

# $\mathbb{X}$ Resolution Correspondence Networks

## Supplementary Material

Georgi Tinchev<sup>1</sup>

gtinchev@robots.ox.ac.uk

Shuda Li<sup>2</sup>

shuda.li@xyzreality.com

Kai Han<sup>3</sup>

khan@robots.ox.ac.uk

David Mitchell<sup>2</sup>

david.mitchell@xyzreality.com

Rigas Kouskouridas<sup>2</sup>

rigas.kousk@xyzreality.com

<sup>1</sup> Oxford Robotics Institute

University of Oxford  
Oxford, UK

<sup>2</sup> XYZ Reality

London, UK

<sup>3</sup> Visual Geometry Group

University of Oxford  
Oxford, UK

This supplementary material provides extra details which are not presented in the main paper due to space limitations. In the following document we discuss the effects of using various re-sampling resolutions during testing in Sec. 1. In Sec. 2, we provide more in depth comparison on HPatches and propose a novel evaluation criterion to demonstrate the effectiveness of the  $\mathbb{X}$ RRCNet. In Sec. 3, we show more qualitative results on InLoc and Aachen Day-Night dataset, which further demonstrate the quality of the proposed model. Moreover, in Sec. 4 we demonstrate the accuracy of our  $\mathbb{X}$ RRCNet in the challenging task of 3D reconstruction using the Aachen Day-Night dataset. We conclude with a brief description of the source code released in Sec. 5.

First of all, we visualise the output feature maps and correlation maps of the key modules in  $\mathbb{X}$ RRCNet in Fig 1 to illustrate the effectiveness of the key modules when solving a correspondence task. We plot 5 examples of various training and testing images. Each is superimposed with the colour map representing the output feature maps or correlation maps of the corresponding modules. From left to right, we plot the coarse features maps from the FPN decoder, the fine feature maps, the 2D coarse correlation map calculated by querying the key point in the source image into the 4D correlation tensor, the same coarse correlation map querying into the 4D tensor after the first mutual matching layer and after the second mutual matching layer respectively. In the end, we plot the final 2D coarse correlation map and the fine correlation map after the re-weighting. For feature maps, we simply visualise the max values along the channels. It can be noticed that the coarse and fine feature map contains similar patterns except the resolution difference possibly due to the original design of FPN layers. The raw 4D correlation tensor does show a peak around the ground truth point location but also contains significant amount of noise. After two rounds of mutual matching filtering, most of the noise are suppressed except a few ambiguous candidates, and the final re-weighting allows the network to look into the local area in detail so that  $\mathbb{X}$ RRCNet can make correct predictions in the end.

# 1 Effects of Different Re-sampling Resolutions

In this section we present both qualitative and quantitative analysis on HPatches when re-sampling the testing images into various resolutions. As shown in Fig. 2, we varied the input image resolution from 720 to 3840 (4K) with a step size of about 200. From left to right, we show the Mean Matching Accuracy (MMA) [14, 15] plots for the cases of illumination challenges, viewpoint challenges, and overall. The native resolution of the HPatches dataset is reported in Tab. 2 of the main paper.

We observe that the low re-sampling resolution has a major impact on the accuracy in the viewpoint challenges. In contrast, for illumination challenges, low resolution performs relatively well for the low error band ( $< 3$  pixels). However, the increased re-sampling resolution leads to better performance on the illumination challenge at the cost of a small decrease at the low error band. This is possibly due to stronger ambiguity in the local region in illumination scenarios. For example, the lighting changes introduce blur around many key points when transitioning from day to night. As the resolution increases, the predicted key point locations are more likely to converge towards more repeatable but less accurate areas. As far as the large error band is concerned, the performance of our method saturates for the illumination scenario while increasing for the viewpoint challenges as the re-sampling resolution increases. The area under the MMA curve is also provided to measure the overall accuracy. It can be seen that the performance gain using higher re-sampling resolution saturated around 2600 to 3400 with the peak performance at resolution 3000. Note that Fig. 8 in the main paper provides a clear visualisation of the overall performance and Fig 2 provides individual plots for each tested resolution.

We have also evaluated different re-sampling resolutions on the InLoc [16] dataset in Fig. 3. It can be seen that high-resolution images result in better relocalisation accuracy in terms of the translation error.

Fig. 4 shows the heatmaps of predicted target point using input images of various resolutions. The ground truth match is marked with a white dot. It can be seen that higher re-sampling resolutions consistently reduce the uncertainty indicated by the size of the coloured blob. However, as the resolution further increases over 3000, the prediction becomes over-confident towards a close but inaccurate location. This is possibly because of the reduced receptive field of the feature backbone relative to the original image.

In addition to evaluating the re-sampling impact for inference, we also trained our correspondence network using various training image resolutions. Surprisingly, increasing the input resolution during training does not improve performance, as shown in Fig. 5. We hypothesise this is because various training image resolutions contain a fixed amount of information that a correspondence network can use. Therefore, we choose to use 400 px resolution during training in order to achieve a fair comparison with other baseline methods. Please note that all methods are trained with a batch size of 16 to accommodate higher resolution in the feature maps.

## 2 Qualitative Analysis — HPatches

In Fig. 6 and 7, we select six individual testing pairs to demonstrate that  $\mathbb{X}$ RCNet outperforms DualRCNet [17] and SparseNC [18] respectively in terms of the ratio of correct matching

predictions of top 2000 outputs<sup>1</sup>. It can be seen that  $\mathbb{X}$ RCNet is capable of producing more reliable results than previous works. In this section we propose a novel evaluation criterion in supplement to the main results along with the qualitative comparison of Fig. 6. This new evaluation criterion can be formulated as:

$$\mathbb{N}(\tau^-; \tau^+, +, -) = \sum_i^N \mathbf{1}(c_i^+ > \tau^+ \cap c_i^- < \tau^-), \quad (1)$$

where  $c_i^+$  and  $c_i^-$  is the ratio of the correct matches out of all the predicted matches of two comparing methods denoted as '+' and '-' respectively.  $\tau^+$  and  $\tau^-$  are thresholds of the corresponding ratio.  $i \in \{1, 2, \dots, N\}$  is the index of the testing pairs in the dataset.  $\mathbf{1}(\cdot)$  is a binary indicator function such that  $\mathbf{1}(\text{True}) = 1$  and  $\mathbf{1}(\text{False}) = 0$ . As long as there exists the pixel-wise ground truth label, we can always adopt Equation 1 to calculate the number of pairs that favour the '+' method against the '-' method.

Equation 1 measures the number of testing image pairs that favours method '+' with respect to ratio  $\tau^-$  at a specific positive  $\tau^+$ . In other words,  $\mathbb{N}(\cdot)$  is a histogram of the testing pairs where the first method '+' achieves accuracy higher than the threshold  $\tau^+$  but the second method '-' achieves accuracy lower than  $\tau^-$ . In the top row of Fig 8, we illustrate plotting both the  $\mathbb{N}(\tau^-; \tau^+, \mathbb{X}\text{RCNet}, \text{DualRCNet})$  the blue curve vs  $\mathbb{N}(\tau^-; \tau^+, \text{DualRCNet}, \mathbb{X}\text{RCNet})$  the red curve over the range of  $\tau^- \in [0, \tau^+]$  with a step size of 0.1, and  $\tau^+$  is set to 0.75, 0.85 and 0.95, respectively. Similarly, in the bottom row of Fig. 8 we present both  $\mathbb{N}(\tau^-; \tau^+, \mathbb{X}\text{RCNet}, \text{SparseNC})$  as the blue curve vs  $\mathbb{N}(\tau^-; \tau^+, \text{SparseNC}, \mathbb{X}\text{RCNet})$  as the red curve. The three sub-figures in Fig. 8 compare the number of testing data that favours  $\mathbb{X}$ RCNet against those favouring DualRCNet/SparseNC. It demonstrates that the proposed  $\mathbb{X}$ RCNet consistently outperforms DualRCNet and SparseNC for all combination of  $\tau^-$  and  $\tau^+$  values as the number favouring  $\mathbb{X}$ RCNet is significantly higher than the number favouring DualRCNet/SparseNC.

### 3 Qualitative Analysis – InLoc and Aachen Day-Night

Fig. 9 illustrates the performance of  $\mathbb{X}$ RCNet on the InLoc dataset. Similarly to HPatches, the increase in resolution from 1600 to 3840 (4K) results in better performance. The 4K up-sampling resolution for InLoc dataset performs better in terms of relocalisation accuracy than the rest. As mentioned in the main article, we hypothesise this due to the native resolution of testing images in InLoc is much higher than that of HPatches.

Fig. 10 shows visual examples of the proposed model evaluated on the Aachen Day-Night dataset [9].

### 4 3D Reconstruction Using Dense Correspondences

To demonstrate a potential application of using the correspondence network, we plot the 3D point cloud reconstructed using the  $\mathbb{X}$ RCNet in Fig. 11 on the Aachen Day-Night reference images. We also compare the quality of the 3D reconstruction using  $\mathbb{X}$ RCNet and DualRCNet in Fig. 12. It can be seen that the quality of the reconstructed models are fairly close for the two methods.

<sup>1</sup>2000 is a arbitrary number. Following previous works of D2Net, DualRCNet and SparseNC, we also adopt 2000 for a fair comparison.

## 5 Code

We include code as part of the supplementary material to allow for reproducibility of the results as well as retraining the models. We believe it is crucial to open-source this project to encourage comparison and facilitate future research along this direction.

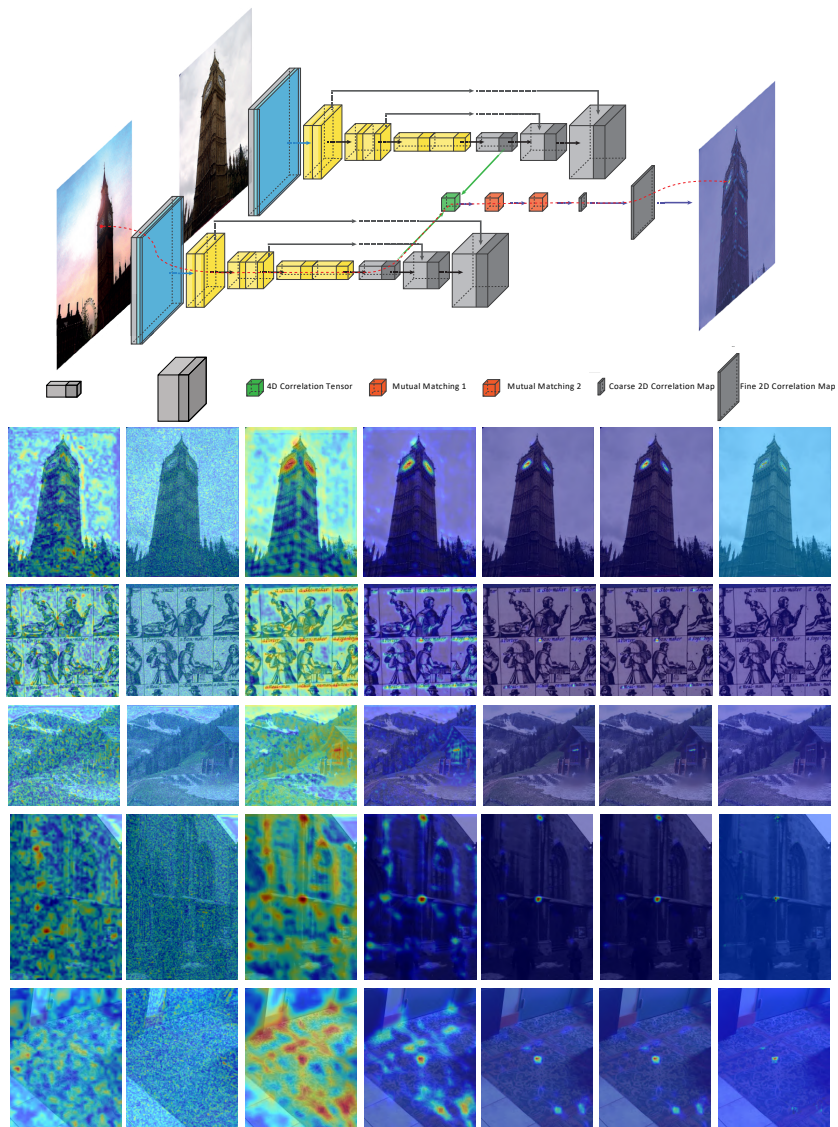


Figure 1: Visualisation of the feature maps and correlation maps of key components in  $\mathbb{X}$ RcNet.

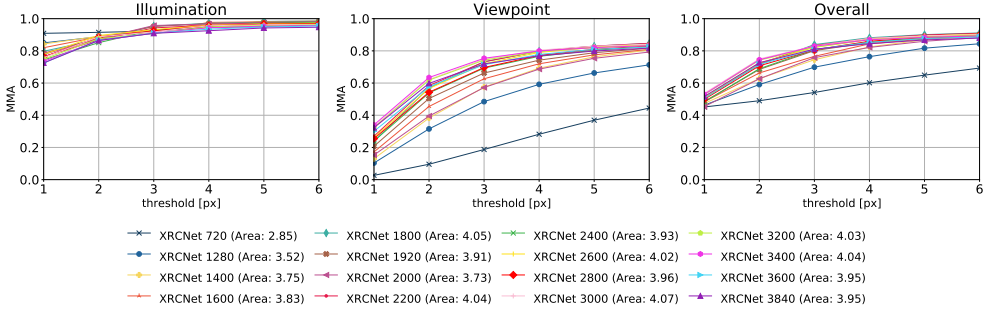


Figure 2: Comparison of  $\mathbb{X}$ RCCNet with respect to the up-sampled input image resolution evaluated on the HPatches dataset.

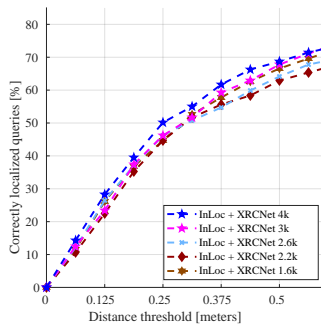


Figure 3: Comparison of  $\mathbb{X}$ RCCNet with respect to the up-sampled input image resolution on the InLoc dataset after geometric verification.

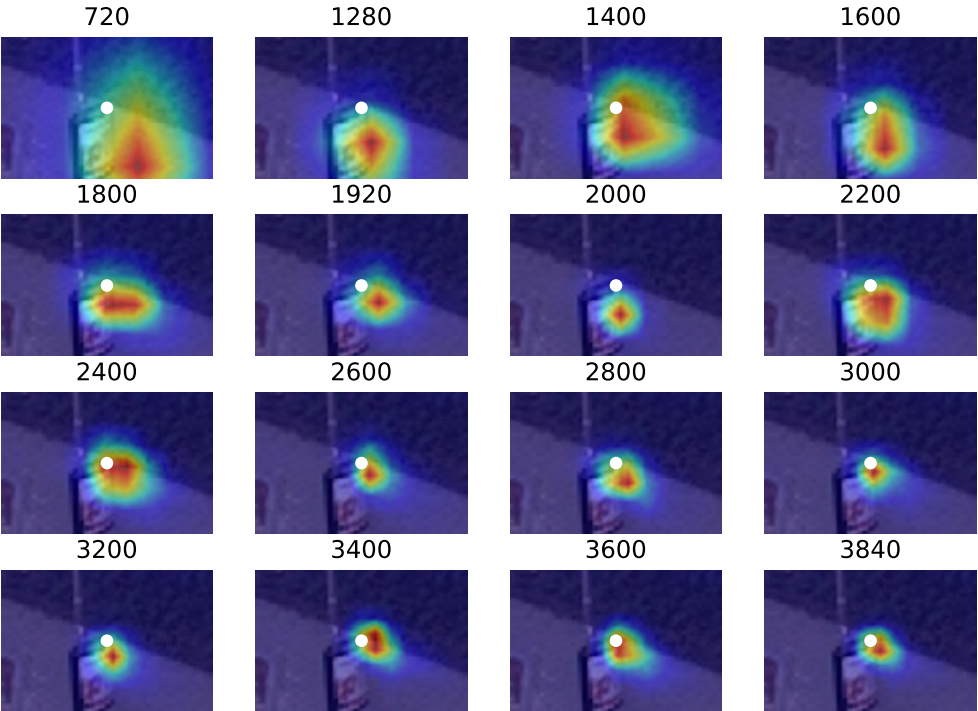


Figure 4: Produced keypoint heatmap from the correlation tensor overlaid at a reference image. The ground truth location of the query keypoint is denoted in white.

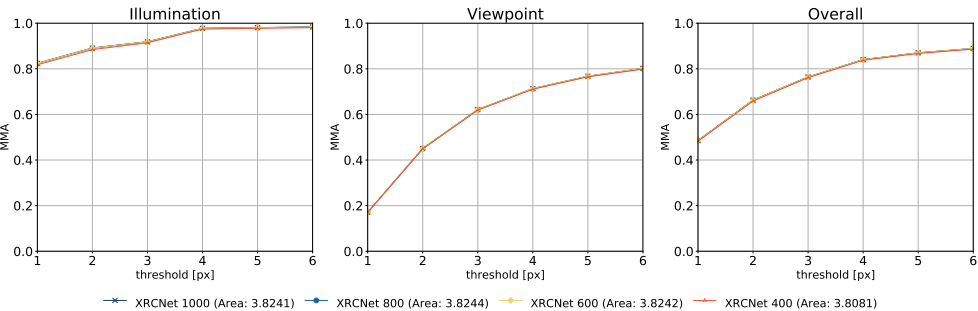


Figure 5: Training XRCNet with re-sampled image resolution of 400px to 1,000 px at every 200px.



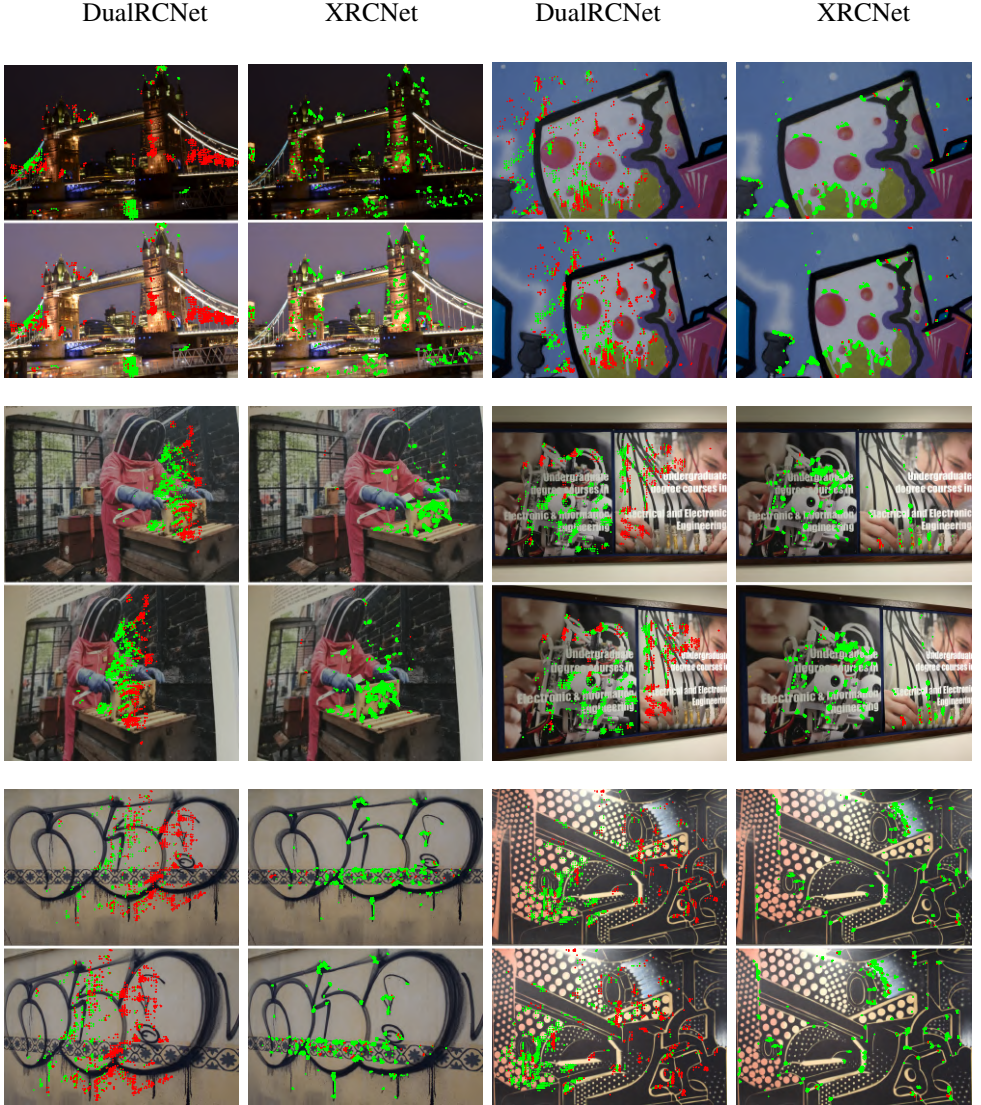


Figure 6: Qualitative comparison between  $\mathbb{X}$ RCNet and DualRCNet on HPatches. The green dots represent the correct matches whose errors are within 3 pixels, and red dots the incorrect matches.  $\mathbb{X}$ RCNet produces more correct matches out of the top 2000 matches than DualRCNet.



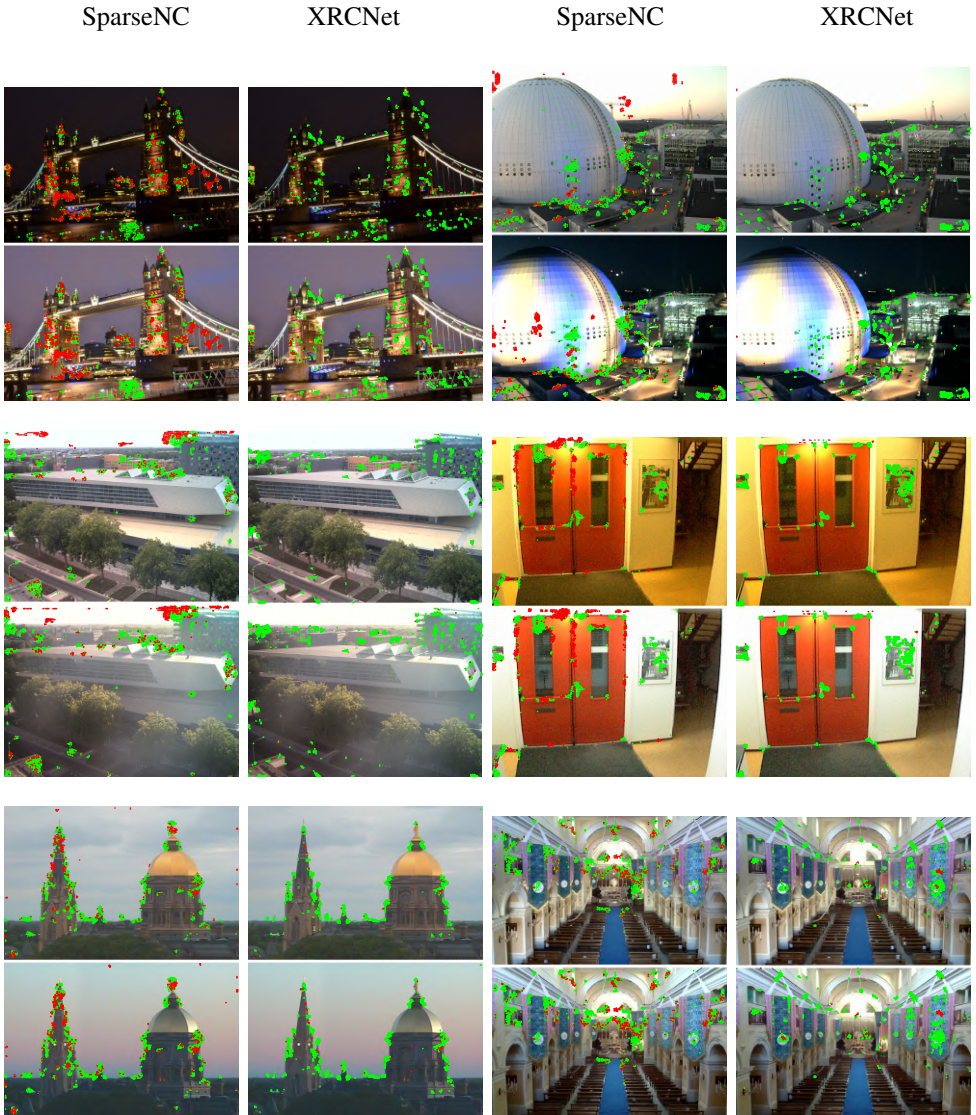


Figure 7: Qualitative comparison between  $\mathbb{X}$ RCNet and SparseNC on HPatches. The green dots represent the correct matches whose errors are within 3 pixels, and red dots the incorrect matches.  $\mathbb{X}$ RCNet produces more correct matches out of the top 2000 matches than SparseNC.

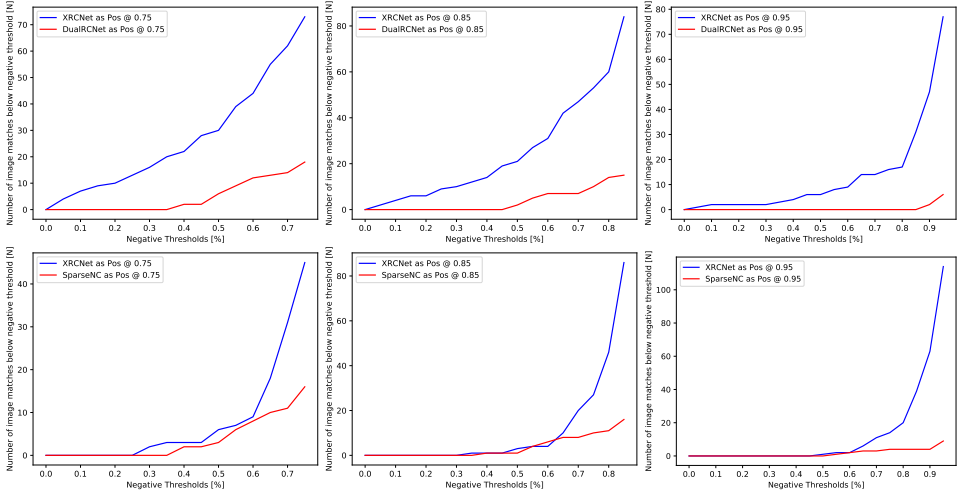


Figure 8: **Top row:** The comparison of the number of testing pairs that  $\mathbb{X}$ RCNet outperforms DualRCNet (blue curve) and DualRCNet outperforms  $\mathbb{X}$ RCNet using Equation 1. **Bottom row:** Similar comparison between  $\mathbb{X}$ RCNet and SparseNC. For all comparisons, the  $\tau^+$  is chosen as 75%, 85%, and 95% for both curves.  $\tau^-$  in Equation 1 is denoted as the Negative Threshold. 'Pos' denotes '+' method and 0.75 represent the  $\tau^+$  ratio threshold.



Figure 9: Examples of  $\mathbb{X}$ RCNet running on the InLoc dataset. To simplify the rendering, we choose top 100 matches before the geometric verification to demonstrate the quality of the matches.

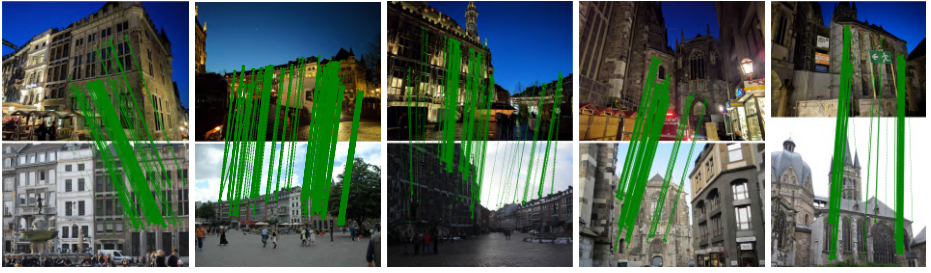


Figure 10: Examples of  $\mathbb{X}$ RCNet running on the Aachen Day-Night dataset - top 2000 matches are displayed. It is worth pointing out the output matches with high reliability scores are heavily clustered in relatively small regions and may overlap each other.

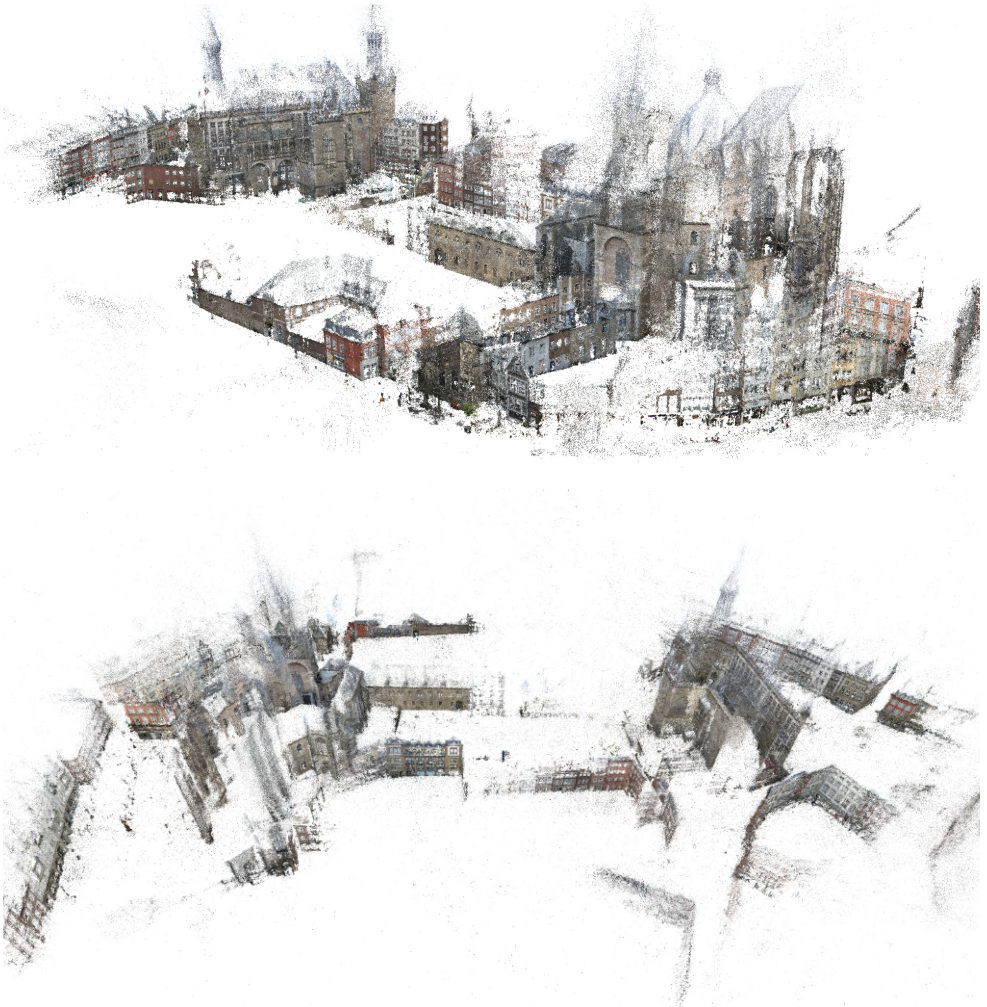


Figure 11: 3D model reconstructed using correspondences obtained by  $\mathbb{X}$ RCNet for the Aachen Day-Night dataset.





Figure 12: Qualitative comparison of  $\mathbb{X}$ RCNet (top) and DualRCNet (bottom) 3D model reconstructions on the Aachen Day-Night dataset.

## References

- [1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [2] Xinghui Li, Kai Han, Shuda Li, and Victor Adrian Prisacariu. Dual-Resolution Correspondence Networks. In *Proceedings of Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.
- [3] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *arXiv preprint*, 2020.
- [4] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *Proceedings of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [5] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.