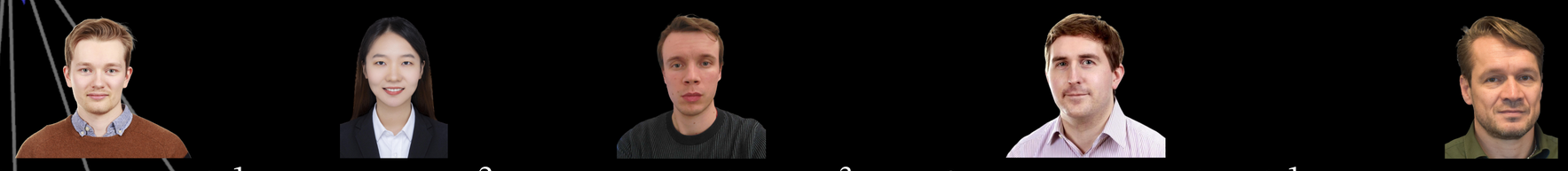
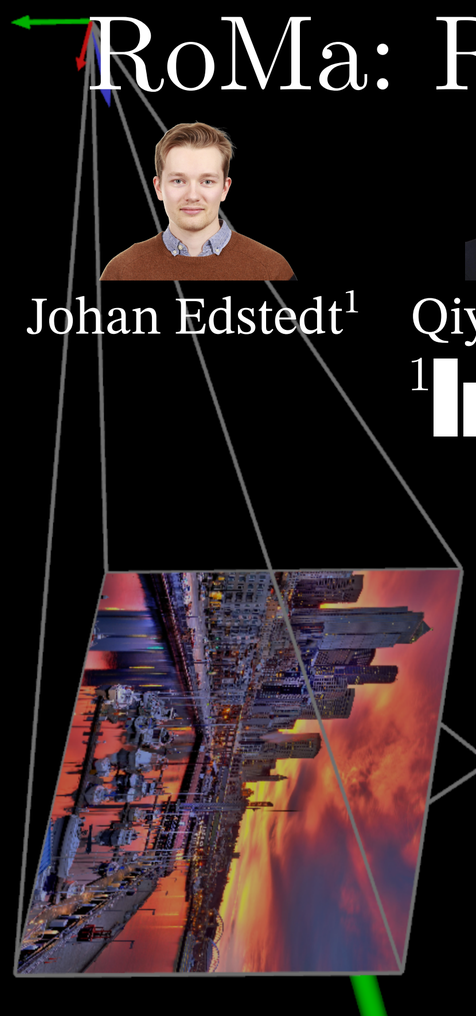
RoMa: Robust Dense Feature Matching



Johan Edstedt[1] Qiyu Sun[2] Georg Bökman[3] Mårten Wadenbäck[1] Michael Felsberg[1]

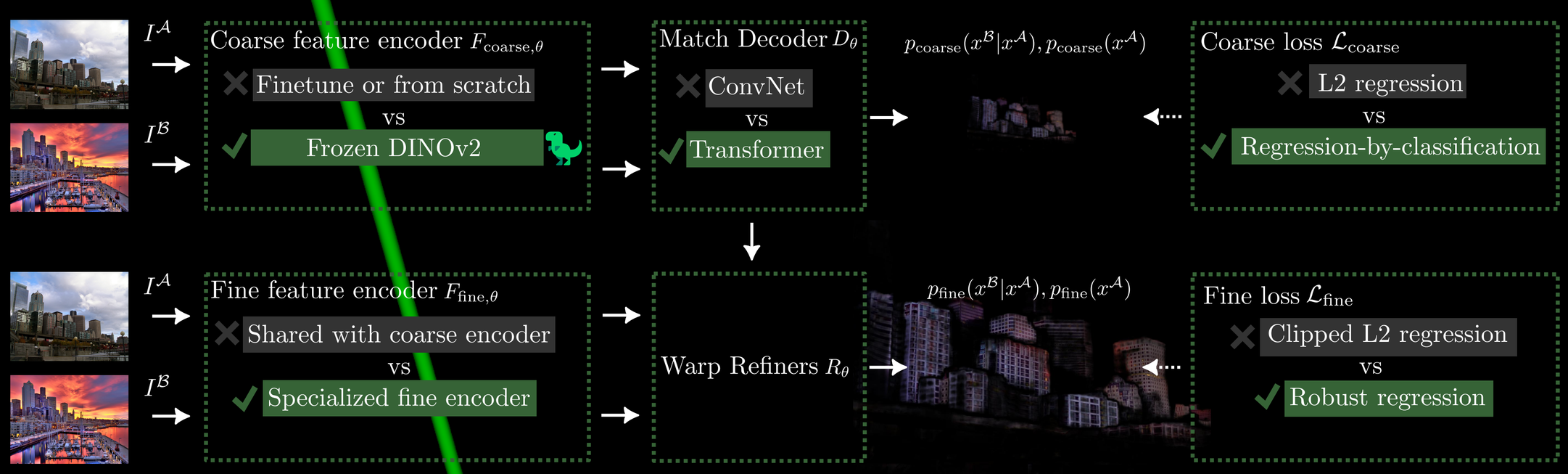
[1] LINKÖPING UNIVERSITY [2] EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY [3] CHALMERS UNIVERSITY OF TECHNOLOGY



Summary:
RoMa estimates dense correspondences between almost any two images of a 3D scene

github.com/parskatt/roma

Method



Insights

- Fine feature encoder $F_{\text{fine},\theta}$ — ✓ Specialized fine encoder
  Insight: Fine localizability and coarse discriminativity are in tension, specializing the fine features help
- Match Decoder D_θ — ✓ Transformer
  Transforms match encodings (e.g. corr. vol.) to predictions
  No position encodings (avoids oversmoothing)
  Predicts anchor probabitilies $p_{\text{coarse},\theta}(x^{\mathcal{B}}|x^{\mathcal{A}}) = \sum_{k=1}^{K} \pi_k(x^{\mathcal{A}})\mathcal{B}_{m_k}$
- Coarse feature encoder $F_{\text{coarse},\theta}$ — ✓ Frozen DINOv2
  DINOv2 features discriminative enough to freeze
  Enables our matcher to generalize significantly better
- Coarse loss $\mathcal{L}_{\text{coarse}}$ — ✓ Regression-by-classification
  $$\mathcal{L}_{\text{coarse}} = \sum_{x^{\mathcal{A}},x^{\mathcal{B}}\in\mathcal{C}} -\log \pi_{k^{\mathcal{B}}}(x^{\mathcal{A}}) - \lambda \log p_{\text{coarse},\theta}(x^{\mathcal{A}})$$
- Fine loss $\mathcal{L}_{\text{fine}}$ — ✓ Robust regression
  $$\mathcal{L}_{\text{fine}} = \sum_i \sum_{x^{\mathcal{A}},x^{\mathcal{B}}\in\mathcal{C}} (||\mu_\theta(x_i^{\mathcal{A}}, \hat{W}_{i+1}^{\mathcal{A}\to\mathcal{B}}) - x_i^{\mathcal{B}}||^2 + s^2)^{1/4} - \lambda \log p_{i,\theta}(x_i^{\mathcal{A}}|\hat{W}_{i+1}^{\mathcal{A}\to\mathcal{B}})$$

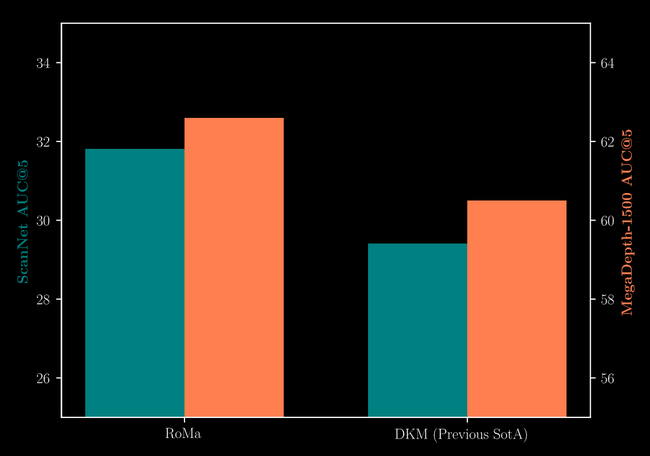
Setup ↓	100-PCK@ →	1px	3px	5px
I (Baseline): DKM		17.0	7.3	5.8
II: I, $F_{\text{coarse},\theta}$ = RN50, $F_{\text{fine},\theta}$ = RN50		16.0	6.1	4.5
III: II, $F_{\text{fine},\theta}$ = VGG19		14.5	5.4	4.5
IV: III, D_θ = Transformer		14.4	5.4	4.1
V: IV, $F_{\text{coarse},\theta}$ = DINOv2		14.3	4.6	3.2
VI: V, $\mathcal{L}_{\text{coarse}}$ = reg.-by-class.		13.6	4.1	2.8
VII (Ours): VI, $\mathcal{L}_{\text{refine}}$ = robust		**13.1**	**4.0**	**2.7**
VIII: VII, D_θ = ConvNet		14.0	4.9	3.5

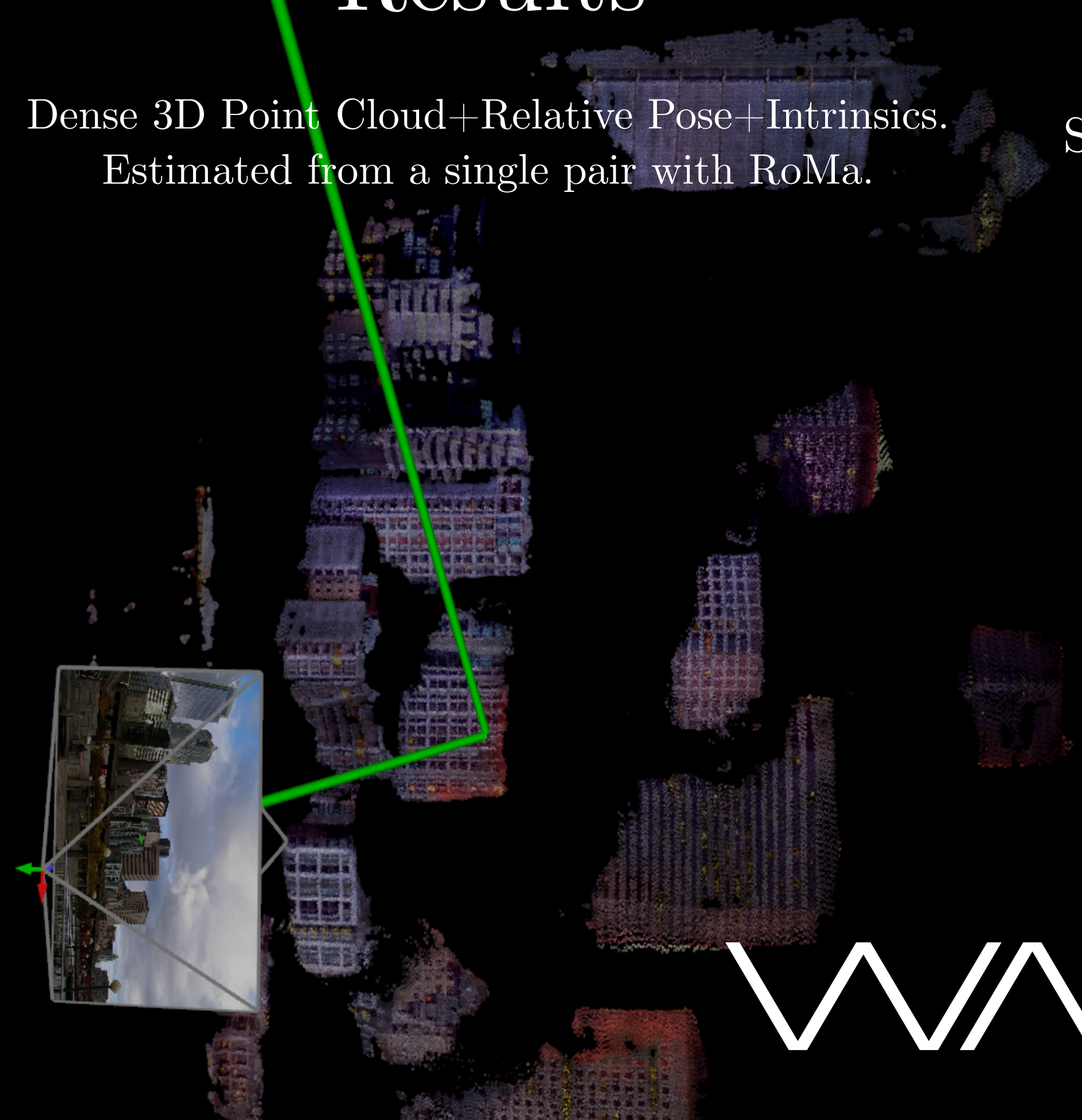
Results

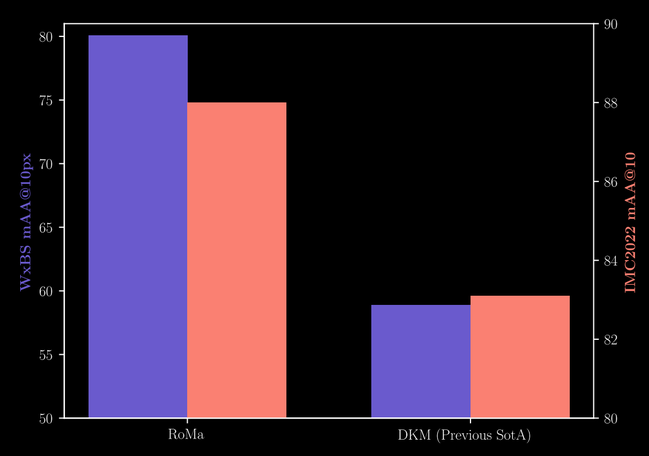
SotA on standard benchmarks.

Dense 3D Point Cloud+Relative Pose+Intrinsics.
Estimated from a single pair with RoMa.

Shines on difficult benchmarks!







ELLIIT Excellence Center at Linköping – Lund in Information Technology

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM