Supplementary Material: Unsupervised Multi-Modal Image Registration via Geometry Preserving Image-to-Image Translation

Moab Arar Yiftach Ginger Dov Danon Amit H. Bermano Daniel Cohen-Or
Tel Aviv University

1. Geometric-Preserving Translation

In this section we provide qualitative results that supports our claims which were provided in the ablation study (see Geometric-Preserving Translation Network in the ablation study). More specifically, Figure 2 showcases the geometric-preserving quality that the translation network has under our proposed two-flow training scheme (i.e jointly training $T \circ R$ and $R \circ T$). Additionally, in Figure 3 we show that training T and R such that the translation task is performed first ($R \circ T$), yields poorly reconstructed images. Finally, we show in Figure 4 that in the *registration first* training flow (i.e $T \circ R$), the translation network supersedes the registration network by performing the translation as well the registration, while the registration network degenerates to performing minor spatial transformation.

2. CycleGAN As Geometry Preserving Translator

In the paper (Section 5.1 - Quantitative Evaluation), we used CycleGAN [6] in order to translate the input image $I_a \in \mathcal{A}$ onto modality $\mathcal{B}$. This in turn, allowed us to extract SIFT [4] features from the generated image and the target image $I_b \in \mathcal{B}$, and later apply feature matching algorithms in order to estimate the spatial correspondence between the images. One of our claims was that CycleGAN is expected to be geometry preserving because it isn't explicitly trained to match the ground-truth data. We show a sample result in Figure 7 where clearly the translated image share the same coordinate system with the input image I_a (see left images in Figure 7a and Figure 7b). Furthermore, extracting the SIFT features from the generated image allows better feature matching than extracting the SIFT features from the input image I_a. To see this, we show the top-10 matched features between I_a and I_b (Figure 7a) and the top-10 matched features between the generated image and I_b (Figure 7b).

3. Qualitative Evaluation - Additional Results

In this section we provide further qualitative results for Figure 4 in the paper. Specifically, we show additional visual results on the registration between RGB and Depth modalities (Figure 8), and between RGB and Thermal modalities (Figure 9).

Name	Input	Output size ($H \times W \times C$)	Comments
I_a	-	$288 \times 384 \times 3$	Input
I_b	-	$288 \times 384 \times 1$	Input
Encoder			
Conv1	$I_a \odot I_b$	$288 \times 384 \times 32$	3×3 Conv/Stride 1
LR1	Conv1	$288 \times 384 \times 32$	Leaky ReLU
Res1	LR1	$288 \times 384 \times 32$	Residual Block
Pool1	Res1	$144 \times 192 \times 32$	Max Pooling
Conv2	Pool1	$144 \times 192 \times 64$	3×3 Conv/Stride 1
LR2	Conv2	$144 \times 192 \times 64$	Leaky ReLU
Res2	LR2	$144 \times 192 \times 64$	Residual Block
Pool2	Res2	$72 \times 96 \times 64$	Max Pooling
...	...	...	...
Conv6	Pool5	$4 \times 6 \times 64$	3×3 Conv/Stride 1
LR6	Conv6	$4 \times 6 \times 64$	Leaky ReLU
Res6	LR6	$4 \times 6 \times 64$	Residual Block
Pool6	Res6	$4 \times 6 \times 64$	Max Pooling
Conv7	Pool6	$4 \times 6 \times 64$	3×3 Conv/Stride 1
LR7	Conv7	$4 \times 6 \times 64$	Leaky ReLU
Res7	LR7	$4 \times 6 \times 64$	Residual Block
Decoder			
UP1	Res7	$9 \times 12 \times 64$	$2\times$ Up Sample
Conv8	Up1 $\odot$ Res6	$9 \times 12 \times 64$	3×3 Conv/Stride 1
LR8	Conv8	$9 \times 12 \times 64$	Leaky ReLU
...	...	...	...
UP6	LR12	$288 \times 384 \times 32$	$2\times$ Up Sample
Conv13	Up6 $\odot$ Res1	$288 \times 384 \times 32$	3×3 Conv/Stride 1
LR13	Conv13	$288 \times 384 \times 32$	Leaky ReLU
Res13	LR13	$288 \times 384 \times 32$	Residual Block
Conv14	Res13	$288 \times 384 \times 32$	1×1 Conv/Stride 0
LR14	Conv14	$288 \times 384 \times 32$	Leaky ReLU
Conv15	LR14	$288 \times 384 \times 2$	3×3 Conv/Stride 1

Table 1: **Detailed Description of $\mathbf{R_\Phi}$.** We omitted full details due to space constraint. All Leaky ReLU layers are with negative slope $\alpha = 0.2$. The output layer Conv15 is initialized with weight distribution of $\mathcal{N}(0, 0.0001)$. All other layers are initialized with Kaiming [2] initialization method. We use residual block in the encoder-path and output path, which are based on the implementation in [3]. We use $\odot$ to represent the channel-wise concatenation operation.

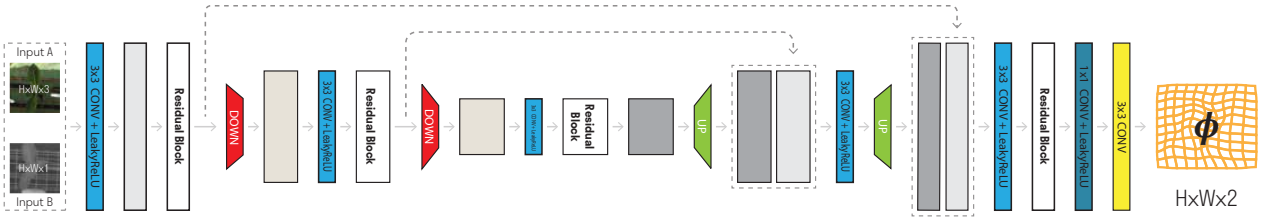


Figure 1: $\mathbf{R_\Phi}$ **- Deformation Field Generation Network.**

4. Loss Ablation - Visual Results

In the paper we claimed that training the registration network R on both the reconstruction loss and the cGAN loss (i.e $\mathcal{L}_{recon}$ and $\mathcal{L}_{cGAN}$) provides smooth and accurate results. Specifically, we claimed that the cGAN loss is necessary in order to produce smooth spatial transformation. In Figure 10, we show the effect of optimizing R with respect to each loss individually. As can be seen, when R is trained with respect to the $L1$ loss, the registration network outputs results which are very distinct but not smooth. However, when we optimize the registration network with respect to the adversarial loss, then we achieve much smoother results.

5. Registration Network Architecture

Recall that our registration network R consists of a resampler R_S, and a deformation field generation network R_Φ. The network R_Φ is UNET-based [5] with residual connections [1] in the encoder and output path (see illustration in Figure 1). For full description of the network, please refer to Table 1.

6. NCC As A Metric

As claimed in the paper, using Normalized Cross Correlation (NCC) as a similarity measure to optimize our registration network R produces less accurate registration. Further, we claimed that in some-cases, the registered image is noisy and contains many artifacts. We show sample failure cases, when our network R is optimized with respect to NCC only. The results are shown in Figure 5.

7. Translation Network Capacity

In this section, we show that there must be a balance between the capacity of the translation network T, and the registration network R. In particular, the capacity of the network T refers to the number of layers or the number of filters that are used in the network. To test how changing the capacity of T affects the registration network R, we train T in three different configurations. Recall, the network T is an encoder-decoder network with residual blocks applied to the encoded features. We use three different architectures for T, where each one has different number of residual blocks. Specifically, we use 9 residual blocks (resnet_9), 6 residual blocks (resnet_6) and 3 residual blocks (resnet_3). As can be seen from Figure 6, the registration accuracy is degrading when T's capacity is increased.

References

[1] Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 2

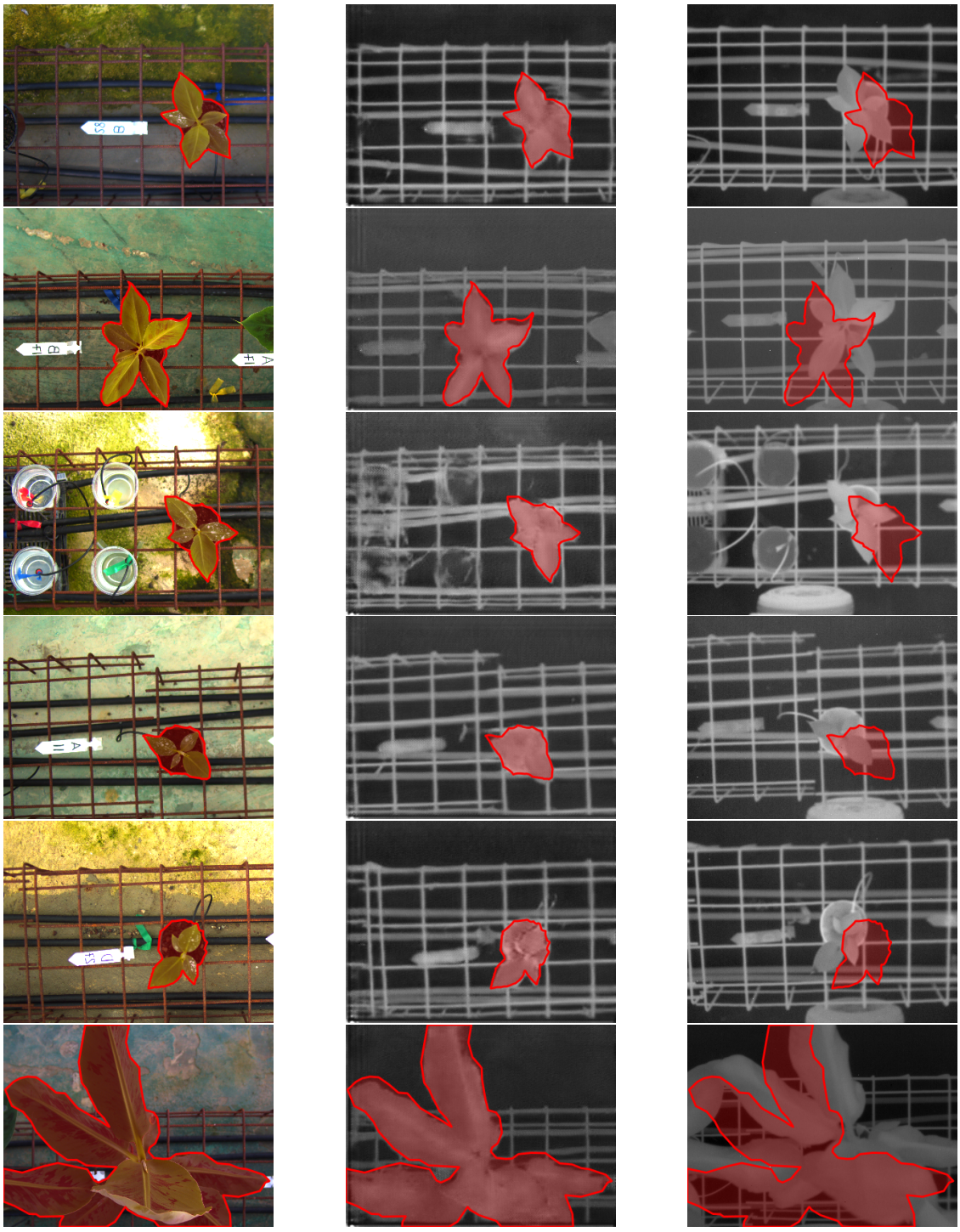
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society. 1

[3] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 1

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 1

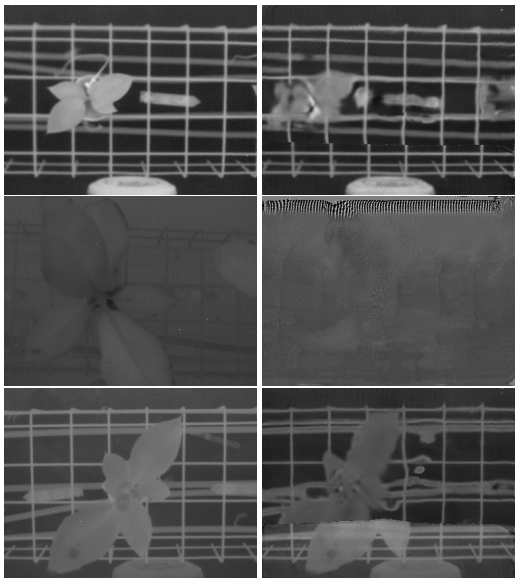
[5] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). 2

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1



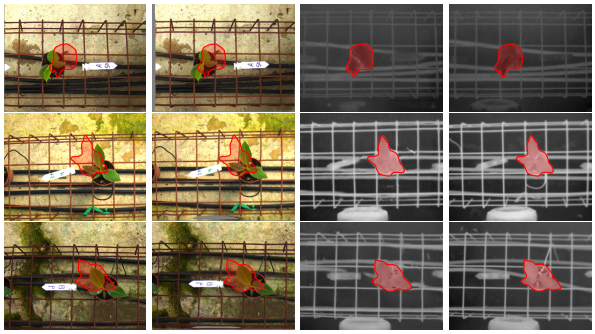
Input A | $T(I_a)$ - Translated | Input B

Figure 2: **Geometric Preserving Translation Network - Visual Results.** On the left are the input RGB images, on the right the target IR images and in the middle are the results of applying only the translation network T on the RGB image. We overlay the silhouette of the salient object in the RGB image on top of all of the images to emphasize the preservation of the geometry during the translation.



Input B Result

Figure 3: **Translation First - Sample Results.** We show real samples on the left and the generated samples using $R \circ T$ on the right. As can be seen, the registration network R takes on the role of a translation network, resulting in low quality images.



Input A Registered Translated Input B

Figure 4: **Registration First - Sample Results.** In all images we place the silhouette of the salient object in the real sample from modality $\mathcal{B}$ on top of the images. As can be seen, the registration network applies only minor spatial transformation on the input image (see Input A and Registered columns), while most of the registration task is done implicitly by the translation network (see Translated and Input B). Please note that the third column (Translated), represent the output of $T \circ R(I_a, I_b)$.

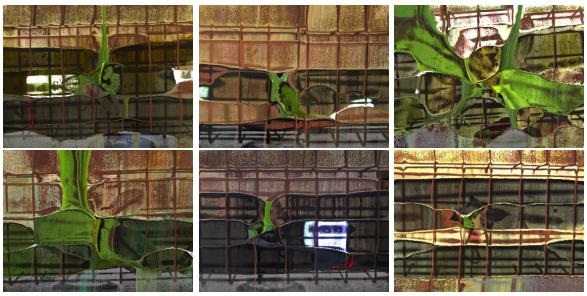


Figure 5: **NCC Failure Cases.** In some cases, optimizing the registration network with respect to normalized cross produces artifacts.

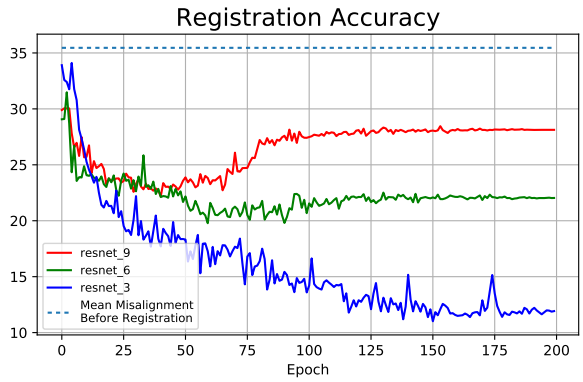
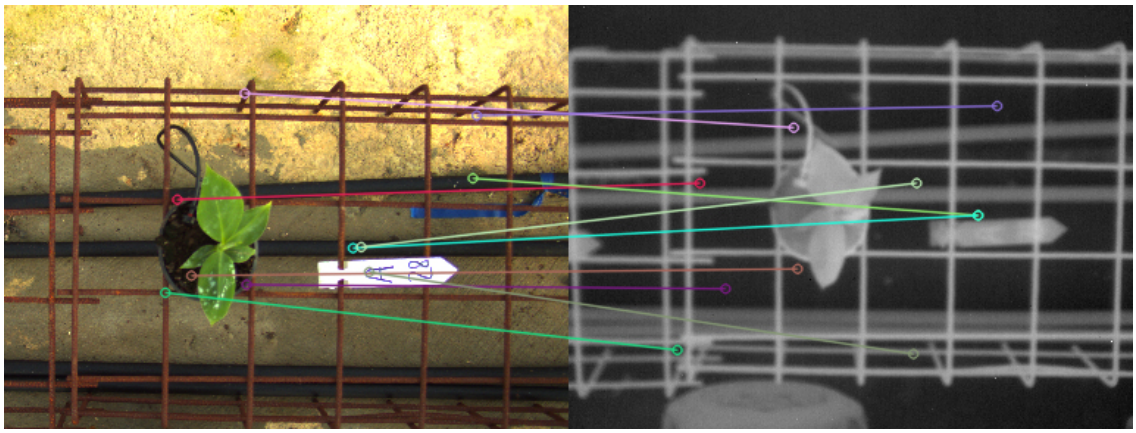
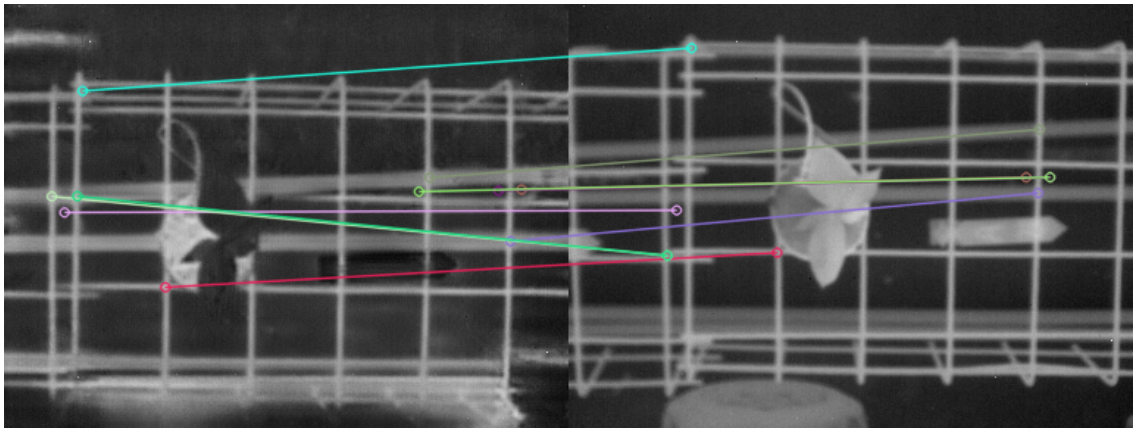


Figure 6: **Translation Network Capacity vs Registration Accuracy.** As can be seen, increasing the number of layers in the translation network T, negatively affects the registration accuracy of R.



(a) Matched SIFT Features Of Input A (left) and Input B (right).



(b) Matched SIFT Features Of CycleGAN's Output (left) and Input B (right).

Figure 7: **Top 10 Matched SIFT Features.** We show the matched SIFT features (a) between I_a and I_b and (b) the translated image (using CycleGAN as a translator) and I_b. As shown in the figure, CycleGAN is geometry preserving (see the left images in (a) and (b)). However, using CycleGAN yields less accurate translation, as suggested by (b). In particular, we see that the translation is not accurate in the salient object, thus most of the matched features in (b) are part of the background.

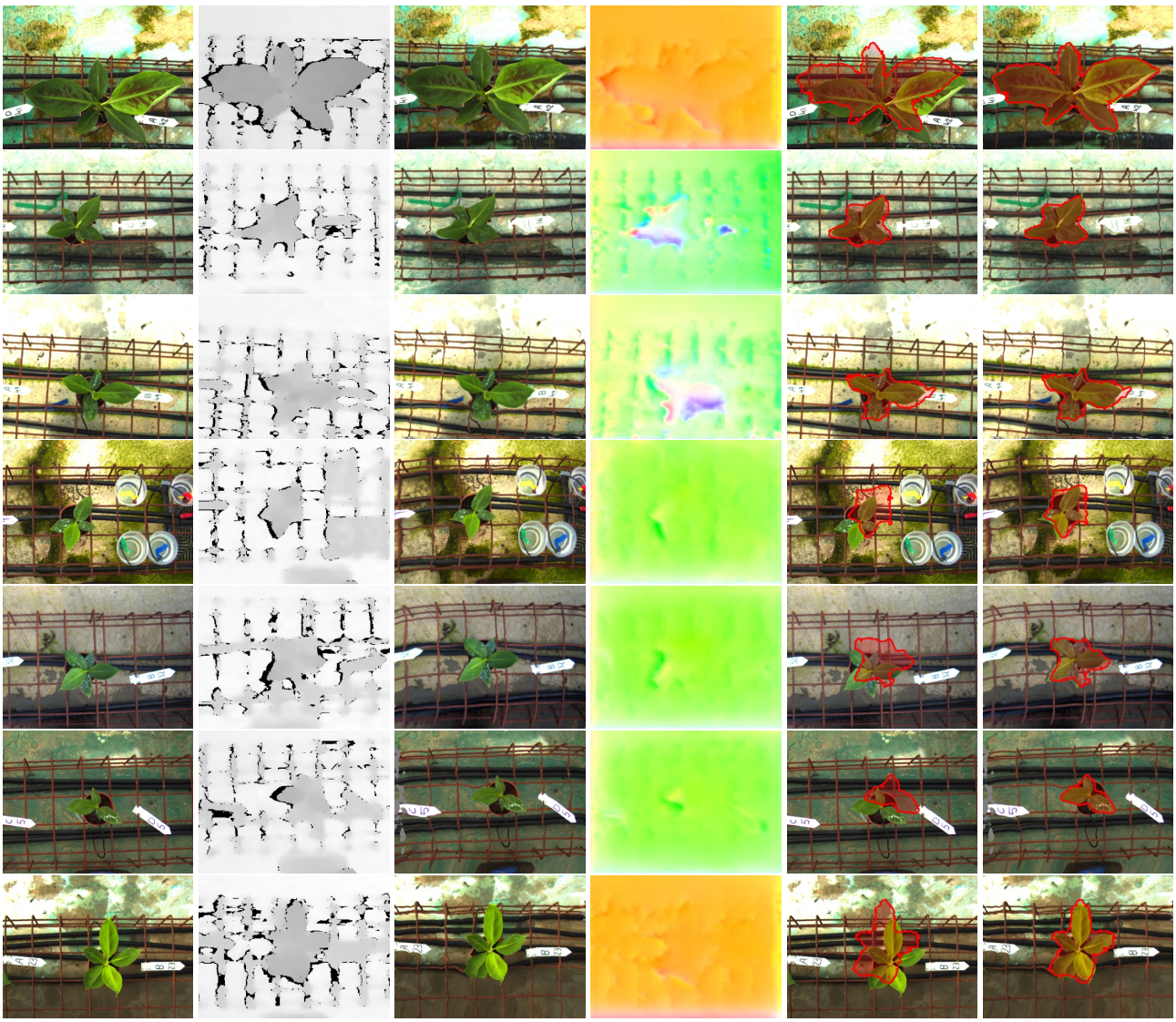


Figure 8: **Image registration between RGB and Depth modalities.** Further results for Figure 4a in the Paper.

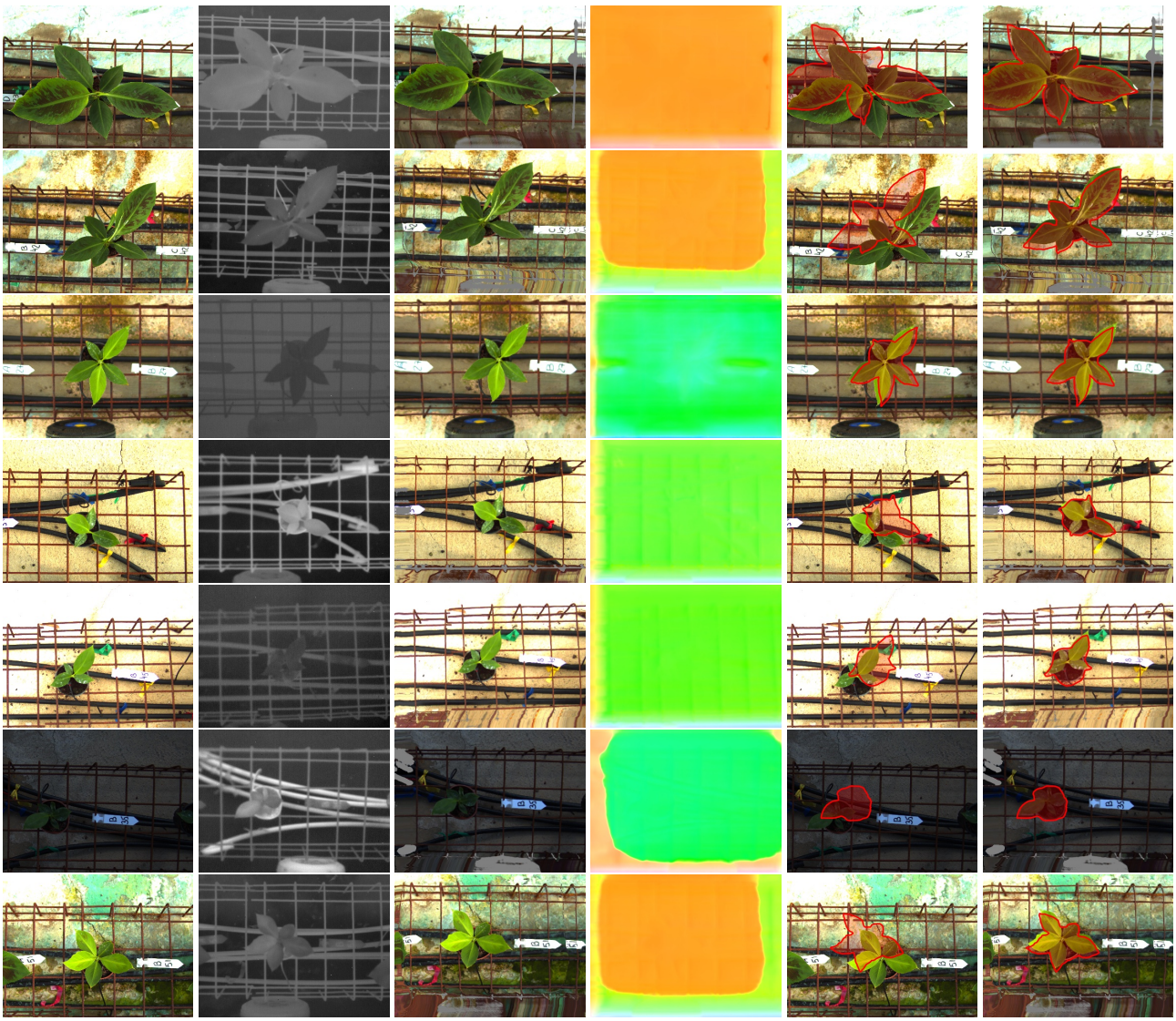


Figure 9: **Image registration between RGB and IR modalities.** Further results for Figure 4b in the Paper.

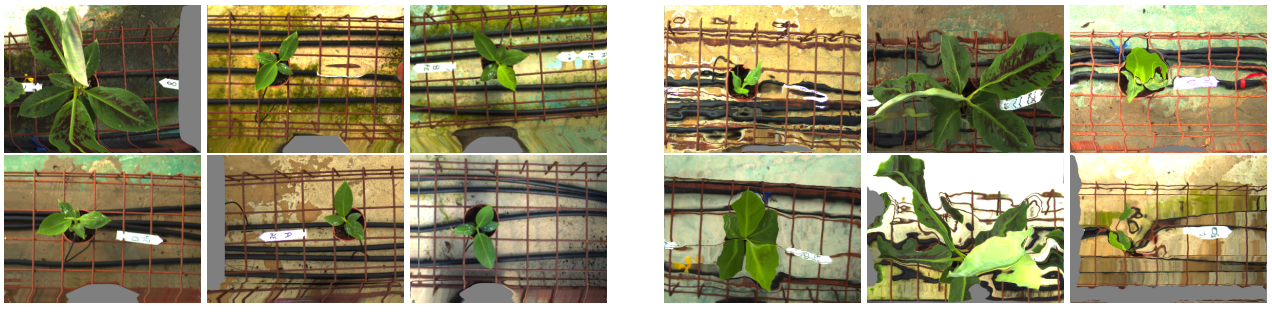


Figure 10: **Loss ablation - visual results.** On the left are the registration results when the registration network R is trained only with respect to the adversarial GAN loss. On the right are the registration results when R is trained only with an $L1$ loss.