Controlled Multi-modal Image Generation for Plant Growth Modeling

Miro Miranda*‡, Lukas Drees*†, Ribana Roscher*†

* IGG, Remote Sensing, University of Bonn, Niebuhrstr. 1a, 53113 Bonn, Germany
† Data Science in Earth Observation, Technical University of Munich, Lise-Meitner-Str. 9, 85521 Ottobrunn, Germany
‡ German Research Center for Artificial Intelligence (DFKI), Trippstadter Str. 122, 67663 Kaiserslautern, Germany

***Abstract*—Predicting plant development is an important task in precision farming and an essential metric for decision-making by researchers and farmers. In this work, we propose a novel generative modeling technique for plant growth prediction based on conditional generative adversarial networks. We formulate plant growth as an image-to-image translation task and predict the appearance of a plant growth stage as a function of its previous stage. We take into account that plant growth is inherently multi-modal, depending on numerous and highly variable environmental factors, and thus a single input belongs to a distribution of potential outputs. We encode the ambiguity in an interpretable and low-dimensional latent vector space representing the various factors of variation that are influencing plant growth. We use a novel encoder-based data fusion technique and combine information contained in remote sensing imagery of different cropping systems with data containing the factors of variation to adequately model plant growth. This offers several advantages over existing methods: (1) we show that we can model a distribution of potential appearances and simultaneously outperform existing methods in providing more realistic predictions, (2) the complexity of plant growth is more adequately captured, as various factors influencing plant growth can be included, (3) predictions are controllable by being conditioned by an interpretable latent vector representing the factors of variation along with an input image of a previous growth stage.**

I. Introduction

Plant growth is multimodal by nature, depending on diverse environmental factors such as climate conditions, nutrient supply, or pest pressure [1], [2]. Therefore, adequately modeling of plant growth as a function of diverse environmental conditions is important for decision-making in agriculture and allows to adapt to changing or fluctuating environmental conditions, pest and weed control, fertilization, or harvest time prediction. Classical models often describe plant development as a function of environmental factors such as light and temperature by using differential equations [3]. Due to the limited accuracy of these approaches, the literature recently relied more on the use of neural networks (NN). Approaches like the one presented in [4] model plant growth by predicting an image of what a plant will look like in the future conditioned on an image of its previous growing stage. In contrast to classical growth prediction models, generated images of future plant growth offer several advantages. They illustrate the complete above-ground coverage, and various parameters can be derived by statistical or machine learning (ML)-based methods such as semantic segmentation [5], [6]. Nevertheless, generating images of plant growth remains a challenge as available methods are deterministic and thus can only produce a single prediction without providing any variability in the generative process [4], [7], [8]. Since plants perform differently under varying environmental and management conditions [1], [2], their growth should be better described by a distribution of potential appearances. These conditions that influence the process of plant growth are termed as factors of variation (FOV). In contrast, available methods that model a distribution of outputs are stochastic, interfering with a controlled and deterministic output generation [9].

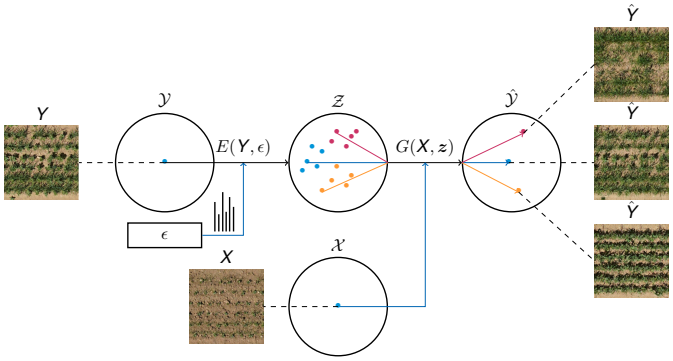

Fig. 1. Temporal plant growth prediction using close-range remote sensing imagery. An encoder-based data fusion of the image domain Y and the domain of factors of variation (FOV) $\boldsymbol{\epsilon}$ is used to span a disentangled latent space Z, $E(Y, \boldsymbol{\epsilon}) \rightarrow \boldsymbol{z}$. Note, that Z follows the distribution of the FOV. Along with a sample of the respective latent space, $\mathcal{Z}$ and an input image X, a generator produces a deterministic plant growth predictions, $G(X, \boldsymbol{z}) \rightarrow Y$. By controlled sampling from $\mathcal{Z}$, one input image is mapped to a distribution of possible outputs (each described by certain FOV).

In this paper, we overcome existing limitations and introduce a framework called `contGAN` for controlled diverse output generation based on conditional generative adversarial networks (cGANs). We formulate plant growth as an image-to-image translation task between two image domains, where each of them represents a different temporal growth stage. In particular, we present the following contributions:

- We explicitly consider the multimodal nature of plant growth by incorporating multiple conditions and provide a bijective mapping between high-dimensional inputs and

high-dimensional outputs. This includes an input image and a low-dimensional latent vector representing the FOV. At inference time, one input image is mapped to a distribution of possible outputs, where each output is determined by a unique set of FOV.
- We present an encoder-based data fusion technique to model a distribution of latent vectors by combining information contained in images with information about FOVs.
- We demonstrate that the low-dimensional latent space is disentangled, resulting in an improved model performance while achieving a more refined control. In addition, we demonstrate higher interpretability and explainability of a prediction due to the multimodal input data. This is in line with the increasing demand for explainable machine learning to assure the value of a prediction [10].

The capability of our method is demonstrated using a combination of remote sensing imagery and FOVs for different cropping systems.

II. Related Work

Conditional generative modeling: Machine learning-based generative modeling is a common approach for synthetic image generation. The GAN framework [11] achieved impressive results in image generation and was extended in various ways. A conditional setting for cross-domain mapping between two image domains was proposed by [7] and extended to high-resolution image generation [12]. However, most methods are deterministic and lack diversity as they can only map to a single target, known as one-to-one mapping [4]. In addition to modeling plant growth, many generative modeling problems are ambiguous. A single input may correspond to a distribution of possible outputs while demanding a deterministic and not a stochastic output generation. Mapping between multimodal inputs and an output distribution is challenging and was addressed by [9]. The multimodality is distilled in a low-dimensional latent vector of noise that is internalized during training and that can be randomly sampled during test time. In contrast to this, an unsupervised multimodal framework for image-to-image translation was proposed by [13]. Here, the underlying idea is to decompose the internal latent representation into a domain-invariant content code and a style code. The content code can be recombined with randomly sampled style codes to generate diverse outputs. However, instead of modeling a distribution, the output is restricted to the set of combination between content code and style code. The available methods that provide diverse outputs have two main limitations. They are either stochastic, resulting in an uncontrolled output generation, or do not model a probability distribution, interfering with spatial interpolation. We are not aware of a method for a deterministic mapping between high-dimensional inputs and high-dimensional outputs by considering diverse conditions.

Plant growth modeling: There has been substantial interest on plant growth modeling in the last decades and many methods have been derived since then including linear- and polynomial regression models [14]–[17]. Recently, the literature addressed plant growth by using NN as they capture patterns and correlations even in noisy and high-dimensional data and represent good approximators for any complex and non-linear function [18].

With the rise of remote sensing technology, images are now available at low-cost and large scale and can be further used for plant growth modeling, and yield prediction [19]. While, as a consequence, many approaches have emerged that have derived specific plant-related parameters from images, there is only little research on predicting images for plant growth modeling. Nonetheless, it has been recently shown, that GANs are an appropriate method to model growth or semantic segmentations of plants [4], [20], [21]. However, most existing methods are restricted by predicting a single or a limited diverse output instead of a probability distribution. A methodology for plant growth prediction by using conditional generative adversarial networks (cGANs) on both field and laboratory image data was proposed by [4]. Here, future growth stages are conditioned on earlier ones. However, the presented framework is realized as a one-to-one mapping ignoring the multimodality underlying plant development. To our knowledge, there are no approaches using cGANs for controllable and multimodal plant image generation with multiple data sources, neither for plant growth prediction - especially in complex cropping systems - nor for general purposes.

III. Image-to-Image Translation

A. Conditional GAN

Conditional generative adversarial networks (cGANs) [11] are an established method for image-to-image translation. GANs simultaneously train two neural networks: a generative and a discriminative model. While the generative model G tries to generate data indistinguishable from reality, the discriminative model D tries to distinguish between real and generated data. Contrary to GANs that learn a mapping from a random noise vector to an output image, cGANs instead learn a mapping from an input image of domain $\mathcal{X}$ and random noise $\boldsymbol{z}$ to an output image of domain $\mathcal{Y}$, $G : \{\boldsymbol{X}, \boldsymbol{z}\} \rightarrow \boldsymbol{Y}$. An outstanding approach was proposed by the Pix2Pix architecture [7] with the formalism

$$\begin{aligned}\mathcal{L}_{\mathrm{cGAN}}(G, D) = \mathbb{E}_{\mathcal{X},\mathcal{Y}}[\log D(\boldsymbol{X}, \boldsymbol{Y})] \\ + \mathbb{E}_{\mathcal{X},\mathcal{Z}}[\log(1 - D(\boldsymbol{X}, G(\boldsymbol{X}, \boldsymbol{z})))],\end{aligned} \tag{1}$$

which G tries to minimize, while D tries to maximize. Further, an $L1$ loss is included to force the generator not only to fool the discriminator but also to produce samples which are pixel-wise close to the target image.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{\mathcal{X},\mathcal{Y},\mathcal{Z}}[||\boldsymbol{Y} - G(\boldsymbol{X}, \boldsymbol{z})||_1]. \tag{2}$$

This leads to the final objective described by:

$$G^*, D^* = \arg\min_G \max_D \ \mathcal{L}_{\mathrm{cGAN}}(G, D) + \lambda_I \mathcal{L}_{L1}(G) \tag{3}$$

where λ_I is a hyperparameter used to control the weighting of the $L1$ loss. As a drawback, introducing stochasticity by

random noise was found to be ignored by the generator [7], [9], limiting the approach to a one-to-one mapping.

B. Multi-Conditional GANs

We define the objective as learning the information contained in both image domains as well as additional FOV. Precisely, we learn a multimodal mapping function between two image domains $\mathcal{X}$ and $\mathcal{Y}$, given paired training data $\{(\boldsymbol{X} \in \mathcal{X}, \boldsymbol{z} \in \mathcal{Z}), \boldsymbol{Y} \in Y\}$ with $\boldsymbol{z}$ following a joint probability distribution $p(\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{\epsilon})$. Here, $\boldsymbol{z}$ is not just random noise, but a low-dimensional description of the reference image and thus a second condition along with the input image. In addition, $\boldsymbol{\epsilon}$ are the additional FOV. Similarly, $p(\epsilon)$ refers to the distribution of the additional FOV. If not specified, the term test time distribution refers to a sample from the space of FOV. The multimodal mapping, therefore, become deterministic by $G(\boldsymbol{X}, \boldsymbol{z}) \rightarrow \boldsymbol{Y}$, resulting in ambiguous outputs when sampling from $\mathcal{Z}$ during test time. Our framework is built on the BicycleGAN model proposed by [9].

We encode useful information into a low-dimensional latent space $\mathcal{Z}$ by using a multi-layer autoencoder network (AE) [22] formulated in a probabilistic fashion. This has been proposed by the variational autoencoder framework [23]. In a conditional setting, the distribution $Q(z|\boldsymbol{Y}, \boldsymbol{\epsilon})$ of the latent vector $\boldsymbol{z}$ is given by using an encoder network E, $Q(z|\boldsymbol{Y}, \boldsymbol{\epsilon}) := E(\boldsymbol{Y}, \boldsymbol{\epsilon})$. Sampling from $\mathcal{Z}$ is enabled by a reparametrization trick, allowing direct back-propagation [23]:

$$z \sim E(\boldsymbol{Y}, \boldsymbol{\epsilon}) = \mu + \sigma\boldsymbol{\epsilon}, \tag{4}$$

with $\boldsymbol{\epsilon} \sim p(\epsilon)$ given as the FOV uniquely describing $\boldsymbol{Y}$. We explicitly aim to reproduce the distribution of the FOV, using a Kullback-Leibler-divergence (KL) loss. This allows an explainable and interpretable sampling during test time:

$$\mathcal{L}_{\mathrm{KL}}(E) = \mathbb{E}_{\mathcal{Y}}\left[\mathcal{D}_{\mathrm{KL}}(E(\boldsymbol{Y}, \boldsymbol{\epsilon}) \parallel p(\epsilon))\right] \tag{5}$$

Along with this low-dimensional latent representation of the reference image and the input image, a cGAN should learn to reconstruct the reference image. The formalism of the conditional variational autoencoder GAN (VAE-GAN) is given by:

$$\begin{aligned} G^*, D^*, E^* = \arg\min_{G,E}\max_{D} \mathcal{L}^{\mathrm{VAE}}_{\mathrm{cGAN}} \\ + \lambda_I \mathcal{L}^{\mathrm{VAE}}_{L1}(G) + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}(E). \end{aligned} \tag{6}$$

Although the encoder-based data fusion is a way to make the latent space influential, it may not be considerably close to the test time distribution because the KL-loss may be poorly optimized. As a solution, a conditional Latent Regressor GAN (cLR-GAN) was proposed by [9]. The cLR-GAN forces the latent distribution to follow the test time distribution, which can be randomly sampled. This is realized by using an L1 loss in the latent space:

$$\mathcal{L}^{\text{latent}}_{L1}(G) = \mathbb{E}_{\mathcal{X},\mathcal{Z}}[||\boldsymbol{z} - E(G(\boldsymbol{X}, \boldsymbol{z}))||_1] \tag{7}$$

Together with a randomly sampled latent vector, a generator produces a prediction $\hat{\boldsymbol{Y}}$ which should be realistically close to the target domain. The predicted image is then encoded back into a low-dimensional representation. We extend the cLR-GAN by additionally adding an $L1$ loss term between the prediction $\hat{\boldsymbol{Y}}$ and the ground truth, resulting in more realistic results. The full loss term is given by:

$$\begin{aligned} G^*, D^* = \arg\min_{G,E}\max_{D}\ \mathcal{L}^{\mathrm{cLR}}_{\mathrm{GAN}}(G, D) \\ + \lambda_{\text{latent}}\mathcal{L}^{\text{latent}}_{1}(G, E) + \lambda_I\mathcal{L}^{\mathrm{cLR}}_{L1}(G) \end{aligned} \tag{8}$$

Then the final objective of the proposed contGAN is given by:

$$\begin{aligned} G^*, E^*, D^* = \arg\min_{G,E}\max_{D} \mathcal{L}^{\mathrm{VAE}}_{\mathrm{GAN}}(G, D, E) + \lambda_I\mathcal{L}^{\mathrm{VAE}}_{1}(G, E) \\ + \mathcal{L}^{\mathrm{cLR}}_{\mathrm{GAN}}(G, D) + \lambda_I\mathcal{L}^{\mathrm{cLR}}_{1}(G, E) \\ + \lambda_{\text{latent}}\,\mathcal{L}^{\text{latent}}_{1}\,(G, E) + \lambda_{\mathrm{KL}}, \mathcal{L}_{\mathrm{KL}}(E) \end{aligned} \tag{9}$$

where the parameters λ are used to control the weighting of all terms.

A U-Net architecture with skip connections is used as the generator network [24]. Two discriminators are used for each cLR-GAN and VAE-GAN, defined as a PatchGAN [25] implementation with an overlapping patch size of $[70 \times 70]$. For the encoder network, we use a ResNet architecture with several residual blocks [26]. We use a Least Squares GAN (LSGAN) [27] that uses the least squares loss function for the discriminator instead of a binary cross-entropy [28].

C. Model Evaluation

Images are qualitatively evaluated using human perception by a three-point guideline:

- Realistic appearance: Generated images should look sharp, artifact-free, and have visible morphological details.
- Consistency of conditions: The images should be consistent with both input conditions, the image and the chosen factor of variation.
- Generalization: The model should generalize on unseen data and produce diverse but realistic results without mode collapse.

For quantitative evaluation an unbiased estimator of the Fréchet Inception Distance (FID_∞) is used, termed FID infinity score [29]. While the classic FID [30] is a robust and efficient score that is considered to correlate with human judgment of image diversity, and quality [31], it also has some disadvantages. For a finite set of samples, FID deviates from the true score, is characterized by a high bias [32], and depends on the model being evaluated [29]. Therefore, a comparison between different models is unreliable. The FID_∞, is a bias-free estimation of the FID score computed with an infinite number of samples.

We use the Ground Cover Fraction (GCF) to estimate the fractional green canopy between predictions and references. The GCF is a good estimator of canopy development and correlates with the above ground biomass [33]. Moreover, with the GCF, image semantics are evaluated and accessed to determine whether generated images are suitable for inferring phenotypic

TABLE I
FACTORS OF VARIATION (FOV) BELONGING TO AN IMAGE FROM EACH OF THE THREE CLASSES MIXTURE (MX), SPRING WHEAT (SW), AND FABA BEAN (FB).

class	Total Yield (g)	Harv. Weight (g)	Seed Numbers SW	Seed Numbers FB	Dried Biomass (g) SW	Dried Biomass (g) FB	Mean Height (cm) SW	Mean Height (cm) FB
SW	0.61	0.29	1.14	0.12	0.20	-1.07	-1.46	-1.40
MX	-0.13	1.03	-0.06	-0.07	0.09	0.56	-0.13	0.91
FB	-5.02	-0.23	-1.91	-2.66	-2.83	2.28	0.81	1.17

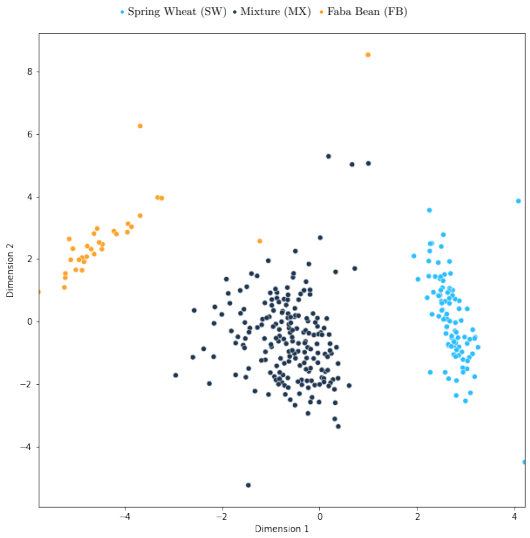

Fig. 2. FOV of all images embedded in a two-dimensional feature space using Multidimensional Scaling (MDS). Three clusters arises according to the underlying classes of MX, SW, and FB.

traits from predictions. We use an automatic color threshold classification (ACT) method based on [33] to estimate the GCF by color ratios in RGB space. For fairness, we compare the proposed model against a classic cGAN (Pix2Pix: [7]) that has been proven to be a good estimator of plant growth [4], and with a multimodal cGAN (BicycleGAN: [9]).

IV. DATA

We use multimodal data acquired by a mixed-cropping experiment at the site Campus Klein-Altendorf in 2020, Germany. Here, mixtures of varieties spring wheat (*Triticum aestivum*) and faba bean (*Vicia faba L.*), as well as monoculture reference plots, were planted with different FOV on 320 plots, each with a size of $10\,\mathrm{m} \times 1.5\,\mathrm{m}$.

RGB images: On a weekly basis during the growing season, RGB images were taken, which are processed into georeferenced orthophotos with a ground resolution of $3\,\mathrm{mm}$. This allows the extraction of spatially aligned image pairs from an early (4 weeks after plant emergence, domain X) and a mid-growth stage (8 weeks, domain Y) that serve as input for our model. We select a balanced data set equally containing images of all three classes, namely mixture MX, spring wheat SW and faba bean FB monoculture. In total, a training set of 750 images and a spatially separated test set of 307 images were selected.

Factors of variation (FOV): Each image is associated with a feature vector $\boldsymbol{\epsilon}$, containing FOV. Those describe dozens of environmental and management conditions as well as plant characteristics. A subset is shown in Tab. I. Fig. 2 visualizes a low-dimensional embedding of this feature space by using a Multidimensional Scaling (MDS) method for dimensionality reduction [34]. The figure illustrates three distinct clusters that are in concordance with the crop varieties.

V. EXPERIMENTS

We use remote sensing imagery, and additional FOV to explicitly analyze the ability to (1) generate sharp and realistic temporal predictions, (2) investigate if we can model the spatial distribution of potential outputs, and (3) analyze if predictions are controllable and explainable.

A. Experimental setup

For efficient processing, all images are patched to a size of $[256 \times 256]$. A set of FOV is selected by using a random-forest feature importance estimation [35], which helps to filter highly correlated factors and thus to create a class-separated but dense feature space. The selected FOV are associated with yield, one of the most essential traits while being easy to evaluate.

1) Latent code injection: We test different ways of injecting the latent code into the generator. Injecting the latent code into all layers by spatial replication and concatenation was found to produce the most realistic results. Moreover, we analyze varying shapes and their influence on the predictions. A high-dimensional latent code may encode more information resulting in more diverse results but at the cost of realism. In contrast, a low-dimensional latent vector may result in less diverse outputs [9]. We find that the optimal shape of the latent code is $z = 8$ as it yields diverse but still realistic results.

2) Hyperparameter: The hyperparameters λ are set to $\lambda_{\mathrm{I}} = 0.5$, $\lambda_{\mathrm{latent}} = 10$, $\lambda_{\mathrm{KL}} = 0.01$. We train on 400 epochs with a learning rate of $2\mathrm{e}{-4}$.

B. Latent Space Dissection

This section demonstrates the effectiveness of including additional FOV for a latent space disentanglement. We show the latent space of the encoder-based data fusion in Fig. 3. Here, we use a t-Distributed Stochastic Neighbor Embedding (t-SNE) [36] with a perplexity of 50 for dimensionality reduction, as we want to preserve local structures to better analyze the latent space distribution. The right image displays how the latent code is disentangled in an interpretable and explainable way, forming three distinct classes in concordance to the used crop varieties. In contrast, when only adding random Gaussian noise, the latent space is randomly distributed. This is shown in the left image. It cannot be interpreted in such a way that it can be used for a controlled output generation and neither for comprehensive plant growth prediction.

C. Temporal plant growth prediction

In Fig. 4 we compare the performance of our proposed framework contGAN achieving controlled mapping with a

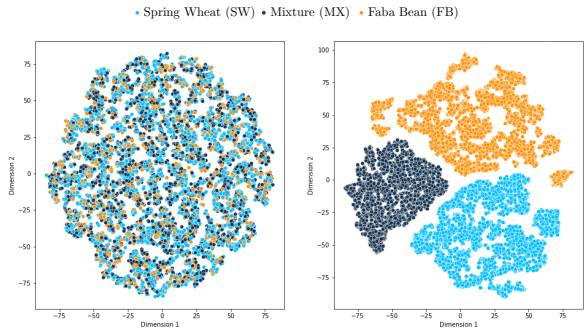

Fig. 3. t-Distributed Stochastic Neighborhood (t-SNE) embedding of the latent vectors $z = E(Y, \epsilon)$ sampled during training. On the left: ϵ is a random sample from a Gaussian normal distribution $\epsilon \sim \mathcal{N}(0, 1)$. On the right: in contrast, ϵ is the vector of FOVs uniquely describing Y.

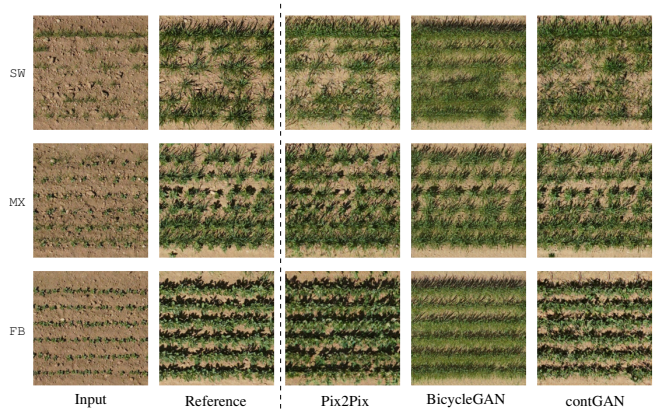

Fig. 4. Example results for 1) Pix2Pix, 2) BicycleGAN, and 3) the proposed contGAN. The most left column shows the generator input and the reference image being next to it. Example results are given for all three classes namely 1) `MX`, 2) `SW`, and 3) `FB`.

classic cGAN (Pix2Pix: [7]) and with a multimodal cGAN (BicycleGAN: [9]). We decide for this comparison to evaluate if the proposed framework can combine the advantages of both models.

While in Pix2Pix there is no additional input z to the image, and in the BicycleGAN z is sampled as a random realization of a Gaussian normal distribution, in our controlled method, z represents the FOV describing the displayed reference image Y. Pix2Pix achieves impressive results for plant growth prediction in complex cropping systems. Structures are well preserved and consistent with the input image. However, the prediction for the `FB` (last row) seems slightly too dense compared to the reference images. In contrast, the uncontrolled mapping with BicycleGAN yields more diverse results. We note that predictions deviate from the reference image but are still realistic. However, the prediction for `FB` is not faithful to the input as we observe inter-class transformations from `FB` to `SW` without explicitly aiming for this. The last column displays the performance of our controlled framework. The displayed results are realistic and faithful to the input image. Structures are well preserved, plant-related details like single leaves are visible, and soil holes in the canopy are close to the target image. In contrast to both other methods, the prediction for the class `FB` looks more realistic and closer to the reference.

TABLE III
OVERVIEW ABOUT CALCULATED FID INFINITY SCORES ($\mathcal{N}_r$,$\mathcal{N}_g$) FOR DIFFERENT EXPERIMENTS.

$\text{FID}_{\text{Pix2Pix}}$	$\text{FID}_{\text{BcyGAN}}$	$\text{FID}_{\text{contGAN}}$ FID_{spec}	$\text{FID}_{\text{class}}$	FID_{rand}
17.9473	25.80	**16.32**	19.38	21.02

TABLE IV
L1 LOSS FOR GFC GENERATED IMAGES.

L1 Loss	Pix2Pix	BicyleGAN	contGAN
`SW`	0.27	0.42	**0.18**
`MX`	0.25	0.40	**0.23**
`FB`	0.27	0.39	**0.21**

D. Influence of FOV on a prediction

In Fig. 5, controllable and interpretable output generation is demonstrated, conditioned on both an input image (left) and a learned latent vector that is pointedly sampled from arbitrary FOV. The chosen FOV are illustrated as a bar plot above each image, and respective features are colored according to each class and according to the features associated with the expression of the biomass and yield. We demonstrate that we can control the expression of classes as well as the expression of biomass, yield, and height by simply changing the FOV accordingly. We show that the predictions are faithful to both the input image as well as the FOV. The predictions are impressively realistic as we see morphological details such as leaves. Moreover, we notice that the results are explainable and interpretable according to the given FOVs.

E. Quantitative Evaluation

Besides the qualitative evaluation, we calculate the FID infinity score $\text{FID}_\infty(\mathcal{N}_r, \mathcal{N}_g)$ between real and generated images. In our approach, we distinguish between three different types of generated images, generated from always the same input images, but combined with - at test time - different FOV.

- FID_{spec} $(\mathcal{N}_r, \mathcal{N}_g)$: FOV which belong to the specific reference image.
- $\text{FID}_{\text{class}}$ $(\mathcal{N}_r, \mathcal{N}_g)$: FOV which belong to the same class.
- FID_{rand} $(\mathcal{N}_r, \mathcal{N}_g)$: FOV are randomly sampled.

We moreover compare the results against the Pix2Pix model ($\text{FID}_{\text{Pix2Pix}}$) and the BicycleGAN ($\text{FID}_{\text{BcyGAN}}$). In Tab. III, the calculated FID infinity scores are illustrated. We see that the $\text{FID}_{\text{Pix2Pix}}$ and the FID_{spec} achieve comparably low FID scores underlining the high quality of the generated predictions. Whereas, the $\text{FID}_{\text{BcyGAN}}$ and the remaining $\text{FID}_{\text{class}}$ and FID_{rand} achieve higher FID scores. We conclude that a higher FID scores reflects a higher divergence to the reference image and thus a higher diversity in the predictions. We demonstrate that the contGAN outperforms the Pix2Pix

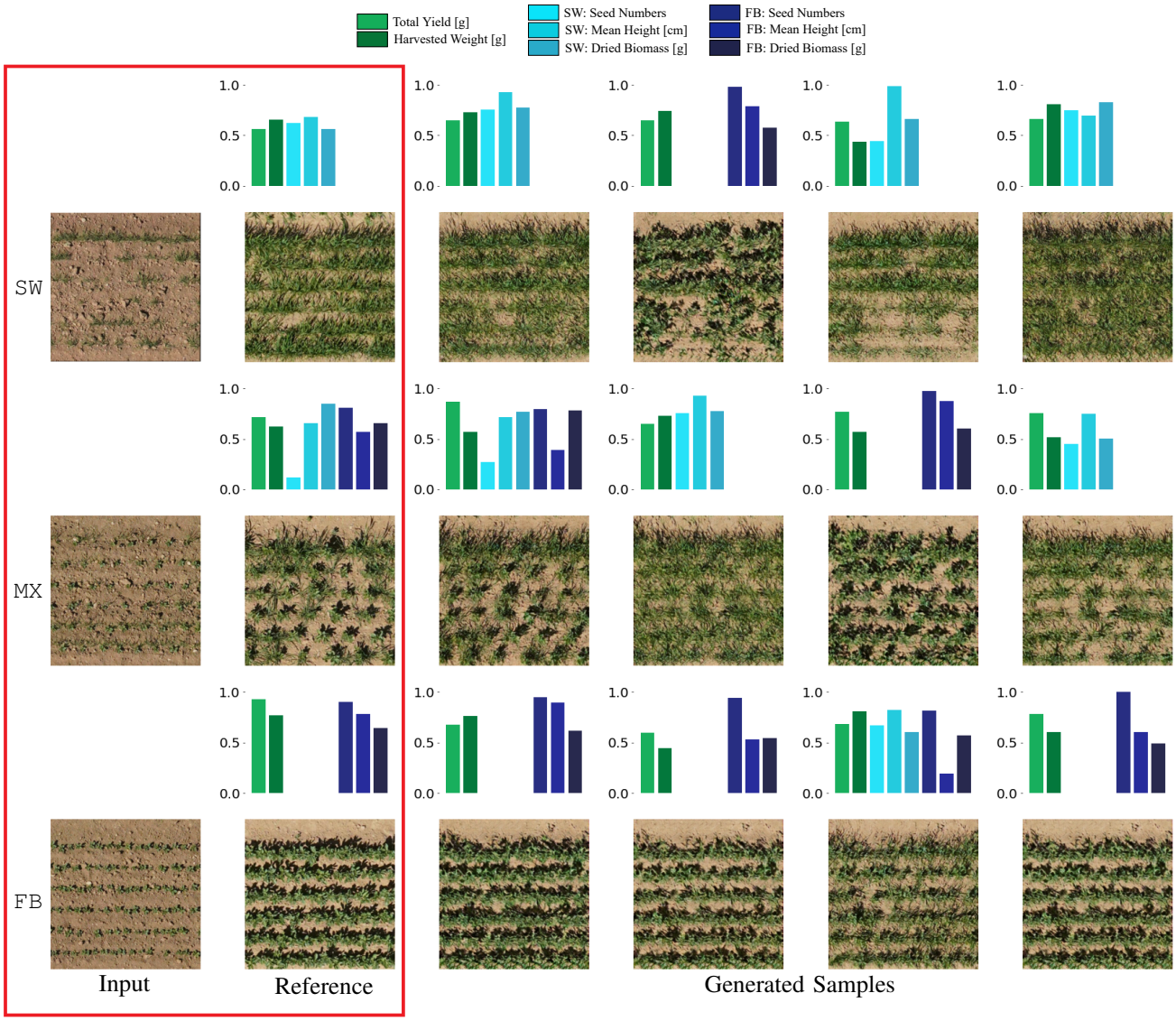

Fig. 5. Example results of a controlled multimodal mapping. The influence of deviating FOV on the prediction is shown. The first column, demonstrated as a red box, shows the input image and the reference images for 1) SW, 2) MX, and 3) FB. On top the reference image we show in a bar chart the normalized FOV that belong exactly to the reference image. Next, we display generated samples using the proposed framework. On top each image the normalized feature vectors of FOV is displayed that was given during test time. Note, that the given feature vectors are interpretable and explainable. We see that the generated samples are effectively controlled by the chosen FOVs.

model, while achieving high diversity depending on the given FOV. We calculate the GCF for both, the reference and the generated images, and compare our proposed framework with the Pix2Pix implementation and a BicycleGAN. In addition to the input image, we give the FOV describing the reference image as input in the case of our model. In Tab. IV, L1 losses between GFC converted reference images, and GFC converted predicted images are shown. The table highlights that both the cGAN and our framework better estimate the fractional green canopy compared to the BicycleGAN implementation. We show that our proposed framework outperforms both models in estimating the green canopy development for all varieties.

VI. Conclusion

Plant growth modeling remains a challenge, especially in complex cropping systems. Nonetheless, the proposed framework for an integrative multi-conditional GAN appears promising. We show that we can outperform existing growth prediction models based on image generation while being able to model a complete distribution of appearances. With the vast amount of available data and image processing tools for monitoring plant ecosystems, the generation of synthetic images appears reasonable for capturing the non-linear nature of plant growth. In addition, it seems necessary that models be more explainable, supporting their acceptance among researchers and farmers.

ACKNOWLEDGMENT

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 – 390732324 and partly funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

REFERENCES

[1] J. S. Boyer, "Plant productivity and environment," *Science*, vol. 218, no. 4571, pp. 443–448, 1982.

[2] W. Gang, W. Zhen-Kuan, W. Yong-Xiang, C. Li-Ye, and S. Hong-Bo, "The mutual responses of higher plants to environment: physiological and microbiological aspects," *Colloids and Surfaces B: Biointerfaces*, vol. 59, no. 2, pp. 113–119, 2007.

[3] R. U. Larsen, "Plant growth modelling by light and temperature," in *Symposium on Bedding and Pot Plant Culture 272*, 1989, pp. 235–242.

[4] L. Drees, L. V. Junker-Frohn, J. Kierdorf, and R. Roscher, "Temporal prediction and evaluation of brassica growth in the field using conditional generative adversarial networks," *Computers and Electronics in Agriculture*, vol. 190, 2021.

[5] L. Zabawa, A. Kicherer, L. Klingbeil, R. Töpfer, H. Kuhlmann, and R. Roscher, "Counting of grapevine berries in images via semantic segmentation using convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 164, pp. 73–83, 2020.

[6] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and electronics in agriculture*, vol. 147, pp. 70–90, 2018.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Multimodal image-to-image translation by enforcing bi-cycle consistency," in *Advances in neural information processing systems*, 2017, pp. 465–476.

[10] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42 200–42 216, 2020.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[12] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[13] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[14] C. Spitters, "An alternative approach to the analysis of mixed cropping experiments. 1. estimation of competition effects." *Netherlands Journal of Agricultural Science*, vol. 31, no. 1, pp. 1–11, 1983.

[15] C. T. Paine, T. R. Marthews, D. R. Vogt, D. Purves, M. Rees, A. Hector, and L. A. Turnbull, "How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists," *Methods in Ecology and Evolution*, vol. 3, no. 2, pp. 245–256, 2012.

[16] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers in plant science*, vol. 10, p. 621, 2019.

[17] R. Rokhana, W. Herulambang, and R. Indraswari, "Machine learning and polynomial – l system algorithm for modeling and simulation of glycine max (l) merrill growth," in *2020 International Electronics Symposium (IES)*, 2020, pp. 463–467.

[18] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020.

[19] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and electronics in agriculture*, vol. 163, p. 104859, 2019.

[20] J. Kierdorf, I. Weber, A. Kicherer, L. Zabawa, L. Drees, and R. Roscher, "Behind the leaves–estimation of occluded grapevine berries with conditional generative adversarial networks," *arXiv preprint arXiv:2105.10325*, 2021.

[21] R. Yasrab, J. Zhang, P. Smyth, and M. P. Pound, "Predicting plant growth from time-series data using deep learning," *Remote Sensing*, vol. 13, no. 3, p. 331, 2021.

[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[24] 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.

[25] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European conference on computer vision*. Springer, 2016, pp. 702–716.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[27] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[28] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.

[29] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[31] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Weinberger, "An empirical study on evaluation metrics of generative adversarial networks," *arXiv preprint arXiv:1806.07755*, 2018.

[32] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[33] A. Patrignani and T. E. Ochsner, "Canopeo: A powerful new tool for measuring fractional green canopy cover," *Agronomy Journal*, vol. 107, no. 6, pp. 2312–2320, 2015.

[34] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[35] J. Rogers and S. Gunn, "Identifying feature relevance using a random forest," in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. Springer, 2005, pp. 173–184.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.