RoMa: Robust Dense Feature Matching

Supplementary Material

In this supplementary material, we provide further details and qualitative examples that could not fit into the main text of the paper.

A. Limitations and Future Work

(a) Our approach relies on supervised correspondences, which limits the amount of usable data. We remedied this by using pretrained frozen foundation model features, which improves generalization.
(b) We train on the task of dense feature matching which is an indirect way of optimizing for the downstream tasks of two-view geometry, localization, or 3D reconstruction. Directly training on the downstream tasks could improve performance.

B. Frozen Feature Evaluation

We use an exponential cosine kernel as in DKM [17] with an inverse temperature of 10. We train using the same training split as in our main experiments, using the same learning rates (note that we only train a single linear layer, as the backbone is frozen). We use the regression-by-classification loss that we proposed in Section 3.4. We present a qualitative example of the estimated warps from the frozen features in Figure 5.

C. Architectural Details

Encoders: We extract fine features of stride $\{1, 2, 4, 8\}$ by taking the outputs of the layer before each 2×2 maxpool. These have dimension $\{64, 128, 256, 512\}$ respectively. We project these with a linear layer followed by batchnorm to dimension $\{9, 64, 256, 512\}$.

We use the patch features from DINOv2 [39] and do not use the `cls` token. We use the ViT-L-14 model, with patch size 14 and dimension 1024. We linearly project these features (with batchnorm) to dimension 512.

Global Matcher: We use a Gaussian Process [40] match encoder as in DKM [17]. We use an exponential cosine kernel [17], with inverse temperature 10. As in DKM, the GP predicts a posterior over embedded coordinates in the other image. We use an embedding space of dimension 512.

For details on D_θ we refer to Section 3.3.

Refiners: Following Edstedt et al. [17] we use 5 refiners at strides $\{1, 2, 4, 8, 14\}$. They each consist of 8 convolutional blocks. The internal dimension is set to $\{24, 144, 569, 1137, 1377\}$. The input to the refiners are the stacked feature maps, local correlation around the previous warp of size $\{0, 0, 5, 7, 15\}$, as well as a linear encoding of the previous warp. The output is a $B \times H_s \times W_s \times (2 + 1)$ tensor, containing the warp and an logit offset to the certainty.

D. Qualitative Comparison on WxBS

We qualitatively compare estimated matches from RoMa and DKM on the WxBS benchmark in Figure 6. DKM fails on multiple pairs on this dataset, while RoMa is more robust. In particular, RoMa is able to match even for changes is season (bottom right), extreme illumination (bottom left, top left), and extreme scale and viewpoint (top right).

E. Metrics

Image Matching Challenge 2022: The mean average accuracy (mAA) metric is computed between the estimated fundamental matrix and the hidden ground truth. The error in terms of rotation in degrees and translation in meters. Given one threshold over each, a pose is classified as accurate if it meets both thresholds. This is done over ten pairs of uniformly spaced thresholds. The mAA is then the average over the threshold and over the images (balanced across the scenes).

MegaDepth/ScanNet: The AUC metric used measures the error of the estimated Essential matrix compared to the ground truth. The error per pair is the maximum of the rotational and translational error. As there is no metric scale available, the translational error is measured in the cosine angle. The recall at a threshold τ is the percentage of pairs with an error lower than τ. The AUC@τ° is the integral over the recall as a function of the thresholds, up to τ, divided by τ. In practice, this is approximated by the trapezial rule over all errors of the method over the dataset.

F. Details on Ablation Validation Set

The validation set is made from random pairs from the MegaDepth scenes $[0015, 0022]$ with overlap > 0 (where the overlap is defined as IoU of SfM 3D tracks). To measure the performance we measure the percentage of estimated matches that have an end-point-error (EPE) under a certain pixel threshold over all ground-truth correspondences, which we call percent correct keypoints (PCK) using the notation of previous work [17, 56].

G. Loss

As in DKM [17] we set warp prediction loss to zero whenever the previous prediction has an error larger than some fixed threshold. For scales $\{1, 2, 4, 8\}$ we set this threshold

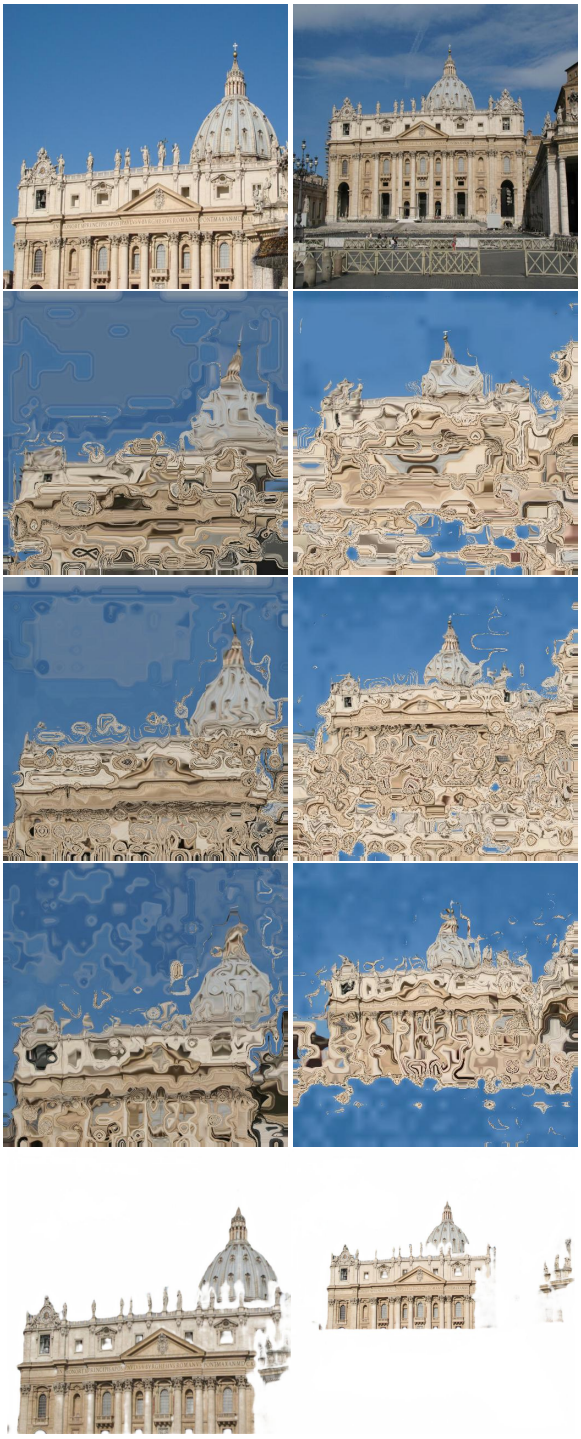

Figure 5. **Evaluation of frozen features.** From top to bottom: Image pair, VGG19 matches, RN50 matches, DINOv2 matches, RoMa matches. DINOv2 is significantly more robust than the VGG19 and RN50. Quantitative results are presented in Table 1.

to $\{4, 8, 16, 32\}$ pixels at a resolution of 560 pixels respectively. In addition, we also set the target for the confidence

Table 9. **SotA comparison on Aachen v1.1 [46].** Measured in AUC (higher is better). HLoc [43] is used for all methods.

Method ↓	Day	Night
SP+SG	90.3 / 96.5 / 99.3	75.9 / 91.1/ **100.0**
LoFTR	88.7 / 95.6 / 99.0	**78.5** / 90.6 / 99.0
PATS	89.6 / 95.8 / 99.3	73.8 / 92.1 / 99.5
RoMa	**90.9 / 96.5 / 99.4**	**78.5 / 92.7 / 100.0**

loss in these regions to be zero.

It can be noted here that this heuristic can in itself be regarded as a type of robust loss, as we set the loss for large outliers to zero. Our generalized Charbonnier loss can be seen as a more smooth version of this. We found that combining the Charbonnier loss with clipping yielded the best results.

Furthermore, as in DKM, the usage of binary cross-entropy in the loss functions requires us to include $x^{\mathcal{A}} \notin \mathcal{C}$, as the marginal could otherwise trivially be maximized. Like in DKM we choose all $x^{\mathcal{A}}$ on the image grid.

H. Theoretical Model

Here we discuss a simple connection to scale-space theory, that did not fit in the main paper. Our theoretical model of matchability in Section 3.4 has a straightforward connection to scale-space theory [27, 32, 63]. The image scale-space is parameterized by a parameter s,

$$L(x, s) = \int g(x - y; s) I(y) dy, \tag{22}$$

where

$$g(x; s) = \frac{1}{2\pi s^2} \exp\left(-\frac{1}{2s^2} \|x\|^2 \right) \tag{23}$$

is a Gaussian kernel. Applying this kernel jointly on the matching distribution yields the diffusion process in the paper.

I. Further Details on Match Sampling

Dense feature matching methods produce a dense warp and certainty. However, most robust relative pose estimators (used in the downstream two-view pose estimation evaluation) assume a sparse set of correspondences. While one could in principle use all correspondences from the warp, this is prohibitively expensive in practice. We instead follow the approach of DKM [17] and use a balanced sampling approach to produce a sparse set of matches. The balanced sampling approach uses a KDE estimate of the match distribution $p_\theta(x^{\mathcal{A}}, x^{\mathcal{B}})$ to rebalance the distribution of the samples, by reweighting the samples with the reciprocal of the

Figure 6. **Qualitative comparison.** RoMa is significantly more robust to extreme changes in viewpoint and illumination than DKM.

KDE. This increases the number of matches in less certain regions, which Edstedt et al. [17] demonstrated improves performance.

J. Runtime Comparison

We compare the runtime of RoMa and the baseline DKM at a resolution of 560×560 at a batch size of 8 on an RTX6000 GPU. We observe a modest 7% increase in runtime from 186.3 $\rightarrow$ 198.8 ms per pair.

K. Visual Localization on Aachen v1.1

We evaluate RoMa on the Aachen Day-Night v1.1 benchmark using the pipeline of HLoc [43] (v1.4). We present the results in Table 9.