LM-Reloc: Levenberg-Marquardt Based Direct Visual Relocalization: Supplementary Material

Lukas von Stumberg[1,2*] Patrick Wenzel[1,2*] Nan Yang[1,2] Daniel Cremers[1,2]
[1] Technical University of Munich [2] Artisense

A. Video

As mentioned in the paper, we provide a video of the qualitative relocalization demo, which is available at https://vision.in.tum.de/lm-reloc.

B. Network Architecture

CorrPoseNet. The CorrPoseNet takes 2 images (I and I') as the input and outputs the relative pose R, t between those images. The overall network architecture of the CorrPoseNet is depicted in Figure 1. The convolutional blocks consist of in total 9 convolutional layers followed by ReLU activations. The architectural details of the convolutional blocks are listed in Table 1. The correlation layer which takes the output of the convolutional blocks as input is described in the main paper. The correlation layer is followed by the regression block which regresses the relative pose. The layers of the regression block are listed in Table 2. The output of the network is the rotation R as Euler angles and translation t.

Table 1: Network architecture and parameters of the convolutional blocks. **k** denotes kernel size, **s** stride, and **p** padding.

Convolutional blocks						
layer	**in-chns**	**out-chns**	**k**	**s**	**p**	**activation**
conv0	3	16	16	2	3	ReLU
conv1	16	32	5	2	2	ReLU
conv2	32	64	3	2	1	ReLU
conv3	64	64	3	1	0	ReLU
conv4	64	128	3	2	2	ReLU
conv5	128	128	3	1	1	ReLU
conv6	128	256	3	2	1	ReLU
conv7	256	256	3	1	1	ReLU

LM-Net. We adopt U-Net [1] as the encoder of LM-Net. However, we change the decoder part of the architecture in the following way. Starting from the coarsest level, we upsample (with bilinear interpolation) the feature maps by 2 and concatenate those feature maps with the feature map of the higher level. This is followed by 1×1 convolutional filters. This procedure is repeated 4 times. This results in the feature pyramid maps as described in Table 3.

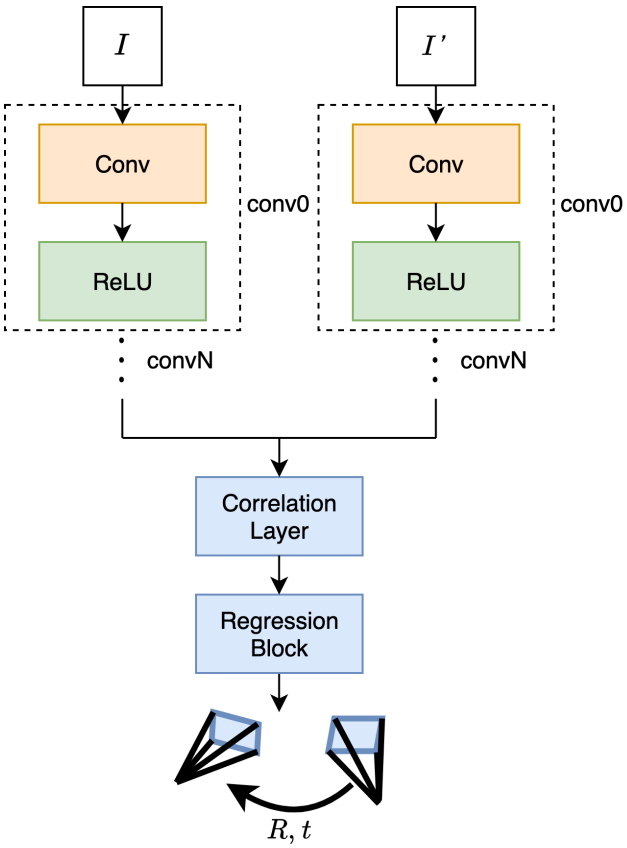
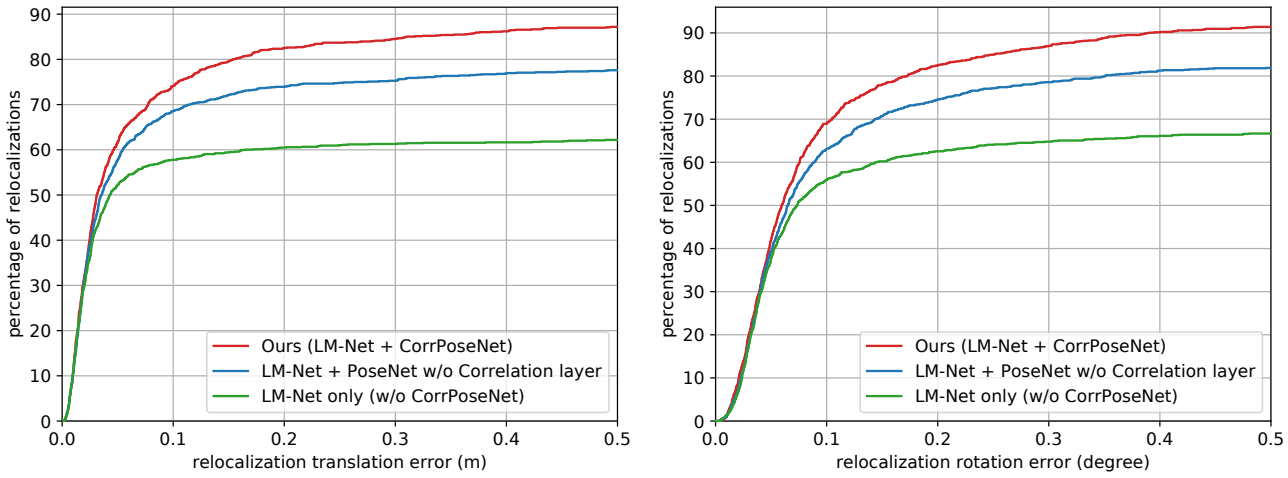


Figure 1: Network architecture of CorrPoseNet.

C. Ablation Study Correlation Layer

We demonstrate the impact of the Correlation layer in the proposed CorrPoseNet. We compare it to a simpler pose estimation network where the correlation and regression layers are replaced with two 1×1 convolutions with 3 output-channels each, which directly regress rotation and Euler angles. This simpler PoseNet has one more convo-

*Equal contribution.



(a) Translation error.

(b) Rotation error.

Figure 2: Cumulative error plot for relocalization on the CARLA relocalization benchmark validation data [2]. It can be seen that the correlation layer in CorrPoseNet has a large impact on the performance.

Table 2: Network architecture and parameters of the regression block. **k** denotes kernel size, **s** stride, and **p** padding. $N_c = 256$ denotes the input channels for the CARLA model, and $N_o = 260$ denotes the input channels for the Oxford model, respectively.

Regression block						
layer	**in-chns**	**out-chns**	**k**	**s**	**p**	**activation**
conv0	N_c / N_o	128	7	1	0	ReLU
BN	128	128	-	-	-	
conv1	128	64	5	1	0	ReLU
BN	64	64	-	-	-	
FC	2304	6	-	-	-	

Table 3: Output of the decoder of LM-Net. H, and W denote height and width of the feature maps.

Decoder layer	**Output size**
F_1	$16 \times H/8 \times W/8$
F_2	$16 \times H/4 \times W/4$
F_3	$16 \times H/2 \times W/2$
F_4	$16 \times H \times W$

lutional block conv8 with 512 output channels, kernel size 3, stride 2, and padding 1. Otherwise the network architecture and parameters are the same as for CorrPoseNet. The results on the CARLA validation data are shown in Figure 2. Even the simpler pose estimation network (PoseNet w/o Correlation layer) improves the result over using identity as an initialization for the direct image alignment (LM-Net only). However, utilizing the correlation layer significantly boosts the performance.

Table 4: This table shows the AUC until 0.5 meters / 0.5 degrees for the relocalization error on the Oxford validation sequences. Our data augmentation (which warps the images using random poses) improves both rotation and translation error.

Method	t_{AUC}	R_{AUC}
Ours	**80.45**	**65.11**
Ours w/o data augmentation	80.15	64.58

D. Ablation Study Oxford Data Augmentation

We show the impact of the data augmentation for the Oxford RobotCar Relocalization benchmark, where we warp the images to different poses using dense depths in Table 4. It can be seen that the proposed augmentation improves translation and rotation error on the validation data.

References

[1] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 1

[2] L. von Stumberg, P. Wenzel, Q. Khan, and D. Cremers. GN-Net: The Gauss-Newton Loss for Multi-Weather Relocalization. *RA-L*, 5(2):890–897, 2020. 2