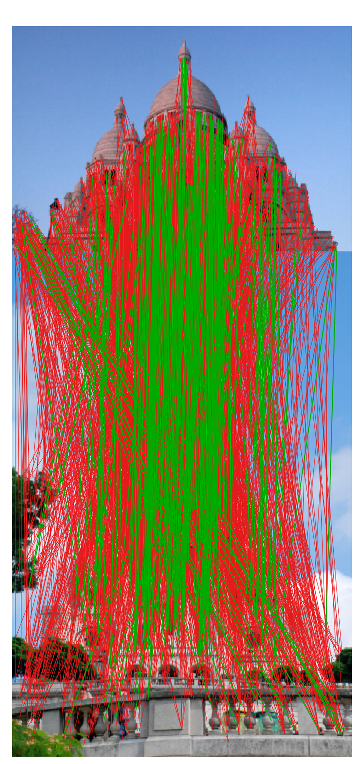
- Problem: **weak to outliers.** Solution: **weighted 8-point**, using the weights from the network: $\mathbf{X}^\top \mathtt{diag}(\mathbf{w})\mathbf{X}$. Fully differentiable.

$$\mathcal{L}_e(\Phi, \mathbf{x}_k) = \min\left\{\|\mathbf{E}_k^* \pm g\left(\mathbf{x}_k, \mathbf{w}_k\right)\|^2\right\}$$

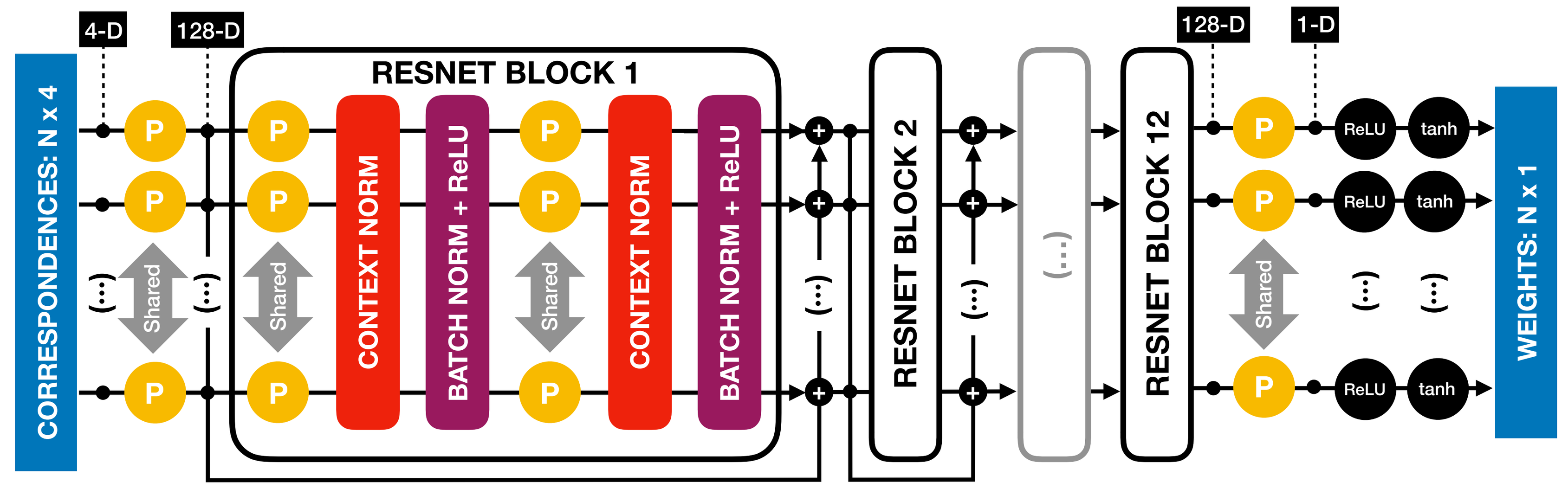
Learning with classification

- Learning outlier rejection implicitly by regressing the pose is too hard. **Network does not converge.**
- Solution: create **training labels from epipolar constraints.** How? threshold over the symmetric epipolar distance. Noisy but good enough!
- Loss: standard binary cross-entropy.

$$\mathcal{L}_x(\Phi, \mathbf{x}_k) = \frac{1}{N}\sum_{i=1}^{N}\gamma_k^i H\left(y_k^i, S\left(o_k^i\right)\right)$$



Outlier Rejection Network



- **Input:** N correspondences $\{u, v, u', v'\}^{1 \le i \le N}$. **Output:** N weights.
- **Problem:** input data is unordered. Output should be permutation-invariant. Not feasible with e.g. convolutional or fully-connected layers.
- **Solution (PointNet):** Multi-layer, weight-sharing perceptrons.
- Deep network: 12 resnet-style blocks. Still **very small!**
- Each point is processed individually! We need contextual information. PointNet solution: **global feature,** pooled with a second network.
- **Our solution:** embed into the feature maps with **Context Normalization.**

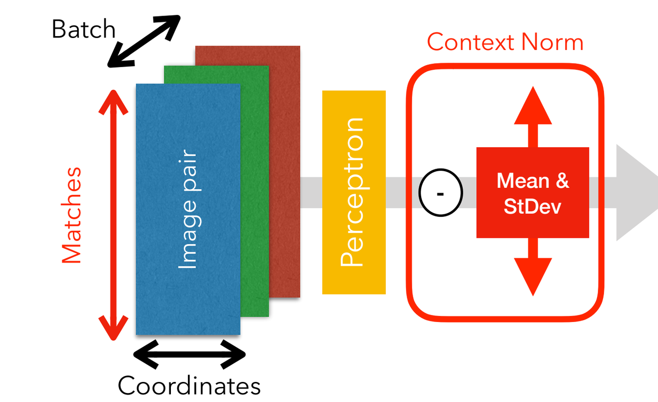
Context Normalization

- Simple, non-parametric normalization. Given features $\mathbf{o}^l_{1 \le i \le N}$ at layer l:

$$\mathrm{CN}\left(\mathbf{o}_i^l\right) = \frac{\left(\mathbf{o}_i^l - \mu^l\right)}{\sigma^l}$$

$$\mu^l = \frac{1}{N}\sum_{i=1}^{N}\mathbf{o}_i^l\,, \quad \sigma^l = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{o}_i^l - \mu^l\right)^2}$$

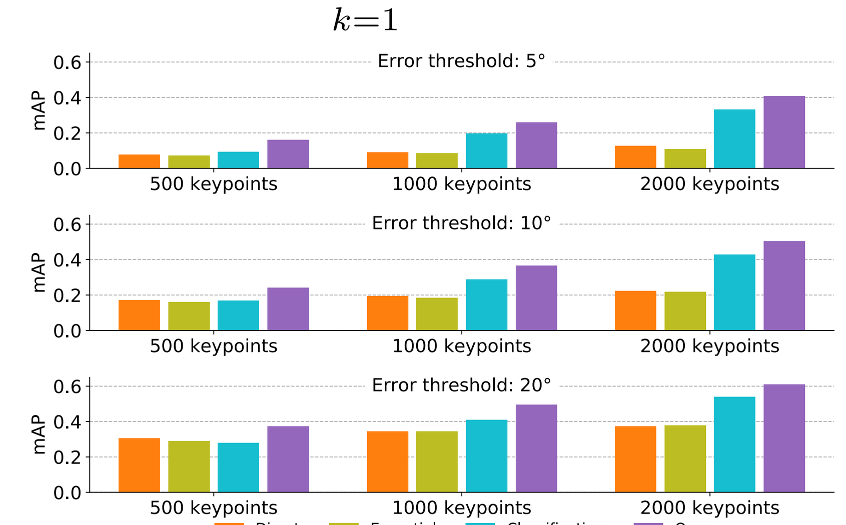
- Similar to BN/LN, but nothing is learned. Same operation for training/inference.
- Operates separately over image pairs:



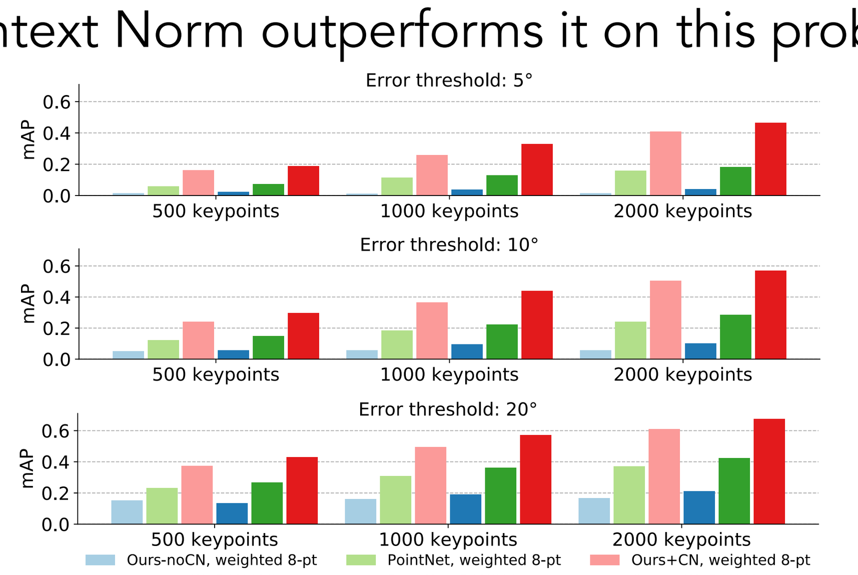
- In image stylization: **Instance Norm.**

Ablation: Loss & Context

- Classification required to converge. Hybrid loss does best: $\mathcal{L}(\Phi) = \sum_{k=1}^{P}(\alpha\mathcal{L}_x(\Phi, \mathbf{x}_k) + \beta\mathcal{L}_e(\Phi, \mathbf{x}_k))$

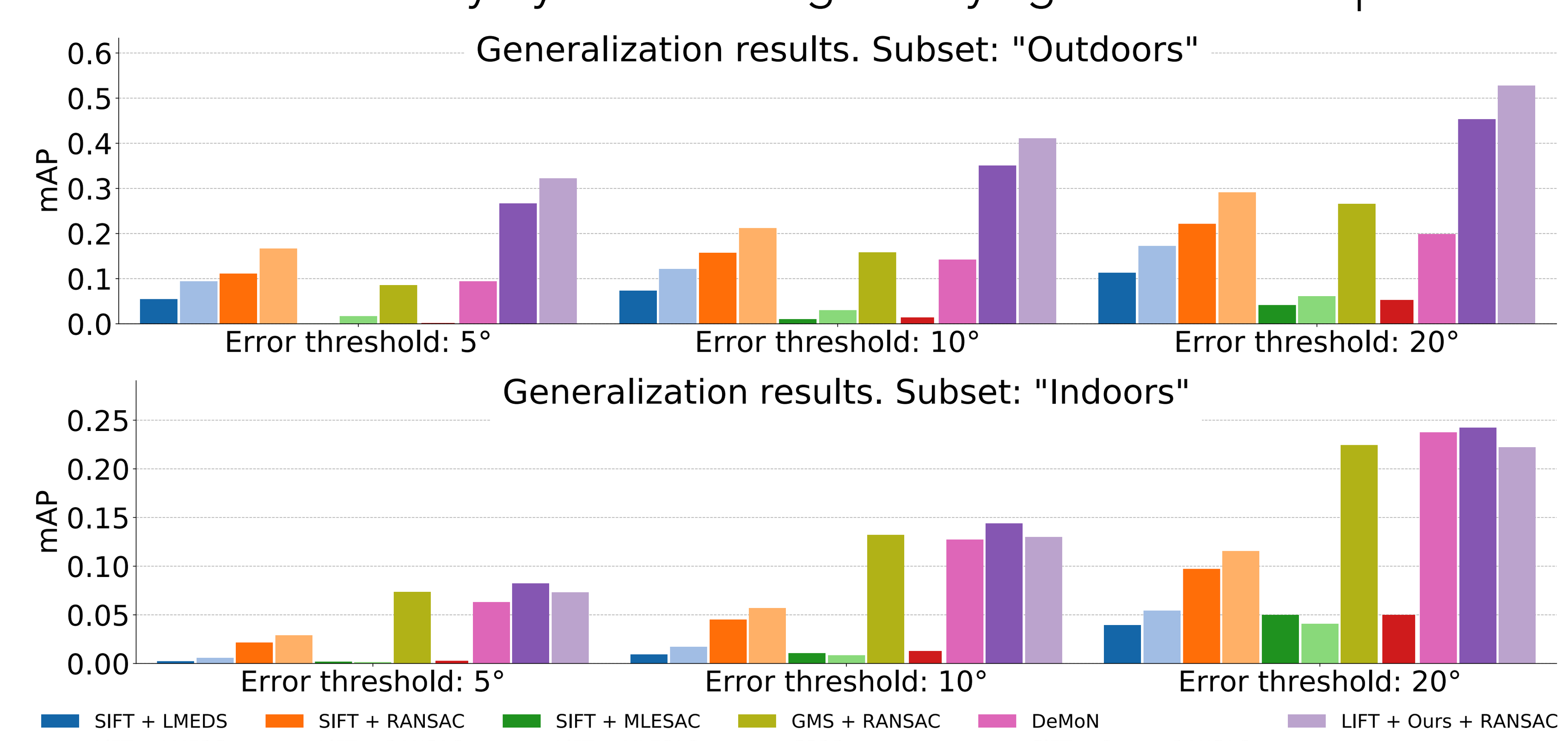


- PointNet-style context works, but our simple Context Norm outperforms it on this problem.



Evaluation

- **Datasets:** indoors (SUN3D) and outdoors (YFCC100+SfM).
- Our models are trained on a **single sequence from each.**
- **Baselines:** sparse (RANSAC variants, GMS) & dense (G3DR, DeMoN).
- **Metric:** angular error between ground-truth & estimated R/T. Determine accuracy by thresholding at varying values & compute mAP.



- **Outdoors:** great improvements. **Indoors:** still better than dense SoA.
- For testing we do not need differentiability! We run **ours (one forward pass) then RANSAC.** Improves performance (2x) and speed (17x!).

Qualitative results

- Top: **RANSAC.** Bottom: **Ours.** Same input.
- Drawing **inliers only.** Pictured in **green** if they are below the ground truth epipolar distance threshold, and in **red** otherwise.