Non-Linear co-registration in UAVs' images using deep learning

Leandro Henrique Furtado Pinto Silva*†, Jocival Dantas Dias Júnior*, João Fernando Mari†,
Mauricio Cunha Escarpinati* and André Ricardo Backes*
*School of Computer Science, Federal University of Uberlândia
Av. João Naves de Ávila, 2121, 38408-100 Uberlândia, MG, Brazil
Email: *arbackes@yahoo.com.br*
†Federal University of Viçosa, Campus Rio Paranaíba, MG, Brazil

***Abstract*—Unmanned Aerial Vehicles (UAVs) has stood out for assisting, enhancing, and optimizing agricultural production. Images captured by UAVs allow a detailed view of the analyzed region since the flight occurs at low and medium altitudes (50m to 400m). In addition, there is a wide variety of sensors (RGB cameras, heat capture sensors, multi and hyperspectral cameras, among others), each with its own characteristics and capable of producing different information. In multi-spectral images acquisition, we use a distinct sensor to capture each image band and at different time, leading to misalignments. To tackle this problem we propose to train a deep neural network to predict the vector deformation fields to perform the registration between bands of a multi-spectral image. The proposed approach has an accuracy ranging from 89.90% to 93.79% in the task of estimating the displacement field between bands. With this field estimated by the network, it is possible to register between the bands without the need for manual marking of points.**

I. Introduction

Agriculture has undergone major changes in recent centuries. Such changes would directly impact the history of humanity. It is worth noting that until the end of the 18th century, there was much pessimism about the Earth's ability to feed a population that grew exponentially. Pessimism, fortunately, did not materialize to catastrophic levels due to advances in agriculture. We can summarize these advances in three great revolutions: first, the mechanization of farms (1900 to 1930), later, the genetic modification of plants (1990 to 2005), and we are currently experiencing the third revolution, defined as Precision Agriculture (PA) [1]–[3].

In general, we understand PA as the individualized treatment of each area that makes up a particular crop. Thus, the individualized treatment allows, for example, the effective application of agricultural inputs, which generates a reduction in costs and an increase in the gains of the agricultural producer. In this sense, PA is heavily dependent on imaging and mapping techniques. Images can be captured from different sources such as satellites, Unmanned Aerial Vehicles (UAVs), and even smartphones [4].

The use of UAVs in the PA has become popular due to the possibility of capturing images from low and medium distances (especially when compared to satellites). Additionaly, it is possible to couple several sensors in a single aircraft, thus enabling us to capture multispectral images. Figure 1 presents an example of a scene containing five bands. However, because the flight of a UAV is subject to environmental impacts (e.g., wind, shadows) and the different sensors are often placed in different parts of the aircraft, it is common misalignments between the bands of the same captured scene. Such misalignments can be grouped into two large groups: (i) linear, which involves a pre-established pattern (e.g., rotation, scale, and perspective change), and (ii) non-linear, where there is no established pattern of deformation. Figure 2 presents an image composed of two bands, through a checkerboard (where light and dark tones differentiate different bands). In this figure, it is possible to see the misalignment between the bands of the same scene. [5]–[9].

For correct analysis of images for decision-making in PA, the bands must be aligned. Thus, the present work presents a proposal based on U-Net to perform alignment between the bands of the same scene without manual marking of points. For this, we considered two datasets and trained a deep-learning approach to the task of estimating the displacement field between bands of the same scene. [11], [12].

II. Materials and methods

A. Deep learning

Deep Learning consists of an area of Artificial Intelligence that simulates the human brain through an artificial neural network containing more than one hidden layer. Through Deep Learning it is possible to perform problems of, for example, classification, segmentation, or regression, in different contexts (e.g., images, text, speech) [13].

Convolutional Neural Networks (CNNs) are a category of deep learning methods. These networks use the concept of the receptive field of biological systems, which gives them networks the ability to learn different filters and features of an image. In this way, CNNs can explore the spatial correlations between pixels in image to extract meaningful image attributes for many tasks, such as image classification and segmentation. In short, a typical CNN contains three types of layers: convolutional, pooling, and dense. The convolutional layer is responsible for extracting significant features from an image. The pooling layer aims to reduce the feature maps

978-1-6654-5385-1/22/$31.00 ©2021 IEEE

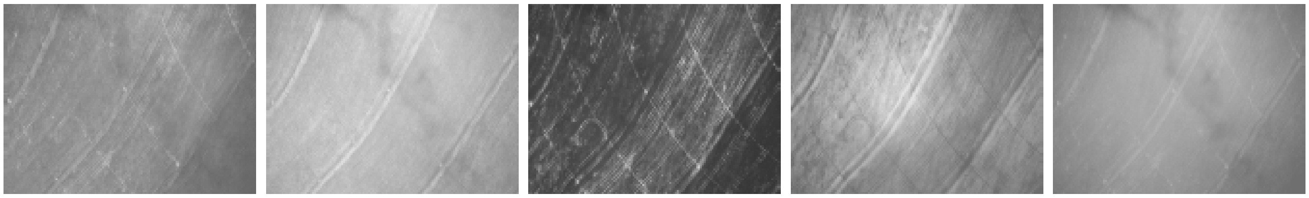

Fig. 1. Example of a scene containing all bands: Blue, Green, Red, Near-Infrared, and Red Edge, respectively. Adapted from [10].

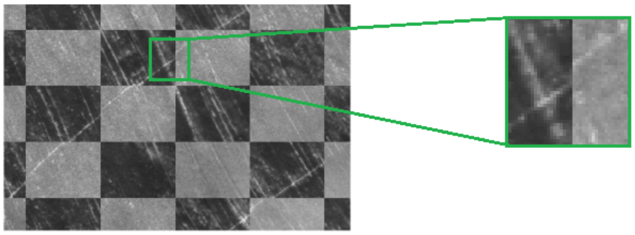

Fig. 2. Checkerboard of two bands of the same image. Note that in highlight there is a misalignment between the bands, where the same line, in its different bands, should follow the same diffraction [10].

computed by the convolutional layer. Finally, a flattening operation converts the feature maps (two or more dimensions - 2D or more) obtained in the previous operations into a one-dimensional vector (1D) which is processed by the dense layer [14]–[16].

U-Net is a CNN proposed by [11] for medical image segmentation. The main idea of U-Net is to apply pooling and upsampling operations. In general terms, the pooling operation reduces the size of the input to highlight the main features. The upsampling operation will progressively expand the feature maps identified in previous pooling operations until the output has the same size as the input [17], [18].

B. Image Dataset

Two different experts manually marked the images in the datasets in order to align them, thus making it a ground truth for our experiments. Each expert worked on a different dataset. In general, the experts marked 12 points in the green band. In the sequence, the experts selected the equivalent points in the other bands [10]. Tables I and II shows the misalignment, in pixels, between the ground truth generated by the experts in the soybean and cotton datasets, respectively.

TABLE I
MISALIGNMENT AVERAGE, IN PIXELS, BETWEEN THE SENSORS PRESENT IN SOYBEAN DATASET.

	Blue	Green	Red	NIR	Red Edge
Blue	-	5.18	15.75	15.07	12.14
Green	5.18	-	15.09	12.33	4.02
Red	15.75	15.09	-	29.25	14.79
NIR	15.07	12.33	29.25	-	16.17
Red Edge	12.14	4.02	14.79	16.17	-

TABLE II
MISALIGNMENT AVERAGE, IN PIXELS, BETWEEN THE SENSORS PRESENT IN COTTON DATASET.

	Blue	Green	Red	NIR	Red Edge
Blue	-	28.30	12.11	33.28	8.48
Green	28.30	-	21.44	24.01	35.66
Red	12.11	21.44	-	14.94	21.30
NIR	33.28	24.01	14.94	-	39.37
Red Edge	8.48	35.66	21.30	39.37	-

For the experiments, we considered two dataset to evaluate the proposed approach. Both datasets have images of 1280×960 pixels size and an average of 80% overlap between images. We obtained the first dataset from a soybean plantation. Soybean dataset contains 1080 images (216 scenes and 5 bands). The second dataset is from a cotton plantation. Cotton dataset contains 830 images (166 scenes and 5 bands). The bands present in both datasets are Blue, Green, Red, Near-Infrared (NIR), and Red Edge. We obtained the images in a single flight without any type of pre-processing, and it took place at an average height of 100 meters, at an average speed of $20m/s$. Under these conditions, the Ground Sample Distance (GSD) is $6.8cm/pixel$ [10].

Due to the nature and purpose of the UAV flight, both dataset have images of heterogeneous content. Figure 3 shows some image samples of the soybean dataset and Figure 4 shows image samples of the cotton dataset, after after the generation of the ground truth.

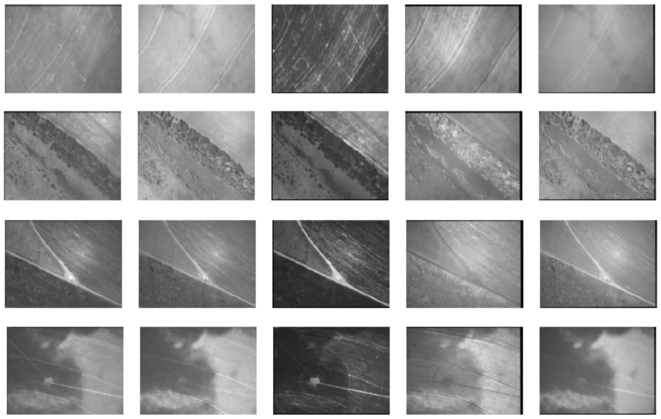

Fig. 3. Examples of images from soybean dataset: each row represents a scene captured by UAV. Columns represent the respective bands (From lef to to right: Blue, Green, Red, NIR, and Red Edge, respectively).

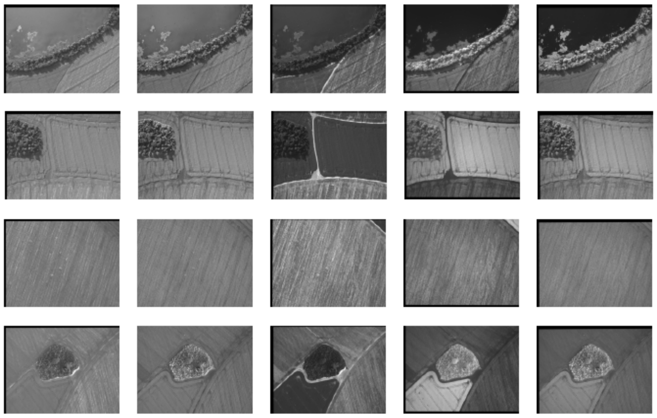

Fig. 4. Examples of images from cotton dataset: each row representsa scene captured by UAV. Columns represent the respective bands (From lef to to right: Blue, Green, Red, NIR, and Red Edge, respectively).

To perform our experiments, we have proportionally reduced the images to 20% of their original size, preserving their content and aspect ratio so that we do not insert any deformation. We performed this operation to reduce the computational cost of further training in relation to their original sizes. In this way, the images will be 256×192 pixels size.

We consider the green band as fixed as in [10]. The authors justifies this choice because the images are usually composed of vegetation, so this band has more content that can be used as reference object for matching with other bands. Thus, we generated deformations in the other bands and we kept the green band preserved for both datasets.

The non-linear deformations consisted of two random grids, one for displacements in the *y-direction*, and the other for displacements in the *x-direction*. In short, the transformation is defined through a 3×3 point B-spline grid, where random displacements are ranging from a uniform distribution with $x, y = [-0.05, 0.05]$ [19].

To make the generated deformations more adherent with real deformations, we also inserted two more rigid deformations often found in a UAV flight: rotation and scale. These deformations were also randomly computed in a controlled range of variations. For the scale, we consider variations of $\pm 2\%$, to the original. For the rotation, we consider a variation of up to 10 degrees around the central point. Figure 5 shows the artificially generated deformations.

In our approach we used supervised learning to learn the displacement field (U) of the moving image to the fixed image. The displacement field (U) is calculated as follows:

$$U = originalGrid - transformedGrid \qquad (1)$$

where *originalGrid* refers to the deformation-free image grid (original image) and *transfomedGrid* refers to the image grid after inserting the deformations. The visual presentation of the displacement fields, with the respective color map, is shown in Figure 6.

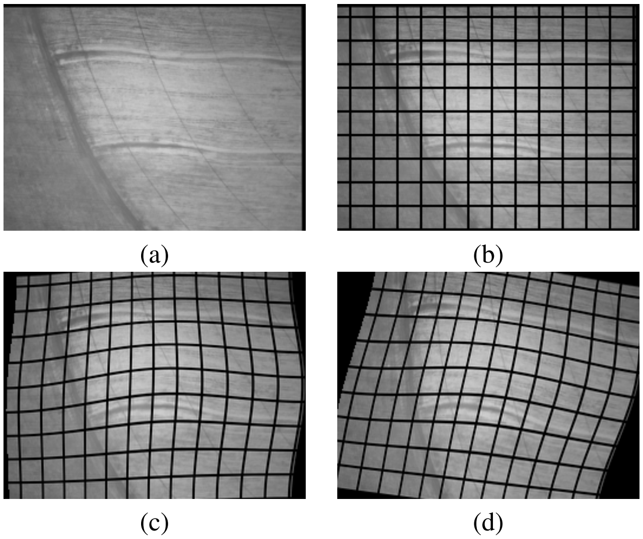

Fig. 5. Example of distortion in near-infrared band. Note that the grids in this figure are for a better visual representation of the deformation and do not represent in size the B-spline transformation grid.

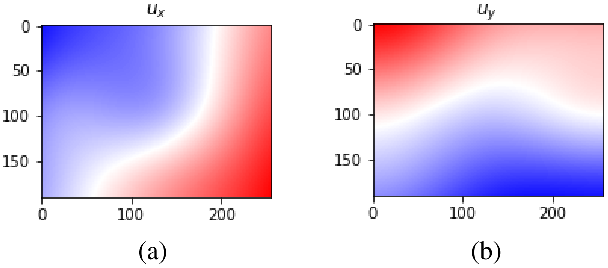

Fig. 6. Example of displacement fields in Figure 5-(d) in relation to Figure 5-(b): (a) displacement field for the x axis; (b) displacement field for the y axis. The colors blue and red represent positive and negative displacements, respectively.

C. End-to-end training

We trained our network to maximize the accuracy between the estimated vector field ($\hat{U}$) and the true vector field (U), in accordance with [12]. TensorFlow framework considers that two floating point numbers equals if difference between them is smaller than 10^{-6}. Thus, we define the accuracy as

$$\text{accuracy} = \frac{100}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta(U_{ij}, \hat{U}_{ij}) \qquad (2)$$

with

$$\delta(i,j) = \begin{cases} 1, & |i-j| <= 10^{-6} \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

In addition, the loss is defined as:

$$loss = \frac{1}{n} \sum_{j=1}^{n} |u_j - \hat{u}_j| \qquad (4)$$

where n refers to number of iterations. We also use the Adam optimizer with a learning rate in 10^{-4}.

According to [12], to adapt the U-Net to the registration task, the input layer has two bands, one for the fixed band and the other for the moved band. The output layer is also adapted to have two bands, where each band corresponds to an axis of the vector field (x and y). Finally, all activation functions are the Leaky ReLU parameterized as $\phi(x) = max(x, 0.01x)$, which restrict neurons from being nulled (zero value) in some applications of the ReLU function [19]. However, the output layer has no activation function to enable real values. Figure 7 shows the proposed architecture.

III. Results and discussion

We conducted the experiments on a Personal Computer with Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz, 32GB RAM, 64-bit Windows OS and GPU NVIDIA GeForce GTX 1050 Ti, 4GB GDDR5. We also used Python 3.6 and Keras 2.1.6-tf with TensorFlow 1.10.0 and CUDA Toolkit 9.0. In this experimental setup, our approach has about 8.5 million parameters. We divided each dataset into training, validation, and test sets in the proportion of 70%, 15%, and 15%, respectively. We performed the training of the network for 2000 epochs.

In our network the input consists of two band, being the green the fixed one. Thus we evaluated how well the other four bands align towards the green band. In short, we have the following combinations: (i) Green and Blue, (ii) Green and Red, (iii) Green and Red Edge, and (iv) Green and NIR. The network output, as previously mentioned, is the displacement field of the band moved in relation to the fixed band.

For the soybean dataset, for each band deformed (Blue, Red, NIR, and Red Edge), we obtained an accuracy ranging from 89.90% to 93.79%. For the cotton dataset, with the same deformed bands as the soybean dataset, we obtained an accuracy ranging from 90.01% to 91.21%. Table III summarizes the accuracy for each of the datasets.

TABLE III
Summary of accuracy for each of the moved bands considered.

Deformed Band	Accuracy	
	Soybean Dataset	Cotton Dataset
Blue	89.90%	90.01%
Red	89.95%	90.11%
NIR	90.50%	91.21%
Red Edge	93.79%	90.54%

After training the network, we evaluated its accuracy using a set of images that were not part of the training and validation sets. With the vector fields predicted by the network, we need to map the points of the deformed image towards the fixed image. In short, we applied the transformation learned by the network, interpolating them in a new grid, as follows:

$$I_R(x, y) = I_M(x + Z(x, y), y + V(x, y)) \tag{5}$$

where I_R is the registered band, I_M is the moved band and $Z(x, y)$ and $V(x, y)$ are the predicted vector fields for the x and y axis, respectively.

We also present a visual analysis of our approach. For that, we present the overlap between the bands before and after applying our approach. Figure 8 shows this visual analysis for the soybean dataset. Figure 9 shows the same visual analysis, but for the cotton dataset. Note that these figures for visual analysis have a highlight, where it is possible to observe the ability of our approach to align the images in the correct direction.

IV. Conclusions

Similar to [19], in this work we adapted the U-Net to problems of co-registration in UAV images in the PA. In this context, it was possible to estimate a deformation field between bands of multispectral images with considerable precision. The accuracy, when considering both datasets, ranged from 89.90% to 93.79%. In this way, our network can perform co-registration after training without manually marking the points.

In this sense, we need to highlight that although the approach behaved similarly in both datasets, we obtained them under different flight conditions and in different crops, demonstrating the ability of our approach to generalize.

Although our proposal is related to that of [12], we take into account multispectral images. In addition to the content of each one of the images in our training set being totally different, in other words, there is no defined pattern of content between our images (see Figures 3 and 4). In [12], the images are exclusive of lungs captured through a CT.

In [20], the co-registration between the same bands considered in this work is carried out, however, [20] does not consider the presence of non-linear deformations, besides needing the manual marking of points for the co-registration task.

Finally, through visual analysis, we can see the ability of our approach to performing co-registration on multispectral images. This capacity is even more evident due to the degree of non-linear deformation considered, in addition to the inclusion of linear deformations, which increases the complexity of learning the network. Therefore, from the scenes aligned with our approach, the images can be used more effectively in the PA for different tasks (e.g., vegetation indices, detection of planting lines, and segmentation).

Acknowledgment

André R. Backes gratefully acknowledges the financial support of CNPq (National Council for Scientific and Technological Development, Brazil) (Grant #307100/2021-9). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) – Finance Code 001.

References

[1] T. R. Malthus, *An Essay on the Principle of Population..*, 1872.

[2] P. B. Hazell, *The Asian green revolution*. Intl Food Policy Res Inst, 2009, vol. 911.

[3] B. Farmer, "Perspectives on the 'green revolution'in south asia," *Modern Asian Studies*, vol. 20, no. 1, pp. 175–199, 1986.

[4] D. Jenkins and B. Vasigh, *The economic impact of unmanned aircraft systems integration in the United States*. Association for Unmanned Vehicle Systems International (AUVSI), 2013.

[5] A. McBratney, B. Whelan, T. Ancev, and J. Bouma, "Future directions of precision agriculture," *Precision agriculture*, vol. 6, no. 1, pp. 7–23, 2005.

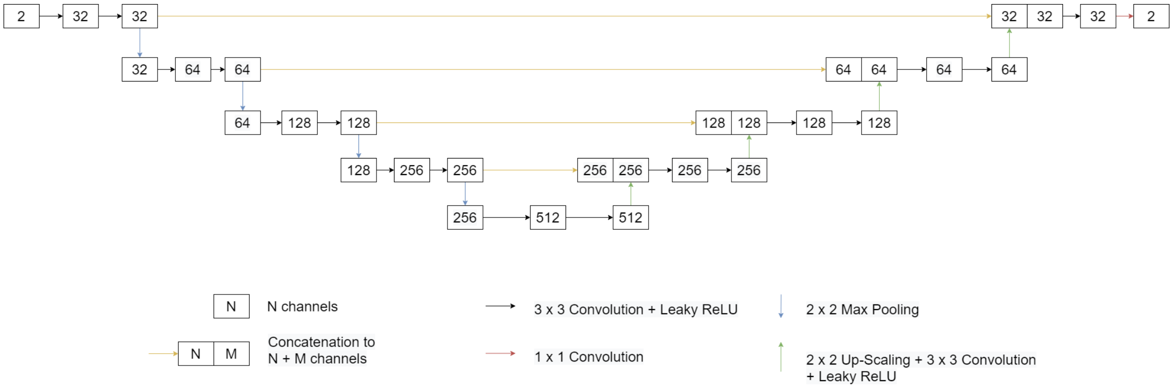

Fig. 7. Proposed network architecture. The network takes two bands as input, and outputs two maps: one for each vector field axis. Adapted from [12].

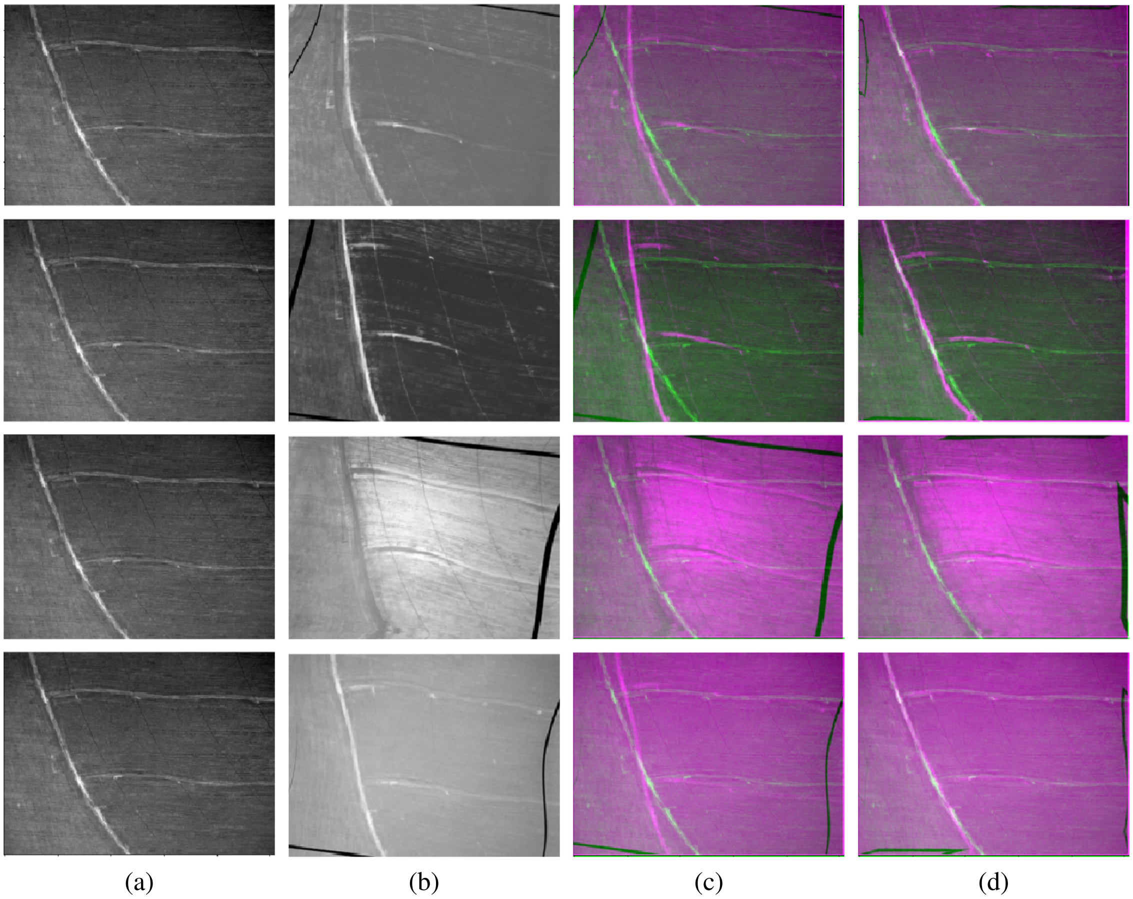

Fig. 8. Example of co-registration in soybean dataset: (a) fixed band (Green); (b) moving band; (c) overlap of the bands; (d) overlap of the bands after the application of the field predicted by the network. Moving band, from top to bottom: Blue, Red, NIR and Red Edge.

[6] A. Milella, G. Reina, and M. Nielsen, “A multi-sensor robotic platform for ground mapping and estimation beyond the visible spectrum,” *Precision agriculture*, vol. 20, no. 2, pp. 423–444, 2019.

[7] T. M. Blackmer and J. S. Schepers, “Aerial photography to detect nitrogen stress in corn,” *Journal of Plant Physiology*, vol. 148, no. 3-4, pp. 440–444, 1996.

[8] S. Sankaran, L. R. Khot, C. Z. Espinoza, S. Jarolmasjed, V. R. Sathuvalli, G. J. Vandemark, P. N. Miklas, A. H. Carter, M. O. Pumphrey, N. R. Knowles *et al.*, “Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping: A review,” *European Journal of Agronomy*, vol. 70, pp. 112–123, 2015.

[9] T. Kataoka, T. Kaneko, H. Okamoto, and S. Hata, “Crop growth estimation system using machine vision,” in *Proceedings 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003)*, vol. 2. IEEE, 2003, pp. b1079–b1083.

[10] J. D. Dias Junior *et al.*, “Uav-multispectral sensed data band co-registration framework,” 2020.

[11] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] K. A. Eppenhof, M. W. Lafarge, M. Veta, and J. P. Pluim, “Progressively trained convolutional neural networks for deformable image registration,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1594–

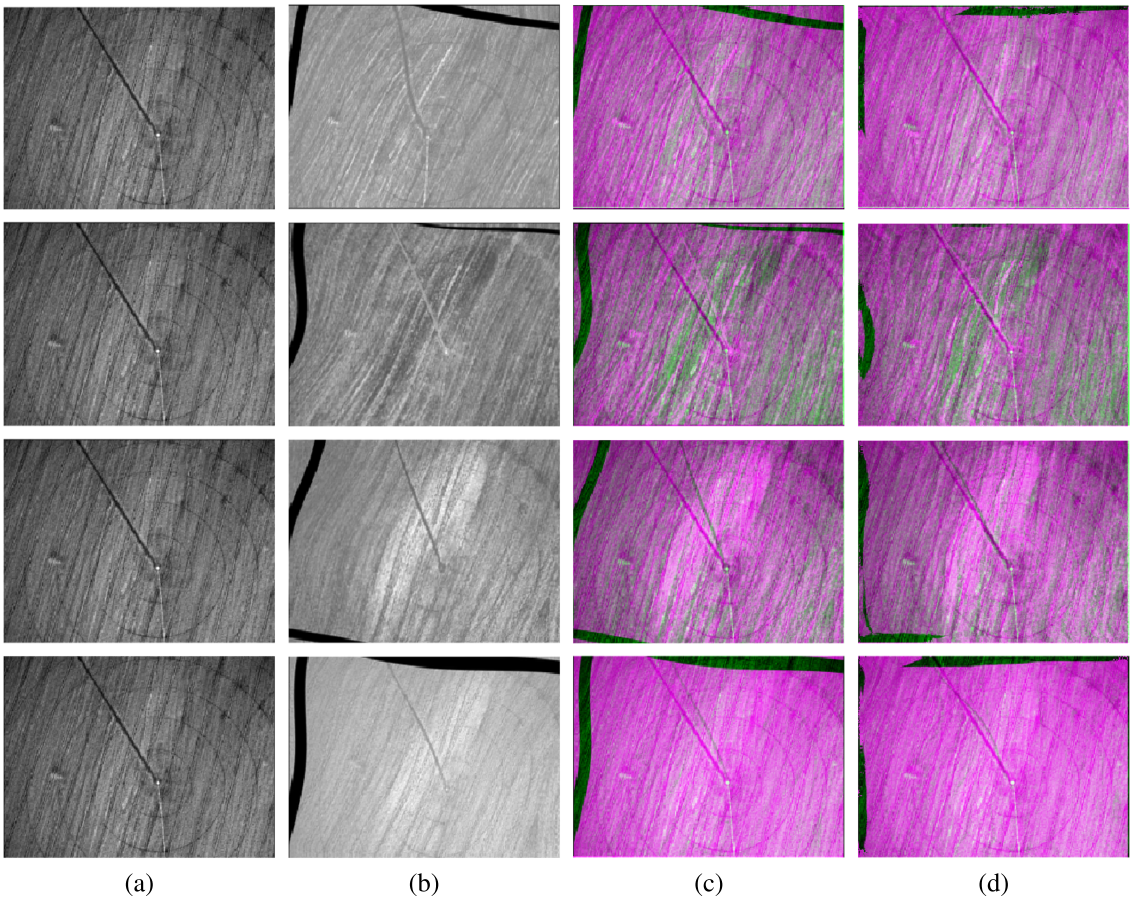

Fig. 9. Example of co-registration in soybean dataset: (a) fixed band (Green); (b) moving band; (c) overlap of the bands; (d) overlap of the bands after the application of the field predicted by the network. Moving band, from top to bottom: Blue, Red, NIR and Red Edge.

1604, 2019.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[14] Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review,” *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[16] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazaré, T. Bui, and J. Collomosse, “Everything you wanted to know about deep learning for computer vision but were afraid to ask,” in *SIBGRAPI Tutorials*. IEEE Computer Society, 2017, pp. 17–41.

[17] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[18] S.-J. Heo, Y. Kim, S. Yun, S.-S. Lim, J. Kim, C.-M. Nam, E.-C. Park, I. Jung, and J.-H. Yoon, “Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers’ health examination data,” *International journal of environmental research and public health*, vol. 16, no. 2, p. 250, 2019.

[19] K. A. J. Eppenhof and J. P. W. Pluim, “Pulmonary ct registration through supervised learning with convolutional neural networks,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1097–1105, 2019.

[20] J. D. Dias Junior, A. R. Backes, M. C. Escarpinati, L. H. F. Pinto Silva, B. C. S. Costa, and M. H. F. Avelar, “Uav-multispectral sensed data band co-registration framework,” in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 223–228.