STUDYING DEEP LEARNING BASED CO-REGISTRATION FOR INCOHERENT CHANGE DETECTION

Yannik Steiniger German Aerospace Center (DLR) - Institute for the Protection of Maritime Infrastructures
Sven Schröder German Aerospace Center (DLR) - Institute for the Protection of Maritime Infrastructures

1 INTRODUCTION

Imaging sonars, like sidescan sonars or synthetic aperture sonars (SAS), are leading systems for capturing images of the sea floor. They are typically mounted on an autonomous underwater vehicle (AUV) to enable an automatic collection of the data. Repeating surveys in a defined time interval allow for the observation of changes within that period. Possible use cases include determining the growth of algae, studying changes of sea floor structures or detecting man-made objects. In general, change detection algorithms are divided into image based and symbolic based methods[1]. Image-based change detection directly compares two images while in symbolic-based change detection objects are detected in both images separately, georeferenced and then compared[2,3,4]. If the images are captured with a sidescan sonar the conventional image based approach is an incoherent change detection in which two intensity images are first co-registered, i.e., aligned with respect to each other[5,6]. In a second step, changes are detected in the difference image between these aligned images (see Figure 1). However, this image based approach requires a precise co-registration as otherwise the subsequent calculations result in a high number of false alarms and potential missed detections. Additionally, differences in the track of the AUV carrying the sonar result in non-linear differences between the images which cannot be recovered using simple affine transformations, making the co-registration task a challenging one.

In recent years, deep learning has become state-of-the-art in most computer vision tasks and has successfully be applied in medical image registration[7,8,9]. Here the common task is to align 3D magnetic resonance images of human brains either captured from the same person at different times or captured from different persons. Current methods for co-registrating sonar images rely on simple affine transformations[10] which cannot cover all relevant shifts between the two images or on correlation based methods[11,12] which can take long to be computed especially for larger images. Deep learning based co-registration methods can in principle learn an arbitrary complex displacement field which, once learned, can easily be applied to the sonar images. Thus, in this work we investigate several deep learning methods which have shown great results in generative tasks as well as medical image registration for the co-registration of two sidescan sonar images in an incoherent change detection processing chain. More specifically, we analyse the generative adversarial network (GAN) pix2pix[13] as well as the U-Net based models Voxelmorph[7] and LKU-Net[8] and compare them to the two conventional methods homography and optical flow. We assess the quality of the alignment results of the different methods as well as the influence on the overall false alarm rate of the change detection processing chain.

The remaining of this paper is organised as follows. Section 2 introduces the change detection processing

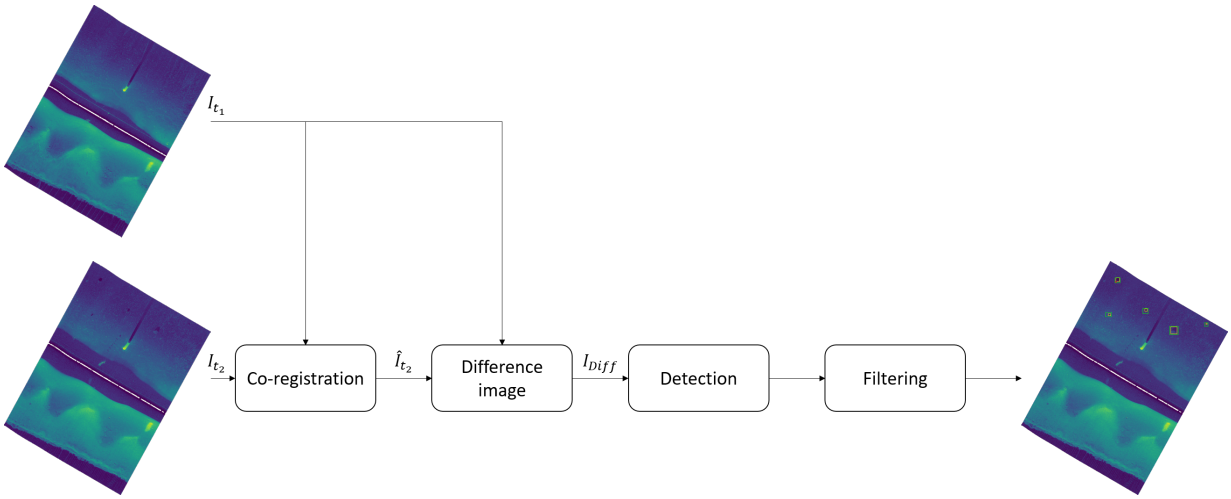


Figure 1: Structure of the baseline change detection.

chain and discusses the methods investigated for the co-registration task. Next, in Section 3 the training and test datasets consisting of real and simulated sidescan sonar images are described. In Section 4 the performance of the different co-registration methods is assessed. Finally, the paper closes with a summary and outlook on future work in Section 5.

2 CHANGE DETECTION PROCESSING CHAIN

2.1 Baseline

For evaluating different methods for co-registering two sidescan sonar images I_{t_1} and I_{t_2} we build upon the change detection processing chain introduced in[10]. Its main steps are shown in Figure 1. For co-registration, first a coarse alignment based on homography is applied. The homography matrix $H \in \mathbb{R}^{3\times3}$ is estimated using matched so-called oriented FAST and rotated BRIEF (ORB) features[14] in the two images. In all following alignment tasks we consider I_{t_1} as the fixed pre-image and I_{t_2} as the post-image to be transformed. After the coarse alignment, the images I_{t_1} and $\hat{I}_{t_2}$ with

$$\hat{I}_{t_2}(m,n) = I_{t_2}\left(\frac{H(1,1)m + H(1,2)n + H(1,3)}{H(3,1)m + H(3,2)n + H(3,3)}, \frac{H(2,1)m + H(2,2)n + H(2,3)}{H(3,1)m + H(3,2)n + H(3,3)}\right), \tag{1}$$

where m and n are the pixel indices, both have the size $M \times N$. Afterwards a fine alignment is applied, where in this work we compare the methods optical flow, pix2pix, Voxelmorph and LKU-Net for this step.

Changes between the aligned images are detected on a pixel-level in the difference image I_{Diff}. Before subtracting the two images, they are normalised and smoothed using a median filter to reduce the present noise. Regions in which changes have occurred are detected by a threshold detector applied to the difference image and marked by bounding boxes. Misalignment or intensity differences between the two input images can thereby lead to a high false positive rate.

To reduce the false positive rate, the detections are filtered using a convolutional neural network (CNN). This CNN is trained to classify snippets into the two categories *object* and *background*[10]. Detected regions whose corresponding snippets are classified as *background* are removed. Furthermore, a size-

based filtering is applied. If a region with width w and height h does not comply with

$$w < 5 \text{ px} \wedge h < 5 \text{ px} \wedge w \times h > 2000 \text{ px}^2 \tag{2}$$

it is filtered.

2.2 Alignment methods

The following section briefly explains the methods analysed for the fine alignment in the change detection processing chain described above. Recall that I_{t_1} is the fixed pre-image and $\hat{I}_{t_2}$ the already coarsely aligned post-image.

2.2.1 Optical flow

To calculate the optical flow between two images the TV-L1 solver[15] is used in this work. The displacement field is calculated based on an image pyramid containing downsampled versions of the input images. The number of levels in the image pyramid depends on the input size and is determined such that at the lowest, i.e., the coarsest, level the shortest image length is between 32 and 16 pixel. At coarser levels larger deformations can be computed. The displacement field is updated from the coarsest to the finest level, i.e., the original image. The final displacement field is then applied to the image $\hat{I}_{t_2}$ to achieve a fine registration with I_{t_1}.

2.2.2 pix2pix

Pix2pix[13] is a GAN originally designed for paired image-to-image translation, e.g., translating a satellite image into a map of the captured area. As all GANs, pix2pix consists of a generator network and a discriminator network. The generator maps the input image into the translated version using a series of convolutional layers. The discriminator is a CNN which takes the original and the translated image, which is either from the training data or generated, and predicts if the input is real or synthetic. For the task of co-registering two sonar images the generator is trained to map the post-image $\hat{I}_{t_2}$ to match the pre-image I_{t_1}. Since the generator and discriminator only accept input images of a predetermined size, in this case 512×512 pixel, the images are first split into patches. After being passed though the trained generator, the aligned patches are reassembled. For the network design we stick to the models proposed in the original pix2pix[13], where the generator is a U-Net[16] with eight downsampling stages and discriminator is a CNN with five convolutional layers. Both networks are trained for 100 epochs with the Adam optimizer and a learning rate of 0.0002.

2.2.3 Voxelmorph

Voxelmorph[7] is a U-Net which was originally designed to co-register two magnetic resonance images. Rather than directly translating one image into the aligned version the U-Net learns to predict a displacement field using both images as an input. This displacement field is then used to align $\hat{I}_{t_2}$ with I_{t_1}. During training the model is optimised such that the transformed post-image is similar to the pre-image and the displacement field is smooth. As with pix2pix, the input size of Voxelmorph is fixed to 512×512 pixel. Thus, the model is applied to patches of the sonar images. The U-Net of Voxelmorph consists of six downsampling stages and the overall network is trained for 500 epochs with the Adam optimizer and a learning rate of 0.001. We investigated deeper and shallower architectures as well as different learning rates and found the settings mentioned above to perform best.

Table 1: Conducted change detection experiments with real SSS images. The star indicates images with manually inserted objects. A, B, C ***and*** D ***indicate different experimental sites.***

Experiment	Pre-image	Post-image	Number of changes	Goal
Exp 1	I_{A,t_1}	I^*_{A,t_1}	5	detect inserted objects in the same image
Exp 2	I_{A,t_1}	I_{A,t_2}	0	generate no false alarms
Exp 3	I_{A,t_2}	I^*_{A,t_1}	5	detect inserted objects in a different post-image
Exp 4	I_{B,t_1}	I^*_{B,t_2}	5	detect inserted objects in a different post-image
Exp 5	I_{C,t_1}	I_{C,t_2}	2	detect objects in a realistic scenario
Exp 6	I_{D,t_1}	I_{D,t_2}	1	detect objects in a realistic scenario

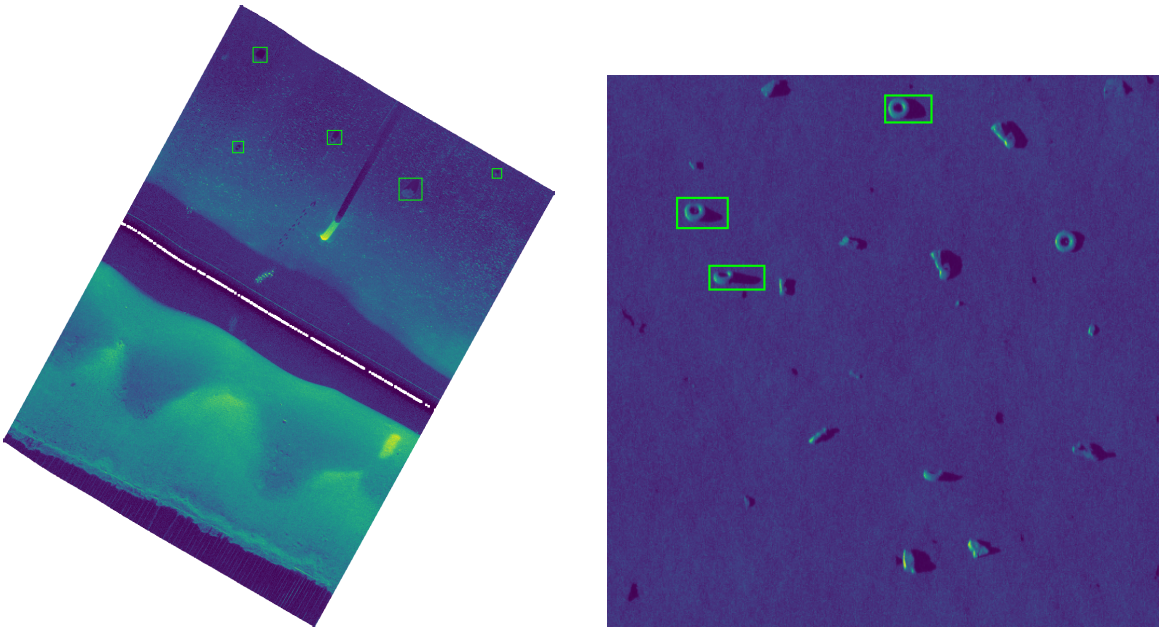
2.2.4 LKU-Net

Similar to Voxelmorph, LKU-Net[8] uses a U-Net to predict a displacement field. However, the convolutional layers in the downsampling path of the U-Net are replaced with a so-called large kernel (LK) block. An LK block processes the input using three parallel convolutional layers with different kernel sizes as well as an identity layer. Afterwards, the four output feature maps are combined using an element-wise addition. This allows the models to have a large receptive field and thus capture spatially distant information while also keeping information at a fine scale. According to[8], the kernel sizes in the LK block are set to 1×1, 3×3 and 7×7. The architecture of LKU-Net consists of five downsampling stages. The overall network is trained for 92 epochs with the Adam optimizer and a learning rate of 0.0001, which correspond to the settings of the original work, which we found to perform best in our study.

3 DATASETS

Over the course of several sea and harbour expeditions, we have collected sidescan sonar data with the SeaCat AUV. The Edgetech 2205 sidescan sonar mounted on the SeaCat AUV operates at a centre frequency of 850 kHz with a bandwidth of 45 kHz. An experimental signal processing chain is used to generate the mosaic images with a pixel resolution of 10 cm. We extend the dataset used in our previous work[10] by an additional measurement campaign in the Baltic Sea explicitly designed for the change detection task. Here the AUV drove the same path twice, but in the second run an object was placed on the seafloor. Table 1 summarises the six experiments that are used in this work to test the change detection processing chains with the different co-registration methods. Furthermore, to enlarge the test dataset, we use the five simulated image pairs from[10]. Figure 2 shows one real and one simulated image from the test dataset with the changes to be detected marked by the green bounding box.

For setting up the dataset used to train pix2pix, Voxelmorph and LKU-Net it is not necessary that the sonar images to contain objects since a general transformation between two images from slightly different AUV paths should be learned. This enlarges the amount of available sonar images for this task. When setting up this training dataset, we selected mission pairs that not necessarily had the exact same AUV path. The paths may vary with respect to the latitude and longitude position but also regarding the depth. This should enable the models to learn different non-linear shifts occurring between two sonar images. We ensured that the overlap of the covered area in the paired images is high enough to feed the models a sufficient amount of information. Only image pairs with an intersection over union, measured using the



(a) ***(b)***

Figure 2: Example of sidescan sonar images from the test dataset. (a) Mosaic image $\tilde{I}_{A,t_1}$. (b) Simulated image. Inserted objects are marked with a green bounding box.

coordinates of the georeferenced images, of at least 0.6 are considered. Also, there is no overlap with images in the test dataset.

Since the deep learning models need to be trained on images of a fixed size, the mosaic sidescan sonar images are split into patches. Before generating these patches, the paired images are coarsely aligned using homography, which corresponds to the first step of the change detection processing chain. To ensure a high amount of information per image, we set the size of the patches, and thus also the input size of the deep learning models, to be 512×512. The patches were extracted with an overlap of 128 pixel. Since due to the mosaicking of the original sonar images some patches have a high number of empty pixel, e.g., patches in the corners of the image in Figure 2a, the patches we manually verified. In total the training dataset contains 847 paired patches form 71 sidescan sonar image pairs captured at four different locations.

4 EXPERIMENTAL RESULTS

The main purpose of improving the co-registration step is to improve the overall performance of the change detection processing chain. Thus, we compare the performance of the processing chains with the different alignment methods by evaluating the number of true-positive (TP) and false-positive (FP) detections. We calculate the distance from the centre of the ground truth (GT) and predicted bounding box and count a TP if this distance is smaller than 1 m. To further quantify the quality of the alignment, we consider the percentage of active pixels in the binary image I_{Bin} after applying the threshold to the difference image I_{Diff} excluding areas with actual changes. Marking these areas using an $M \times N$ masking image I_{Mask}, containing zeros inside the ground truth bounding boxes and ones outside, this

Table 2: Performance of the overall change detection processing chain for different co-registration methods: ORB = ORB feature matching and homography, OF = optical flow, VXM = Voxelmorph, LKU = LKU-Net, P2P = pix2pix. Bold values indicate a false alarm reduction compared to the ORB baseline.

Run	GT	TP						FP					
		ORB	OF	VXM	LKU	P2P	OF+ LKU	ORB	OF	VXM	LKU	P2P	OF+ LKU
Exp 1	5	5	3	4	3	4	3	0	0	0	0	12	0
Exp 2	0	0	0	0	0	0	0	2	3	4	**1**	**0**	4
Exp 3	5	3	3	3	3	2	3	5	**4**	**2**	8	**1**	**4**
Exp 4	6	2	2	2	2	3	2	9	**8**	**6**	11	**4**	**7**
Exp 5	2	0	0	0	0	0	0	27	**17**	33	**16**	86	**18**
Exp 6	1	0	0	0	0	0	0	24	**8**	**17**	**11**	**15**	**7**
Sim 1	3	3	3	3	3	3	3	1	3	3	1	15	2
Sim 2	3	2	3	2	3	3	3	2	2	2	3	14	2
Sim 3	4	4	4	4	4	4	4	2	3	**1**	2	17	2
Sim 4	3	3	3	3	3	3	3	3	6	**2**	**2**	7	7
Sim 5	3	3	3	3	3	3	3	1	4	1	1	6	2

quality factor Q can be computed as

$$Q = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} I_{Bin} \wedge I_{Mask}}{\sum_{n=1}^{N} \sum_{m=1}^{M} I_{Mask}}. \tag{3}$$

The lower this number the better the co-registration.

Table 2 summarises the detection results for the six real and five simulated experiments. Except for the first real experiment, where the additional co-registration methods led to a slight drop in the number of TP detections, this number stays very constant. In the most complex scenarios Exp 5 and Exp 6 none of the designed processing chains detects the object. However, the high number of false alarms in these two cases can be reduced using optical flow and LKU-Net. Due to the good performance of optical flow and LKU-Net we additionally combine them into a stacked alignment routine. For the method OF+LKU the images are first coarsely aligned using homography, then optical flow and afterwards the LKU-Net is applied. In nearly all test runs this results in a lower number of false alarms compared to the co-registration with only the homography. Pix2pix shows a diverse detection result as for Exp 2 - 4 this method generates the least number of false alarms. However, for Exp 1, Exp 5 and the simulated images this number is very high.

It is important to note that looking at the number of false alarms alone might be misleading due to the filtering routine mentioned in Eq. 2.1. The areas which are detected as changes in the difference image are filtered by size. If a co-registration method improves the alignment of two sonar images large areas in the difference image are often split into multiple smaller once. Compared to the baseline these areas are not filtered anymore, in contrast to a large area from a bad alignment result, and generate not one but multiple false alarms. Thus to evaluate the quality of the alignment itself, Table 3 lists the aforementioned percentage of active pixel in the difference image Q. In this analysis optical flow shows the best result among the investigated methods. Considering only deep learning based methods, LKU-Net shows the best and pix2pix the worst result. Pix2pix suffers from misalignments between the individual patches

Table 3: Performance of co-registration methods regarding the alignment quality. Best value marked in bold.

Run	Q					
	ORB	OF	VXM	LKU	P2P	OF+LKU
Exp 1	**0.000**	**0.000**	**0.000**	**0.000**	3.786	**0.000**
Exp 2	6.849	1.657	6.363	4.029	9.512	**1.304**
Exp 3	8.164	1.590	7.671	6.915	9.586	**1.224**
Exp 4	10.231	3.562	7.669	7.668	8.742	**2.640**
Exp 5	12.082	6.814	11.788	10.262	23.411	**5.453**
Exp 6	5.133	2.704	4.361	4.507	4.815	**1.855**
Sim 1	1.438	0.093	1.651	0.190	1.478	**0.051**
Sim 2	2.888	0.023	2.777	0.554	0.840	**0.022**
Sim 3	5.338	0.181	4.442	2.728	2.187	**0.062**
Sim 4	7.791	1.061	6.428	4.702	2.632	**0.449**
Sim 5	6.171	0.167	5.669	0.548	0.842	**0.103**

after stitching them back together. The other methods do not show this behaviour and no artefacts at the border of two patches are visible. Combining optical flow with LKU-Net leads to a further improvement of alignment quality. Regarding the computational overhead, optical flow takes 5.16 s per image of size 512×512 and LKU-Net 0.387 s. Thus, the additional improvement comes at a rather low cost.

Compared to medical images there is more clutter noise in sonar images making the registration task more complex. In addition, although magnetic resonance images of brains from different people look different, they still show the same thing. Regarding sonar images there are more fundamental differences between images from different locations, e.g., ships or quay walls in an harbour environment, which are not present in images from a lake. Thus the deep learning model has to capture a wider variety of changes with respect to style as well as spatial dimension. Our results show that the investigated deep learning models although leading to an improvement over a simple coarse alignment either need to be trained with even more data or be adapted to better capture large and very fine deformations in order to deal with this challenging circumstances.

5 CONCLUSION

In this work the deep learning methods pix2pix, Voxelmorph and LKU-Net were investigated for the co-registration of two sidescan sonar images in a change detection processing chain. Comparing the alignment quality and the effect on the detection performance of the overall process, LKU-Net shows the best and pix2pix the worst results. However, optical flow outperforms LKU-Net. Nevertheless, due to the fast processing speed of LKU-Net this method can be added as an additional alignment step after optical flow to further improve the result. Still not all objects are detected and false alarms occur leaving room for further improvements in all steps of the change detection processing chain. The first step towards this goal is the enlargement of the dataset since it is expected that larger training datasets improve the performance of deep learning models.

REFERENCES

1. F. Nicolas, A. Arnold-Bos, I. Quidu and B. Zerr, “Symbolic Simultaneous Registration and Change Detection Between Two Detection Sets In the Mine Warfare Context”, *OCEANS 2019 MTS/IEEE Marseile*. (2019).
2. E. Coiras, J. Groen, D. Williams, B. Evans and M. Pinto, “Automatic Change Detection for the Monitoring of Cluttered Underwater Areas”, *Proceedings of the 1st International Conference and Exhibition on Waterside Security*, 99-105. (2008).
3. M. Gendron, M. Lohrenz and J. Dubberley, “Automated change detection using Synthetic Aperture Sonar imagery”, *OCENS 2009 MTS/IEEE Biloxi*, 1-4. (2009).
4. J. Ferrand and N. Mandelert, “Change detection for MCM survey mission”, *Proceedings of the 2012 International Conference on Detection and Classification of Underwater Targets*, 193-206. (2012).
5. V. Myers, A. Fortin and P. Simard, “An Automated Method for Change Detection in Areas of High Clutter Density using Sonar Imagery”, *Proceedings of the 3rd International Conference and Exhibition on Underwater Acoustic Measurements*, 287-294. (2009).
6. Ø. Mdtgaard, R.E. Hansen, T.O. Saebo, V. Myers, J.R. Dubberley and I. Quidu, “Change detection using Synthetic Aperture Sonar: Preliminary results from the Larvik trial”, *OCEANS 2011 MTS/IEEE Kona* 1-8. (2011).
7. G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag and A.V. Dalca, “VoxelMorph: A Learning Framework for Deformable Medical Image Registration”, *IEEE Transactions on Medical Imaging*, Vol. 38, No. 8, 1788-1800. (2019).
8. X. Jia, J. Bartlett, T. Zhang, W. Lu, Z. Qiu and J. Duan, “U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration?”, *Machine Learning in Medical Imaging*, 151-160. (2022).
9. J. Chen, E.C. Frey, Y. He, W.P. Segars, Y. Li and Y. Du, “TransMorph: Transformer for unsupervised medical image registration”, *Medical Image Analysis*, Vol. 82. (2022).
10. Y. Steiniger, S. Schröder and J. Stoppe, “Reducing the false alarm rate of a simple sidescan sonar change detection system using deep learning”, *Proceedings of Meetings on Acousics*, Vol. 47. (2022).
11. T. G-Michael, B. Marchand, J.D. Tucker, T.M. Marston, D.D. Sternlicht and M.R. Azimi-Sadjadi, "Image-Based Automated Change Detection for Synthetic Aperture Sonar by Multistage Coregistration and Canonical Correlation Analysis", *IEEE Journal of Oceanic Engineering*, Vol. 41, No. 3, 592-612. (2016).
12. R. Klemm, J. Groen, H. Schmaljohan, “Interoperable image-based change detection”, *International Conference on Synthetic Aperture Sonar and Synthetic Aperture Radar 2023*. (2023).
13. P. Isola, J.-Y. Zhu, T. Zhou and A.A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks", *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 5967-5976. (2017).
14. E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF", *2011 International Conference on Computer Vision*, 2564-2571. (2011).
15. C. Zach, T. Pock and H. Bischof, “A duality based approach for realtime TV-L 1 optical flow”, *Joint pattern recognition symposium*, 214-223. (2007).
16. O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, *Medical Image Computing and Computer-Assisted Intervention*, Vol. 9351, 234-241. (2015).