Translation of digital to multispectral images for automatic plant postharvest decay prediction at early stages to improve health safety (Deep Learning 2021 Course)

Alexander Telepov [1] **Nikita Stasenko** [1]

Abstract

Food quality is a crucial aspect of human health and food quality control helps to avoid heath diseases affected by food poisoning. Optical and hyperspectral imaging are widely used in postharvest biology and technology for fruit quality control while common digital images are useless for decay detection on early stages. However optical and hyperspectral imaging are more costly and time consuming than default digital imaging. This motivates us to investigate the ability of digital to multispectral images translation.

Github repo: github.com
Video presentation: video

1. Introduction

Hyperspectral/Multispectral imaging (HSI/MSI) technologies allow to work with tens to hundreds of spectral channels within the electromagnetic spectrum. Such technologies also exceed the capabilities of human vision because they are based on the principle that every material has a different response (reflection and absorption) to different wavelengths (Grahn & Geladi, 2007). So, it made possible to discriminate different materials between each other at various spectrum ranges and level scales. HSI/MSI imaging methods are applied for different applications. For example, in geology HSI is applied to identify the location of different minerals (Brossard et al., 2016). In medical science MSI/HSI imaging is applied for abdominal organs identification (Akbari et al., 2008), colorectal surgey (Schols et al., 2015), bowel anastomosis detection (Wirkert et al., 2016), billary anatomy identification (Zuzak et al., 2008), Intestinal Ischemia identification (Akbari et al., 2010), gastric cancer identification (Hohmann et al., 2017). Finally, in digital agriculture MSI is applied for crop monitoring (de Oca et al., 2018), postharvest fruit monitoring and decay detection (Li et al., 2019), (Tian et al., 2020), weed detection and mapping (Sa et al., 2018), soil analytics and mapping (Zhou et al., 2020),(Guo et al., 2021).

Food quality is a crucial aspect of human health and food quality control helps to avoid heath diseases affected by food poisoning. Optical and hyperspectral imaging are widely used in postharvest biology and technology for fruit quality control (Lorente et al., 2012), (Qin et al., 2013), (Lu et al., 2020) while common digital images are useless for decay detection on early stages. However optical and hyperspectral imaging are more costly and time consuming than default digital imaging. This motivates us to investigate the ability of digital to multispectral images translation.

Problem of NIR channels restoration from RGB channels is ill conditioned: In arbitrary data case it's impossible to restore signal in infrared spectral bands from visible bands because they almost not intersect. Although for particular data this it may be possible due to some data-specific dependencies.

So, the main purpose of this project is to investigate ability to translate digital images to multispectral. The similar studies were done for satellite images in (de Lima et al., 2019), where estimation of Near Infrared (NIR) bands from a low-cost and well-known RGB camera was presented.

2. Preliminaries

Multispectral image is one that captures image data within specific wavelength ranges across the electromagnetic spectrum. The wavelengths may be separated by filters or detected via the use of instruments that are sensitive to particular wavelengths, including light from frequencies beyond the visible light range, i.e. infrared and ultra-violet. Spectral imaging can allow extraction of additional information the human eye fails to capture with its visible receptors for red, green and blue. In Figure 1 color sensor, which able to capture 3 spectral bands(red, green, blue) and multispectral sensor which capture 4 spectral bands(red, green, blue, yellow), is plotted.

3. Related work

Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between

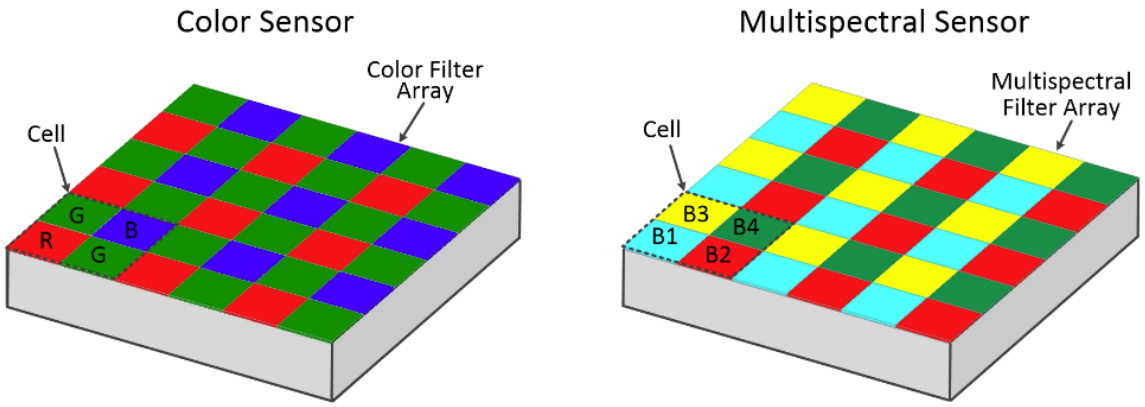


Figure 1. Color and multispectral sensor configurations

an input image and an output image using a training set of aligned image pairs. Common approach for image-to-image translation task - train GAN model.

The GAN architecture is comprised of a generator model for outputting new plausible synthetic images, and a discriminator model that classifies images as real (from the dataset) or fake (generated). The discriminator model is updated directly, whereas the generator model is updated via the discriminator model. As such, the two models are trained simultaneously in an adversarial process where the generator seeks to better fool the discriminator and the discriminator seeks to better identify the counterfeit images.

3.1. Pix2Pix

The Pix2Pix model (Isola et al., 2018) is a type of conditional GAN, or cGAN, which has been demonstrated on a range of image-to-image translation tasks such as converting maps to satellite photographs, black and white photographs to color, and sketches of products to product photographs. In conditional GANs the generation of the output image is conditional on an input, in case of Pix2Pix, a source image. The discriminator is provided both with a source image and the target image and must determine whether the target is a plausible transformation of the source image.

The generator is trained via adversarial loss, which encourages the generator to generate plausible images in the target domain. The generator is also updated via L1 loss measured between the generated image and the expected output image. This additional loss encourages the generator model to create plausible translations of the source image.

$L_{cGAN}(G, D) = E_{x,y}[logD(x, y)] + E_{x,z}[log(1 - D(x, G(x, z))]$

$L_{L1}(G) = E_{x,y,z}[\|y - G(x, z)\|_1]$

$G = arg\min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G)$

3.2. CycleGan

CycleGan (Zhu et al., 2020) goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y using an unpaired set of data pairs. Because this mapping is highly under-constrained, it coupled with an inverse mapping $F : Y \rightarrow X$ and a cycle consistency loss introduced to enforce $F(G(X)) \approx X$ (and vice versa).

For the mapping function $G : X \rightarrow Y$ and its discriminator D_Y , the objective expressed as follows: $L_{GAN}(G, D_Y, X, Y) = E_y[logD_Y(y)] + E_x[log(1 - D_Y(G(x))]$

$L_{cyc}(G, F) = E_x[\|F(G(x)) - x\|_1] + E_y[\|G(F(y)) - y\|_1]$

$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F)$

$G, F = arg\min_{G,F} \max_{D_X,D_Y} L(G, F, D_X, D_Y)$

3.3. Pix2PixHD

Pix2PixHD (Wang et al., 2018) is a modification of pix2pix solution. There are several improvements : Coarse-to-fine generator, Multi-scale discriminators, Improved adversarial loss.

Generator architecture plotted on Fig3. It consist from two part: local enhancer G2 and global generator G1. During training, first the global generator is trained and then train the local enhancer in the order of their resolutions. After that all the networks fine-tuned jointly. This generator designed to effectively aggregate global and local information for the image synthesis task. For effective detail capturing on multiple scales authors propose to use several discriminators,

which operate with output of generator on different scales.

Significant performance boost was provided by loss modification: two extra terms L_{FM} -feature matching loss and perceptual loss were added L_{VGG} to objective. Authors show that feature matching loss stabilizes the training as the generator has to produce natural statistics at multiple scales

$L_{FM}(G, D_k) = \lambda_{FM} E_{s,x} \sum_{i=1} \frac{1}{N_i} [||D_k^{(i)}(s, x) - D_k^{(i)}(s, G(s))||_1]$ where $D_k^{(i)}$ denotes output of i-th layer of D_k discriminator.

$L_{VGG} = \lambda_{VGG} E_{s,x} \sum_{i=1} \frac{1}{M_i} [||F^{(i)}(x) - F^{(i)}(G(s))||_1]$ where $F^{(i)}$ denotes the i-th layer with M_i elements of the VGG network.

4. Data acquisition

We acquired the data that would be similar to data obtained under natural storage conditions. To do this, we selected 16 apples of four types (sets), cultivated in agriculture and used in food production, for this experiment: "Aleysa", "Smirenko", "Golden", "Fujii". Each set contained four apples: one was not subjected to any processing, one was thoroughly washed and wiped, one was subjected to mechanical deformation, and the other was shock-frozen. We also constructed a special testbed for data collection containing frame, table with apples, multispectral camera and lamps. The distance between table and camera was equal to 500 mm. The lamps allowed us to simulate real storage conditions for apples as collection of images under full and partial illumination. Figure **??** shows the testbed.

Multispectral camera "CMS-V1 CMS18100073" (or just "CMS-V") was used for data collection. The camera allows getting images in range 561 - 838 nm, including visible and NIR range. The CMS camera imager has a modified Bayer matrix made of a group of 3x3 pixels, called Macro-pixel, filtering 3x3 (9) spectral bands. So, this camera allows The raw image delivered by the camera is built of 9 interleaved spectral sub-images (8 colors + 1 Panchromatic). The raw image resolution is 1280x1024 pixels. The resolution of the nine sub-images is 426x339 pixels. The camera was installed in testbed frame and connected to personal PC desktop.

We acquired 145 sequential RGB images and 1305 corresponded multispectral images to see the decay dynamics in presented apples. One RGB image was related to 9 images from spectral bands (channel0 = 561 nm, channel1 = 597 nm, channel2 = 635 nm, channel3 = 673 nm, channel4 = 724 nm, channel5 = 762 nm, channel6 = 802 nm, channel7 = 838 nm, channel8(panchromatic channel) = 0 nm). The example of images are given in Figure 5.

5. Algorithms and Models

In this project we investigate the ability to use Pix2Pix, Pix2PixHD and CycleGan approaches in RGB to multispectral image translation. Concrete usage settings described in **Experiments and Results** section.

On our intuition for such translation spatial spatial information not such important as for common style-transfer, or label-to-image task. If it is true, convolutional architectures may be computationally redundant and tend to overfit. Natural alternative for such case - use convolutions with kernel size 1. Specifically, we add to consideration two simple models consist of basic Resnet convolutional blocks, but one of networks also have additional preprocessing layer - it average pixels on 8 neighbours scheme.

6. Experiments and Results

There are two main setups were considered: 1) translation of RGB channels to all 9 channels of multispectral image 2) translation of RGB channels to 3 NIR channels. Since for decay estimation only NIR channels needed we focused on 2) setting to facilitate learning. Our experiments show that learning in 1) produce poor results for NIR channels: best ssim we obtain this channels 0.60 (and its much lower then ssim for other channels 0.82) while for 2) setting 0.95. All obtained models can be downloaded from google disc.

6.1. Augmentations

Since dataset we have quite small (145 RGB images and 1305 MSI images) we intensively imply augmentations: random rotations, shifts, zoom and flips. We don't use transformations like contrast/brightness adjustments since in can destroy physical information.

6.2. Architectures

Discriminator architecture has following structure: 5 layers convolution - batch normalization - leaky relu. Convolution kernel equal 4, stride 2. Number of features gradually increase from 64 to 512. For pix2pixHD model all discriminators have same architecture.

As in original paper, for CycleGan we used Resnet encoder-decoder architecture. It consist of 2 downsampling layers, 6 Resnet bottleneck blocks and 2 upsampling layers. generator. For Pix2Pix model we use UNet generator with 4 downsampling blocks.

6.3. Optimization and inference

As suggested in the original GAN paper and pix2pix paper, rather than training G to minimize $log(1 - D(x, G(x, z))$, we instead train to maximize $logD(x, G(x, z))$. In addition,

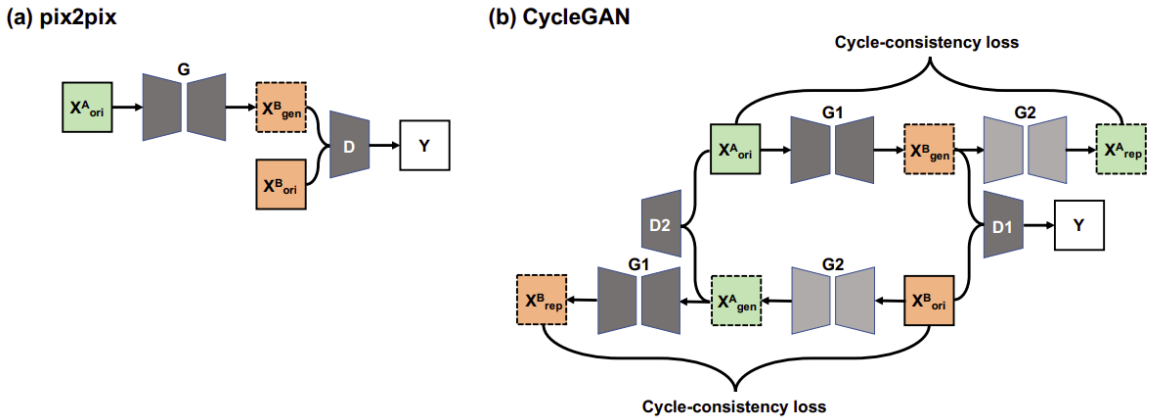


Figure 2. The architecture of (a) Pix2Pix and (b) CycleGAN. (a) pix2pix requires perfectly aligned paired training images. A generator CNN is trained to generate images similar to images in domain B from images in domain A, and discriminator CNN is trained simultaneously to distinguish the generated images from real images in domain B. Reconstruction loss measures how close by the real images in domain B and the generated images. On the other hand, (b) CycleGAN can learn a translation mapping in the absence of aligned paired images. The image generated from domain A to domain B by generator CNN (G1) is converted back to domain A by another generator CNN (G2), and vice versa, in the attempt to optimize the cycle-consistency loss in addition to the adversarial loss.

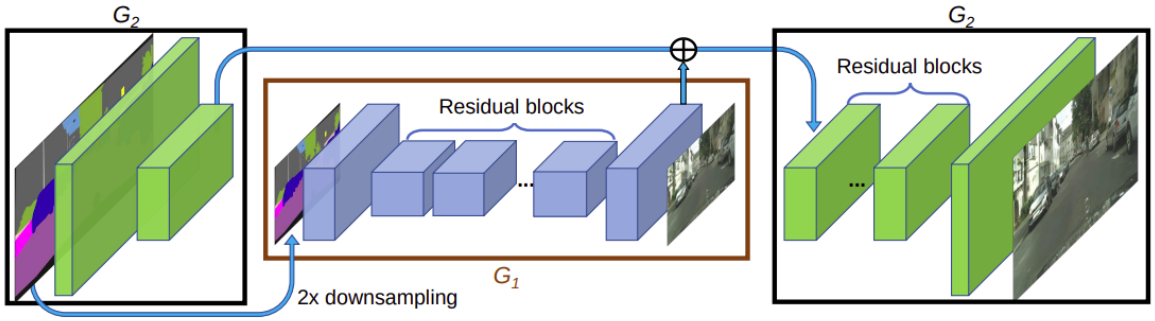


Figure 3. The architecture of Pix2PixHD Generator. First residual network G1 is trained on lower resolution images. Then, another residual network G2 is appended to G1 and the two networks are trained jointly on high resolution images. Specifically, the input to the residual blocks in G2 is the element-wise sum of the feature map from G2 and the last feature map from G1.

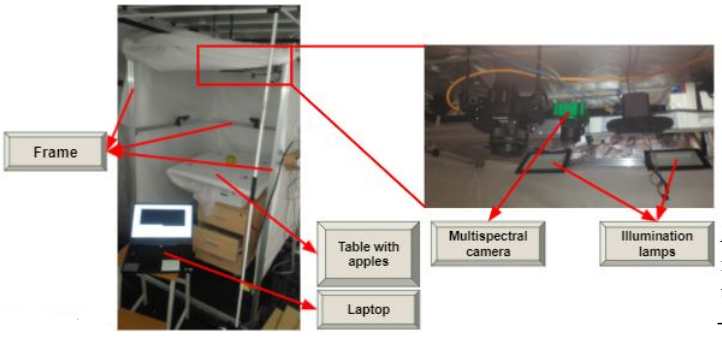


Figure 4. Experimental testbed for images capturing.

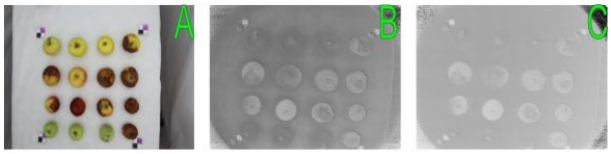


Figure 5. Types of images obtained during experiments: (A) - RGB image of apples acquired under full illumination; (B) - Multispectral image of apples acquired under full illumination (802 nm); (C) - Multispectral image of apples acquired under partial illumination (802 nm)

we divide the objective by 2 while optimizing D , which slows down the rate at which D learns relative to G. Optimization done in two steps: generator loss optimization step and then discriminator loss optimization step. Regularization parameters has following values: $\lambda_{VGG} = \lambda_{Feat} = 10$; $\lambda_{L_1} = 100$.

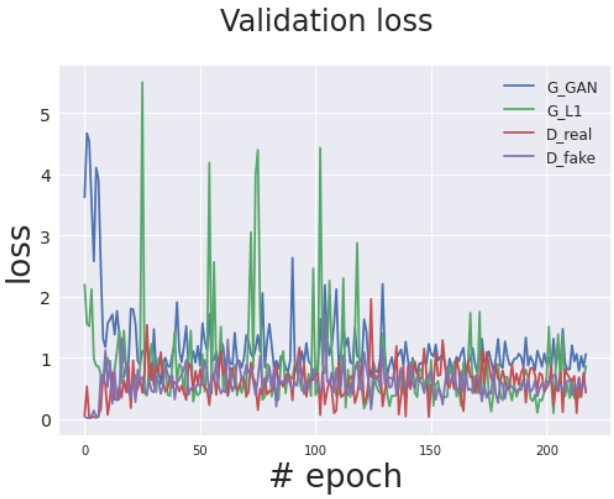
We use Adam optimizer with a learning rate of 0.0002 and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. During training as in original paper we use batchsize equal 1; during inference as in original paper we don't use running statistic for batch-normalization layers. All networks were learned by 200 epoch: first 100 with constant learning rate and rest 100 - with linearly decaying to zero.

Convolutional layers weights initialized with zero-mean normal distribution with std - 0.02, batch-normalization layers scale initialized with normal distribution with mean - 1 and std - 0.02.

6.4. Other details

Experiments were provided on Google Collab cloud service. Due to limited computational resources experiments were not repeated, for same reason we don't use cross-validation. To estimate performance we use train/test random data split (80% for train 20% for test).

6.5. Loss behaviour



Validation loss dynamics during training for Pix2Pix model

In the plot we can see the validation losses of the model against the number of epochs completed in training. We can see that training quite unstable, although both GAN and L_1 losses tend to decrease with time. Same applied to other considered models.

6.6. Results

Resulting quantitive performance of considered models is summarized in Table 6.6. Considering both pixel-based and image metrics we can say that result quite promising.

	mae	mape	mse	psnr	ssim
pointwise	0.008	0.013	0.00011	39.16	0.932
pointwise(avg)	0.006	0.010	0.00009	41.45	0.947
cyclegan	0.067	0.105	0.01127	27.37	0.856
pix2pix	0.004	0.006	0.00003	46.43	0.955
pix2pixHD	0.004	0.006	0.00003	46.85	0.955

Resulted images provided in **Appendix** section. Here we discuss perceptual quality. We can see that reconstructed images looks more or less reasonable (at least on half of cases): images contains apples, overall light instensity similar to groundtrooth and decay region mainly preserved. However all models have particular artefacts. PointWise model have wave-like artefacts around apples boundaries and it captures intensity level poorly - it show worse performance. Pointwise(*) model performs better, but it have artefacts which looks like quantization. CycleGan model have big stamp-like artifacts and there a lot of missed decay regions. Pix2Pix and Pix2PixHD models performs comparable and much better then others, and decay regions preserved relatively well, although intensity level mismatch can be seen.

Actually, to say if considered methods is good or not for decay detection overall quality of full pipeline(translation -¿ segmentation/classification) should be measured. For such scenario labeled data is needed (segmented regions of decay, study of apple decay, etc.) but currently we haven't such data.

Considering that decay regions should be preserved accurately for pipeline setting we can propose two improvements if labels described above available. Loss improvement: we can penalize network for mistakes in decaying regions/apples regions bigger (this will force network to make more accurate prediction in apples regions while background restoration quality degradation hopefully will not affect segmentation part). Architecture improvement: segmentation masks can be utilized to improve performance as attention masks or input to attention layers.

7. Conclusion

Usage of deep learning techniques to translate digital to multispectral images show promising results. Pix2Pix and Pix2PixHD models produce perceptually good images, preserving important for task features and mean error level quite low - 0.6%. Further research should be performed on entire decay detection pipeline construction and incorporat-

ing task specific loss or architecture improvements.

References

Akbari, H., Kosugi, Y., Kojima, K., and Tanaka, N. Wavelet-based compression and segmentation of hyperspectral images in surgery. In *International Workshop on Medical Imaging and Virtual Reality*, pp. 142–149. Springer, 2008.

Akbari, H., Kosugi, Y., Kojima, K., and Tanaka, N. Detection and analysis of the intestinal ischemia using visible and invisible hyperspectral imaging. *IEEE Transactions on Biomedical Engineering*, 57(8):2011–2017, 2010.

Brossard, M., Marion, R., and Carrére, V. Deconvolution of swir reflectance spectra for automatic mineral identification in hyperspectral imaging. *Remote Sensing Letters*, 7 (6):581–590, 2016.

de Lima, D. C., Saqui, D., Ataky, S., Jorge, L. A. d. C., Ferreira, E. J., and Saito, J. H. Estimating agriculture nir images from aerial rgb data. In *International Conference on Computational Science*, pp. 562–574. Springer, 2019.

de Oca, A. M., Arreola, L., Flores, A., Sanchez, J., and Flores, G. Low-cost multispectral imaging system for crop monitoring. In *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 443–451. IEEE, 2018.

Grahn, H. and Geladi, P. *Techniques and applications of hyperspectral image analysis*. John Wiley & Sons, 2007.

Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., et al. Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. *Geoderma*, 398:115118, 2021.

Hohmann, M., Kanawade, R., Klämpfl, F., Douplik, A., Mudter, J., Neurath, M., and Albrecht, H. In-vivo multispectral video endoscopy towards in-vivo hyperspectral video endoscopy. *Journal of biophotonics*, 10(4):553–564, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks, 2018.

Li, J., Luo, W., Wang, Z., and Fan, S. Early detection of decay on apples using hyperspectral reflectance imaging combining both principal component analysis and improved watershed segmentation method. *Postharvest Biology and Technology*, 149:235–246, 2019.

Lorente, D., Aleixos, N., Gómez-Sanchis, J., Cubero, S., García-Navarrete, O. L., and Blasco, J. Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment. *Food and Bioprocess Technology*, 5(4):1121–1142, 2012.

Lu, Y., Saeys, W., Kim, M., Peng, Y., and Lu, R. Hyperspectral imaging technology for quality and safety evaluation of horticultural products: A review and celebration of the past 20-year progress. *Postharvest Biology and Technology*, 170:111318, 2020.

Qin, J., Chao, K., Kim, M. S., Lu, R., and Burks, T. F. Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering*, 118(2):157–171, 2013.

Sa, I., Popović, M., Khanna, R., Chen, Z., Lottes, P., Liebisch, F., Nieto, J., Stachniss, C., Walter, A., and Siegwart, R. Weedmap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sensing*, 10(9):1423, 2018.

Schols, R. M., Alic, L., Beets, G. L., Breukink, S. O., Wieringa, F. P., and Stassen, L. P. Automated spectroscopic tissue classification in colorectal surgery. *Surgical innovation*, 22(6):557–567, 2015.

Tian, X., Fan, S., Huang, W., Wang, Z., and Li, J. Detection of early decay on citrus using hyperspectral transmittance imaging technology coupled with principal component analysis and improved watershed segmentation algorithms. *Postharvest Biology and Technology*, 161: 111071, 2020.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans, 2018.

Wirkert, S. J., Kenngott, H., Mayer, B., Mietkowski, P., Wagner, M., Sauer, P., Clancy, N. T., Elson, D. S., and Maier-Hein, L. Robust near real-time estimation of physiological parameters from megapixel multispectral images with inverse monte carlo and random forest regression. *International journal of computer assisted radiology and surgery*, 11(6):909–917, 2016.

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., and Lausch, A. High-resolution digital mapping of soil organic carbon and soil total nitrogen using dem derivatives, sentinel-1 and sentinel-2 data based on machine learning algorithms. *Science of The Total Environment*, 729:138244, 2020.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

Zuzak, K. J., Naik, S. C., Alexandrakis, G., Hawkins, D., Behbehani, K., and Livingston, E. Intraoperative bile duct visualization using near-infrared hyperspectral video

imaging. *The American Journal of Surgery*, 195(4):491–497, 2008.

A. Team member's contributions

Explicitly stated contributions of each team member to the final project.

Alexander Telepov 70% of work

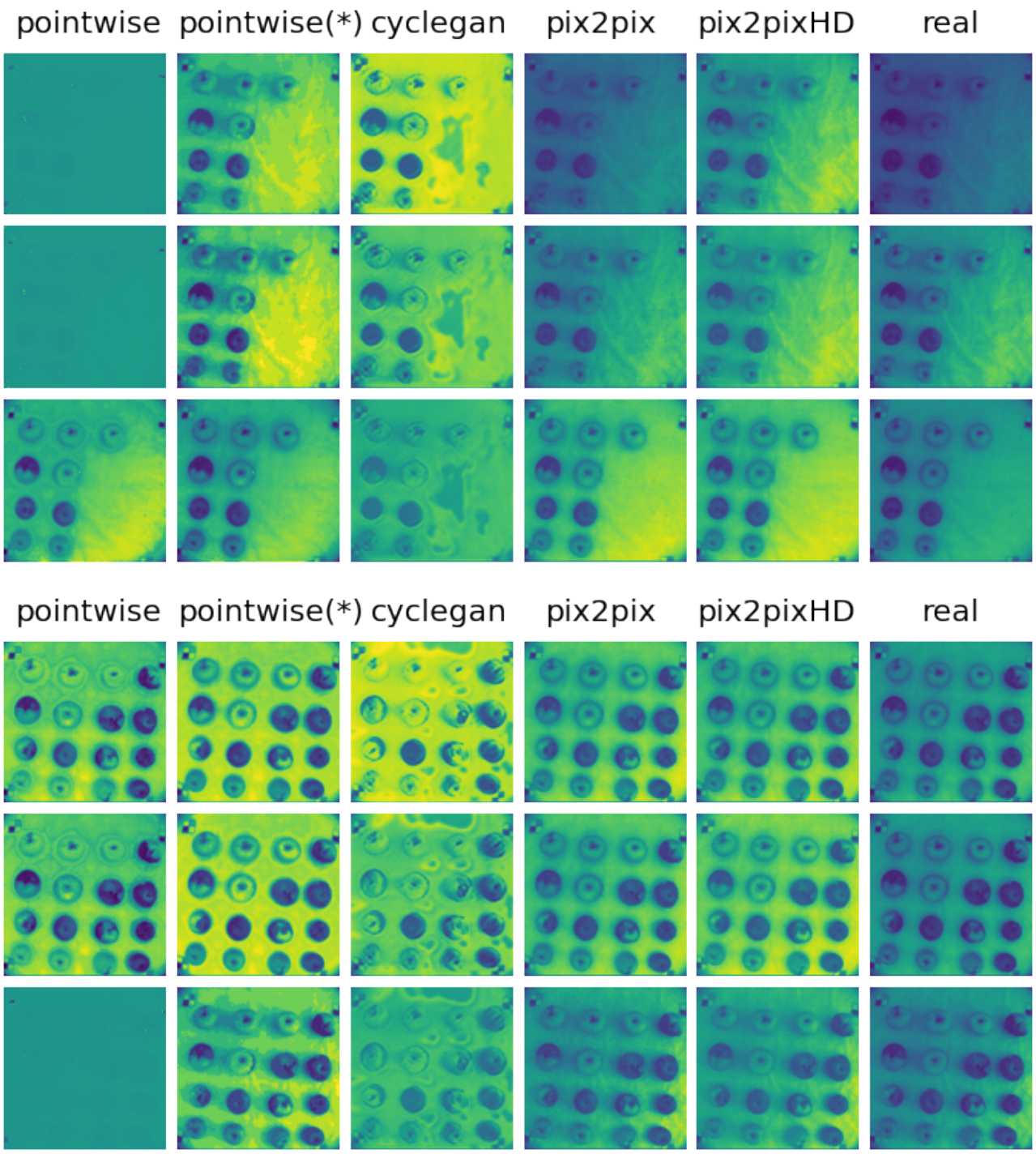
- Rewrite training framework
- Provide numerical experiments
- Preparing the GitHub Repo
- Preparing presentation and writing report

Nikita Stasenko 30% of work

- Making literature review
- Providing experiments for data acquiring
- Data acquiring and processing
- Preparing presentation and writing report

B. External code

Our code based on repository.


pointwise
pointwise(*)
cyclegan
pix2pix
pix2pixHD
real
pointwise
pointwise(*)
cyclegan
pix2pix
pix2pixHD
real