IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 1, 2021, accepted December 27, 2021, date of publication January 5, 2022, date of current version January 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3140464

Shape Based Deep Estimation of Future Plant Images

JOO-YEON JUNG[1], SANG-HO LEE[1], TAE-HYEON KIM[1], MYUNG-MIN OH[2], AND JONG-OK KIM[1], (Member, IEEE)

[1]School of Electrical Engineering, Korea University, Seoul 02841, South Korea
[2]Department of Horticultural Science, Chungbuk National University, Cheongju 361-763, South Korea

Corresponding author: Jong-Ok Kim (jokim@korea.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT), Korean Government, under Grant 2020R1A4A4079705, and in part by the MSIT, South Korea, through the Information Technology Research Center (ITRC) Support Program supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2021-2018-0-01421.

ABSTRACT Plants exhibit dynamic changes as they grow. For example, a new leaf may appear suddenly, and rotate and fold over time. Therefore, it is difficult to predict the growth of plants. For accurate growth predictions, it is important to predict the shapes, colors and textures of the leaves. The conventional methods simply use RGB images to predict the next plant at once. In this paper, we propose a novel deep network which is divided into two subnets of shape estimation and color reconstruction. Four gray time-series images are first aligned to a future target using a spatial transformer network (STN) for shape estimation. They are then fused using U-Net with two LSTMs to generate a future shape image. The color reconstruction subnet fuses the predicted shape with a RGB plant image to restore the color information. In addition, we use gray images with texture information for shape estimation instead of binary images with only simple information and RGB images with too much information. The proposed deep network can robustly generate future plant images for plant growth prediction. It is evaluated using our proprietary dataset as well as two public datasets for different types of plants. The experimental results demonstrate that our proposed network predicts the leaf shape more accurately and restores RGB. As a result, our method can create accurate future plant images.

INDEX TERMS Plant growth prediction, time-series images, shape estimation, spatial transformer network, image generation.

I. INTRODUCTION

It takes much effort to grow plants. For instance, plants need to be constantly supplied with the appropriate amounts of water, humidity, sunlight, temperature, and nutrients, and those environment factors ensure their high-quality growth [1], [24]–[27], [36], [37]. To improve productivity in plant cultivation, we can control the environment factors adaptively, depending on plant growth status. Sensors and monitoring systems have been used in a number of agricultural technologies to increase plant production [2], [28]–[31], [38]. However, they require expensive equipment to measure plant growth, and the measurement also takes much time and is destructive. If plant growth could be predicted with a cheap and remote sensing device such as a vision sensor, that would contribute to cost reduction and automatic environment control without human intervention.

There are very few researches on plant growth prediction. Existing studies focused on literary analysis, especially plant leaf prediction among plant growth, and used existing backbone networks without proposing a new network [3]–[5], [15], [32]. These studies also pass through the network at once and predicts the next plant image. Also, they use only RGB images when predicting plant growth. This paper has a research gap from previous studies. We propose a new network structure which is suitable for plant growth prediction, uses additional gray images without using only RGB images, and is divided into two parts in detail without predicting the plant growth at once to further improve growth prediction accuracy.

In this paper, we propose a novel deep network that generates a future plant image from past and current ones for

The associate editor coordinating the review of this manuscript and approving it for publication was Pinjia Zhang.

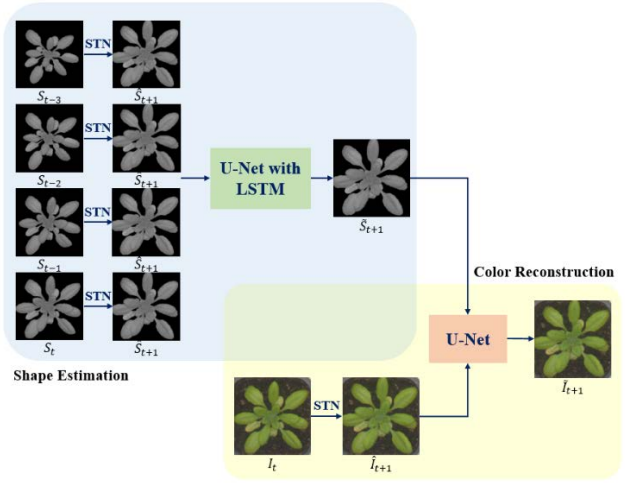

FIGURE 1. The basic concept of the proposed plant growth estimation network architecture.

TABLE 1. Notation used in the paper.

Notation	Definition
S_t	Gray shape image at time t
I_t	RGB image at time t
$\hat{S}_{t+1}$	Gray shape image aligned by STN
$\hat{I}_{t+1}$	RGB image aligned by STN
$\tilde{S}_{t+1}$	Output of the shape estimation subnet
$\tilde{I}_{t+1}$	Output of the color reconstruction subnet (also final output of the network)
$S_{t-3}{\sim}S_t, I_t$	Input of the proposed network

plant growth prediction. To predict the growth accurately, the proposed method consists of two subnets. Fig. 1 illustrates the basic concept of the proposed deep network architecture. First, the overall shape of the plant is predicted from a number of input gray images without a background. Unlike conventional methods that work in the RGB domain, we predict the shape of the plant in gray domain. This reduces the interference of the texture and background in the plant image during the shape prediction [3]–[5]. Each individual input gray image is first aligned to a future target with a spatial transformer network (STN). The resulting multiple aligned images enter the network, and are fused in the shape subnet (U-Net with two LSTMs) to generate a future shape image. Second, the predicted shape image is grayscale, it does not have sufficient texture information. For its colorization and texture replenishment, it passes through the color subnet (auto-encoder net) where it is fused with an RGB image which is simply predicted by a STN from a current RGB image. Finally, the future RGB plant image is generated by the color subnet.

The main contributions of the paper are summarized as follows. First, the growth prediction task is decomposed into the two processes, shape and RGB prediction. The separation of these processes allows the accuracy of the shape prediction to be improved by minimizing the interference of the RGB information in affine transform with STN. In addition, prediction performance was improved using gray images by excluding color information unnecessary for shape prediction. Second, all the shape inputs are geometrically transformed into the future target shape before they enter the network. This leads to more effective shape prediction compared to direct fusion without the STN. Finally, the proposed method is evaluated with our proprietary dataset as well as two public datasets for different types of plants. We built a new dataset which was acquired from plant factory actually.

II. RELATED WORKS

A. STN BASED VIDEO PREDICTION

Although CNN has been widely used, there are still some limitations due to spatial invariance. One of the limitations is that CNN cannot detect objects well when they change in size, rotate, or shift in space. In contrast, spatial transformer network (STN) can manipulate data spatially in the network [6]–[9]. It has been proposed to learn invariance to spatial changes such as size changes, rotations, and position translations. Even today, it is still widely applied to motion estimation and future frame prediction, and is also used in modified forms depending on the purpose [33]–[35], [39], [40]. In [10], a network with dual adversarial training is proposed. The network is divided into the frame and flow branches. The frame branch directly predicts future frames, and the other flow predicts the future flows. The outputs of the two branches are fused to predict the final future frame. In [11], the authors proposed a network that predicts the transformation parameters rather than directly predicting the future frames on the transformation domain. The affine transformation parameters for the final frame prediction are learned via a CNN by fusing a series of successive affine transformation parameters from the input video. That is, the network generates the affine transformation parameters for future frames.

Even in video prediction studies without STN, there have been proposed several methods based mainly on LSTM. In [12], the authors proposed a network that predicts future frames by separating video information into motion and content. The network is trained in an end-to-end manner (rather than separately), and the motion and content parts are combined after passing through different encoders. A network that performs the three processes of pose estimation, pose prediction, and image generation was proposed [13]. After estimating the pose of the time-series input images, the future frames are predicted in the pose domain. It is then reconstructed through image generation subnet from the pose input.

B. PLANT IMAGE PREDICTION

In general, a clean plant image can be obtained only when post-processing such as noise removal is accompanied [20]. However, in as stable environment such as our plant factory, noise or distortion hardly occurs in the acquired image.

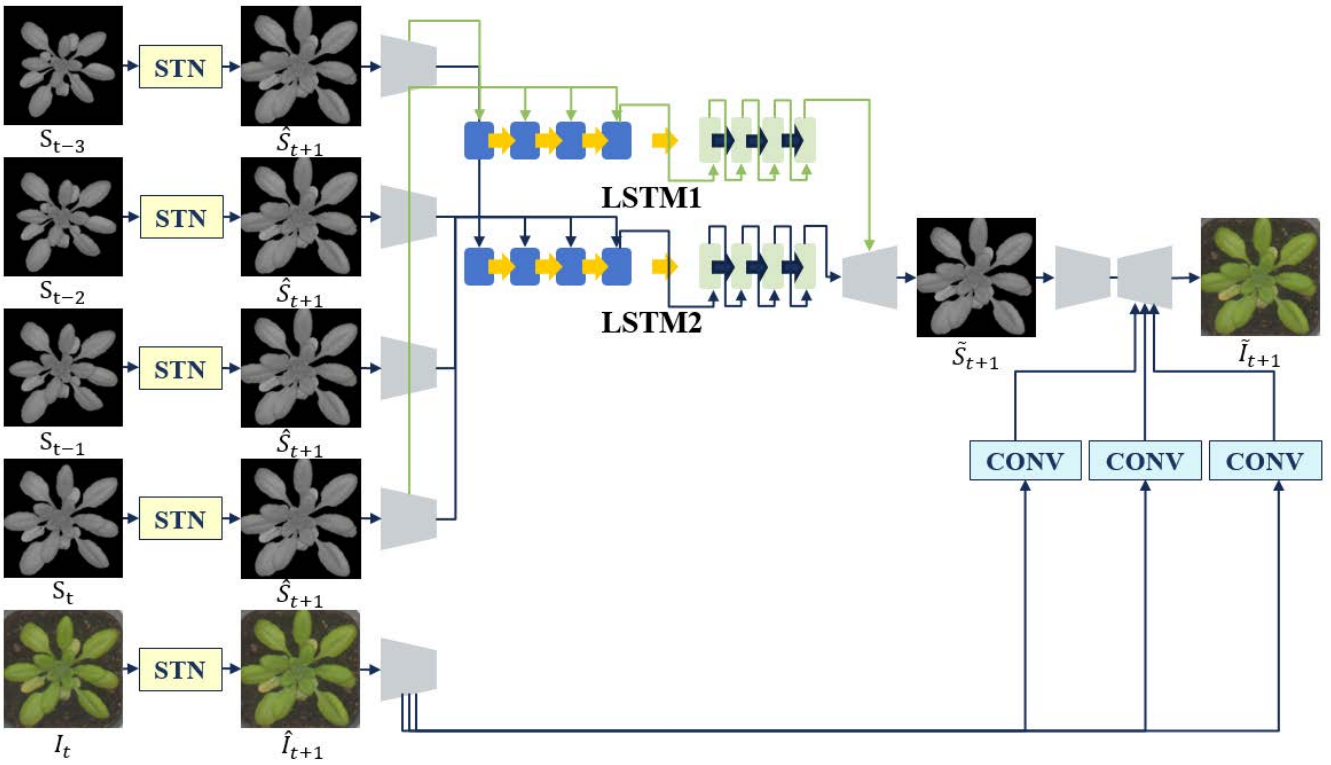

FIGURE 2. The architecture of the shape subnet. The ((*t*-3)-th∼ *t*-th) gray shape images pass through the spatial transformer network (STN) to be aligned to the target ($t+1$)-th image. The results from the STN are passed through two LSTMs to estimate the ($t+1$)-th shape image. Lastly, the color is reconstructed by employing RGB image.

In the meantime, there have been studied a number of deep learning methods to predict future plant images. Research on plant growth prediction mainly adopts the approach of future video generation. The conventional plant growth prediction popularly uses an auto-encoder structure with ConvLSTM, which is often used in video prediction. Multiple auto-encoders corresponding to individual input are fused through ConvLSTM [14]. In [15], the network takes RGB images of a plant and their labels as an input, simultaneously, and again produces both label and RGB images in the output. Although it is a simple network architecture consisting of only an auto- encoder with ConvLSTM, it introduces four loss functions. In [3], the network uses GAN in addition to an auto-encoder with ConvLSTM. Furthermore, by extending [15], several auto-encoders are hierarchically fused with 1/2 and 1/8 resolution via ConvLSTM. The method was also modified by replacing the LSTM based fusion with simple concatenation of CNN-based channels.

On the other hand, there are studies that measure the features or contours of individual leaves rather than the overall appearance of plants and utilize them for growth prediction or recognition [21], [22]. In another aspect, a method of predicting growth has been studied by focusing on the prediction of indicators such as live weight rather than the appearance of plants [23].

The proposed network is highly different from these conventional plant image generation methods. The future frame generation task is divided into the shape and RGB prediction processes, which are performed in both the gray and RGB domains. The predicted gray shape image is then combined with the RGB image estimated by a STN for its colorization. In addition, extensive experiments are performed using three datasets, which consist of our own established dataset as well as two public ones.

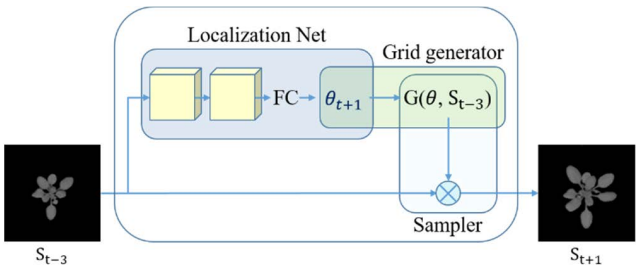

FIGURE 3. Spatial transformer network. Figure reproduced from [6].

III. THE PROPOSED METHOD

As plants grow, their leaves change dynamically. For example, the leaves are gradually enlarged and new leaves are generated. In addition, their movements are diverse from

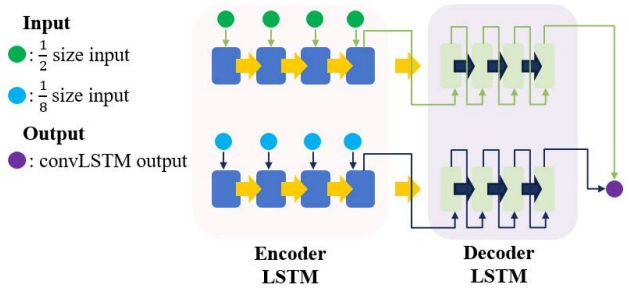

FIGURE 4. Detailed structure of U-Net with two LSTMs. The 1/2-size and 1/8-size shape images pass through the two LSTMs. The outputs from the LSTMs are passed through the decoder to generate the final shape image.

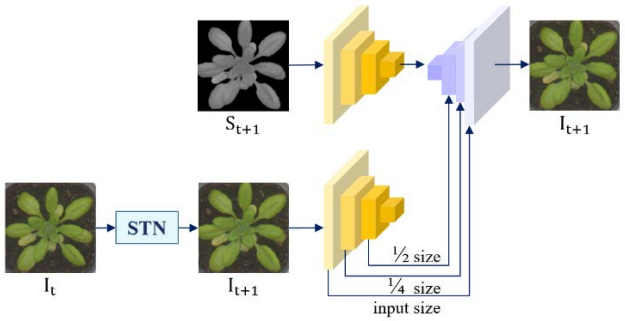

FIGURE 5. Detailed structure of color reconstruction subnet. The plant shape and RGB images are hierarchically fused at the decoder.

bending to spreading with time. In this paper, we attempt to predict the future plant image from a number of past and current images. This technology is particularly useful in plant factories, where the environmental factors for plant cultivation can be intelligently controlled according to the status of plant growth. One of the simplest methods to quantitatively calculate the amount of growth is to measure the leaf area. Thus, it is very important to accurately estimate the shape of the plant. This is the reason why the proposed estimation process is divided into 'shape estimation' and 'color reconstruction' parts.

Fig. 2 illustrates the overall architecture of the proposed deep network. Gray and RGB images are fed into the network as inputs for shape estimation and color reconstruction, respectively. The shape of the leaves is estimated with gray images, while the color information is recovered with an RGB image which is the affine-transform version of a current one. The proposed deep network is divided into two subnets.

First, an affine transformation is performed by the STN to estimate the amount of plant growth from the past or current time to future target time. To avoid interference of background during the transformation, the leaves are segmented in advance. In addition, it was found from diverse experiments that the color information can be an obstacle for accurate affine transformation. Thus, the shape estimation is performed in gray domain. After the STN, the input images at distinct times are geometrically aligned to a future ground truth shape, and the resulting affine-transformed images are fed into the shape subnet to predict the future shape of the input plant. The color information needs to be replenished because the generated image is gray. For the task of color reconstruction, the affine-transformed RGB image is fused with the estimated shape image through an auto-encoder. It is estimated that the higher accuracy of the affine transformation could be obtained from a current image rather than from past images. Thus, we use an RGB image transformed from just a current time for the fusion. The two images are hierarchically fused by connecting the encoder of the RGB to the decoder of the shape as shown in Fig. 2.

A. SPATIAL TRANSFORMER NETWORK

The spatial transformer network (STN) conducts the affine transformation for the input, and finds affine parameters. In STN, there are three parts, localization network, grid generator, sampler. Fig. 3 shows the detailed structure of the STN. The localization network takes the input image and outputs a set of affine transformation parameters, θ. We input each of four gray shape images (from S_{t-3} to S_t) to the localization network, and each individual shape image is affine-transformed to the target image, S_{t+1}. This affine operation with the STN is expressed by

$$\theta_{t+1} = f_{loc}(S_{t-3}) = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 \\ \theta_3 & \theta_4 & \theta_5 \end{bmatrix} \tag{1}$$

where θ_{t+1} is a set of affine parameters, output from the localization network. Note that the t-th RGB image, I_t, is not transformed, and the parameters for S_t are used directly for the STN because the STN has poor performance on the RGB image. In other words, $f_{loc}(I_t)$ is equal to $f_{loc}(S_t)$ as follows.

$$f_{loc}(I_t) = f_{loc}(S_t) \tag{2}$$

The grid generator then creates a sampling grid, which is a set of points to be sampled from the input image. The predicted transformation parameters θ_{t+1} are used to generate the sampling grid. The grid generator applies the affine transformation as

$$\begin{pmatrix} x_i^{t-3} \\ y_i^{t-3} \end{pmatrix} = \theta_{t+1} \begin{pmatrix} x_i^{t+1} \\ y_i^{t+1} \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 \\ \theta_3 & \theta_4 & \theta_5 \end{bmatrix} \begin{pmatrix} x_i^{t+1} \\ y_i^{t+1} \\ 1 \end{pmatrix} \tag{3}$$

where (x_i^{t+1}, y_i^{t+1}) are the target coordinates of the regular grid in the shape image S_{t+1}, (x_i^t, y_i^t) are the source coordinates in the shape image S_t that define the sample points, and θ_{t+1} is the affine transformation matrix, which applies cropping, translation, rotation, scaling, and skewing to the input image. The affine transformation matrix consists of six parameters which are produced by the localization network.

$$\theta_{t+1} = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix} \tag{4}$$

The attention allows cropping, translation, and isotropic scaling by varying s, t_x, and t_y. Finally, the sampler takes the input image and sampling grid, and produces the affine-transformed output by performing bilinear sampling to generate the shape image S_{t+1}.

B. SHAPE ESTIMATION

After performing the STN, the future plant shape is predicted by employing U-Net with two LSTMs. For shape estimation, we use gray plant images rather than binary or RGB images. The plant is segmented from the original plant image, and the resulting image is converted into grayscale. The gray plant image has no background. So, it can focus on leaves only without the interference of its surrounding background.

Eventually, it was found through diverse experiments that using gray images results in better shape prediction. First of all, for shape estimation, each of the four sequential gray images that passes through the STN is fed into the encoder which is connected to a decoder by two LSTMs. In detail, there are two LSTMs that works in 1/2 and 1/8 image sizes on the auto-encoder structure. After passing through the encoder, the 1/2-size image passes through the first LSTM, and the 1/8-size image passes through the second LSTM, as shown in Fig. 4. Four time-series images are sequentially combined by U-Net with two LSTMs. A sequential combination is made at the two distinct hierarchical layers. Likewise, the outputs of the two LSTMs are combined by decoder LSTMs as shown in Fig. 4. Using two LSTMs, we can predict the plant shape, S_{t+1} that is sophisticated and accurate. The final shape output S_{t+1} is obtained by fusing the outputs from the two decoder LSTMs with a U-Net decoder.

C. COLOR RECONSTRUCTION

The previous shape subnet generates only a gray shape image. The color information should be reconstructed additionally, and the estimated shape image still lacks of texture information. We fuse the estimated RGB image, I_{t+1} with the shape from the auto-encoder.

The RGB image, I_{t+1} is the aligned version of a current RGB to a target. Note that the STN is performed with the parameter set θ_{t+1} already obtained by the shape subnet without performing the STN with the RGB image actually. The final shape output S_{t+1} from the shape subnet and the $(t+1)$-th RGB image I_{t+1} are employed for restoring the color information. S_{t+1} and I_{t+1} pass through their respective encoders, and are hierarchically fused at a decoder. In detail, the RGB image I_{t+1} passes through the encoder and its resolution is reduced to 1/2 and 1/4 by convolution. Then, they are concatenated to the shape decoder. The color reconstruction process is detailed in Fig. 5. Finally, we can get the result of predicting the $(t+1)$-th plant image from the $(t$-3)-th$\sim$ t-th plant images. The final output shows good growth tracking and sophisticated color and texture for each leaf as it goes through the shape and color subnets in sequence.

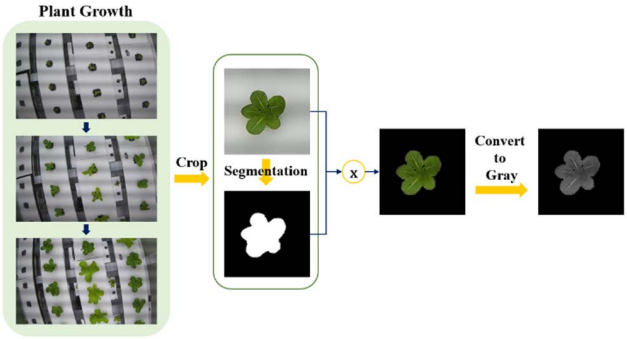

FIGURE 6. Pre-processing step of the Butterhead training dataset. The plants were grown in a plant factory and the RGB, binary, and gray plant images were obtained.

D. LOSS FUNCTION

We used the L1 loss to train our network. The network was trained with other losses such as mean square error and SSIM, and the best results were obtained when trained using L1 loss. There are six L1 losses. Four of these are STN losses, and the remainder are shape and texture losses, as the follows.

$$L = \mu_1 L_{STN} + \mu_2 L_{SHAPE} + \mu_3 L_{COLOR} \tag{5}$$

where μ_1, μ_2 and μ_3 are the coefficients that are experientially determined. The first term, L_{STN}, is a loss after passing through the spatial transformer network. We calculate L_{STN} for the four gray shape inputs, which is given by

$$L_{STN} = \sum_{k=0}^{3} L1(S_{t-k}, S_{t+1}) \tag{6}$$

The second term, L_{SHAPE}, is the loss between the final shape output image (which is the output of the shape subnet) and the $(t+1)$-th gray shape ground truth, and is given by

$$L_{SHAPE} = L1(\tilde{S}_{t+1}, S_{t+1}) \tag{7}$$

The last term, L_{COLOR}, is the loss between the final RGB output of the overall network and the $(t+1)$-th RGB ground truth, and is given by

$$L_{COLOR} = L1(\tilde{I}_{t+1}, I_{t+1}) \tag{8}$$

E. USE SCENARIO

Practical usage is illustrated in Fig. 7. A plant is captured every day (one-day interval). Four snapshots (three past and one current) are required as the network input, and then, the network generates the plant image of the next day. From the network output, we can estimate the plant growth state at the next day. In this way, the network can predict the future plant image from just previous four images.

Although the time interval is set to one-day in this paper, the network can be trained for any interval. However, the prediction performance decreases as the time interval increases. The future work is to increase the time interval while the performance is preserved. Also, see Fig. 13 which compares the performance between one-day and two-days intervals.

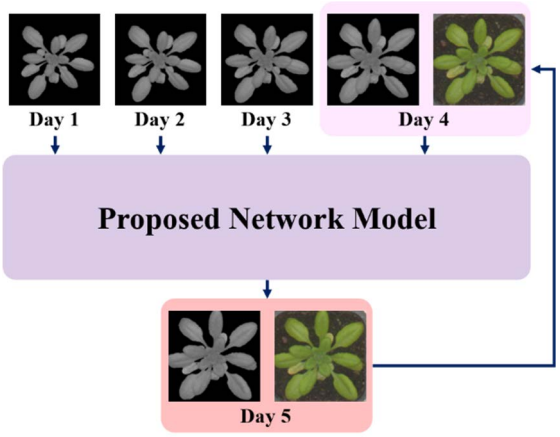

FIGURE 7. The use scenario of the proposed network model.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTING

Our proposed network is trained with the plant datasets. The input image of size 128 × 128 is used as it is without dividing the image into patches. The network is implemented using the PyTorch framework on a PC with 2 NVIDIA RTX 2080ti GPUs [16]. We adopted the Adam optimizer for loss optimization. The batch size is 8 [17]. The initial learning rate is 0.0001 and is divided by 10 for every 30k iterations.

The air temperature, relatively humidity, and CO2 concentration inside a plant factory throughout the experiment (mean ± standard deviation) were 18.5 ± 0.4°C, 65.2 ± 3.1%, and 742 ± 106 ppm, respectively.

B. TRAINING DATASET

We used a variety of datasets for the experimental evaluations on different types of plants. In particular, we grew plants in a plant factory to create our own dataset. The acquired plant images were cropped and segmented for our own plant dataset. The proposed network is evaluated with the three kinds of datasets, Aberystwyth [19], Komatsuna [20] and our own Butterhead. The Aberystwyth and Komatsuna datasets consist of time-series RGB leaf images and their corresponding binary shapes.

We grew green Butterhead lettuce, which is one of the types of lettuce directly in a plant factory. Our own dataset, which is named as "Butterhead", consists of several Butterhead growth images captured at one-day intervals over a period 18 days. The dataset pre-processing step is shown in Fig. 6. First, a camera is attached to the ceiling of a plant factory, Butterhead lettuce seeds are planted in soil, and they are grown for 18 days. The acquired Butterhead images are cropped to 128 × 128 size so that the plant is located in the center of an image. Next, from the resulting RGB image, a binary leaf shape image is obtained by segmentation. The leaves of the Butterhead images can be easily segmented by appropriately thresholding RGB values due to almost uniform background. The binary shape is then multiplied with the RGB image to extract only the foreground plant leaves without the image

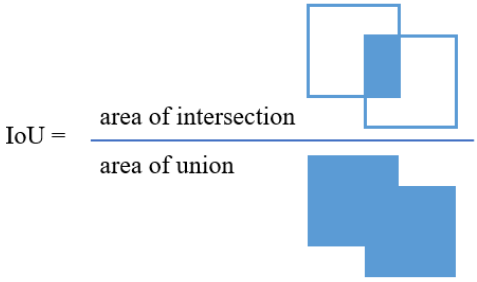

FIGURE 8. Intersection of Union. IoU is used to evaluate the prediction of plant leaf shape.

background. Finally, the images acquired in the previous step are converted to gray. The remaining two datasets also go through the same process.

For data augmentation, we rotated the plant image 90 degrees, 180 degrees, 270 degrees and flipped the plant images along the x- axes and y-axes. As a result, training data in the Aberystwyth dataset is a total of 1,674 images, Komatsuna dataset is a total of 448 images and our Butterhead dataset is a total of 972 images. Averystwyth [19], Komatsuna [20] and Butterhead have one type of plant. Averystwyth has grown for 31 days. The Averystwyth leaf evaluation dataset is composed of Arabidopsis Thaliana plants. Nine plants are used for training and one plant is used for test. A total of ten different plants are used in the experiment. Komatsuna has grown for 56 days. Komatsuna dataset consists of five Komatsuna plants. Four plants are used for training and one plant is used for test. A total of five different plants are used in the experiment. Butterhead has grown in a plant factory for 18 days. For the experiment, eight plants are assigned to the training dataset, and one plant is assigned to the test dataset.

When training, four $(t-3)$-th $\sim$ t-th time series images and one t-th RGB image are used as inputs. At the end of the training, the $(t+1)$-th gray shape image and RGB image are obtained as outputs.

C. COMPARISON WITH THE EXISTING METHODS

Various experiments have been conducted to demonstrate that our network is the best suited method for plant growth prediction. Furthermore, plant growth prediction is evaluated with the existing video prediction networks [12], [13] as well as plant growth methods for comparison. Video frame generation is very similar to plant growth prediction because the movement of an object is predicted over time in both tasks. The task of video prediction estimates the next motion of a moving object, which is very similar to the growth of a plant in our plant prediction task if the moving object is considered as a plant. In other words, object motion corresponds to plant growth in our work.

The future plant images generated by U-Net [3], U-Net-LSTM [3], MC-Net [12], and HP-Net [13] are evaluated. The experiments of the existing methods were conducted by directly putting a sequence of plant images into the

FIGURE 9. Comparison of plant image prediction results from three datasets. The first and second rows from the Averystwyth [19] dataset, the third and fourth rows from our Butterhead dataset, and the last two rows from the Komatsuna [20] dataset. Note that the background is not included to the task of plant image prediction, and the object of a plant only is predicted. Artifact in the background occurs during color reconstruction.

networks instead of video frames. Fig. 9 shows the experimental results. For comparison in detail, take a closer look at Fig. 9. Fig. 9 (a) shows the RGB ground truth while Figs. 9 (b) [3] and 9 (c) [3] show the experimental results from the existing growth prediction methods. The only difference between the two methods is how to fuse multiple encoded inputs. Fig. 9 (d) [12] and (e) [13] show the respective results from MC-Net and HP-Net, which are used for video prediction.

As shown in Fig. 9 (b), in the existing U-Net leaves growth is not predicted correctly. The shapes of the individual leaves are distorted. Moreover, the leaves are blurred and heavy artifacts occur. U-Net fails to detect the growth movement in the time-series data because of the simple concatenation

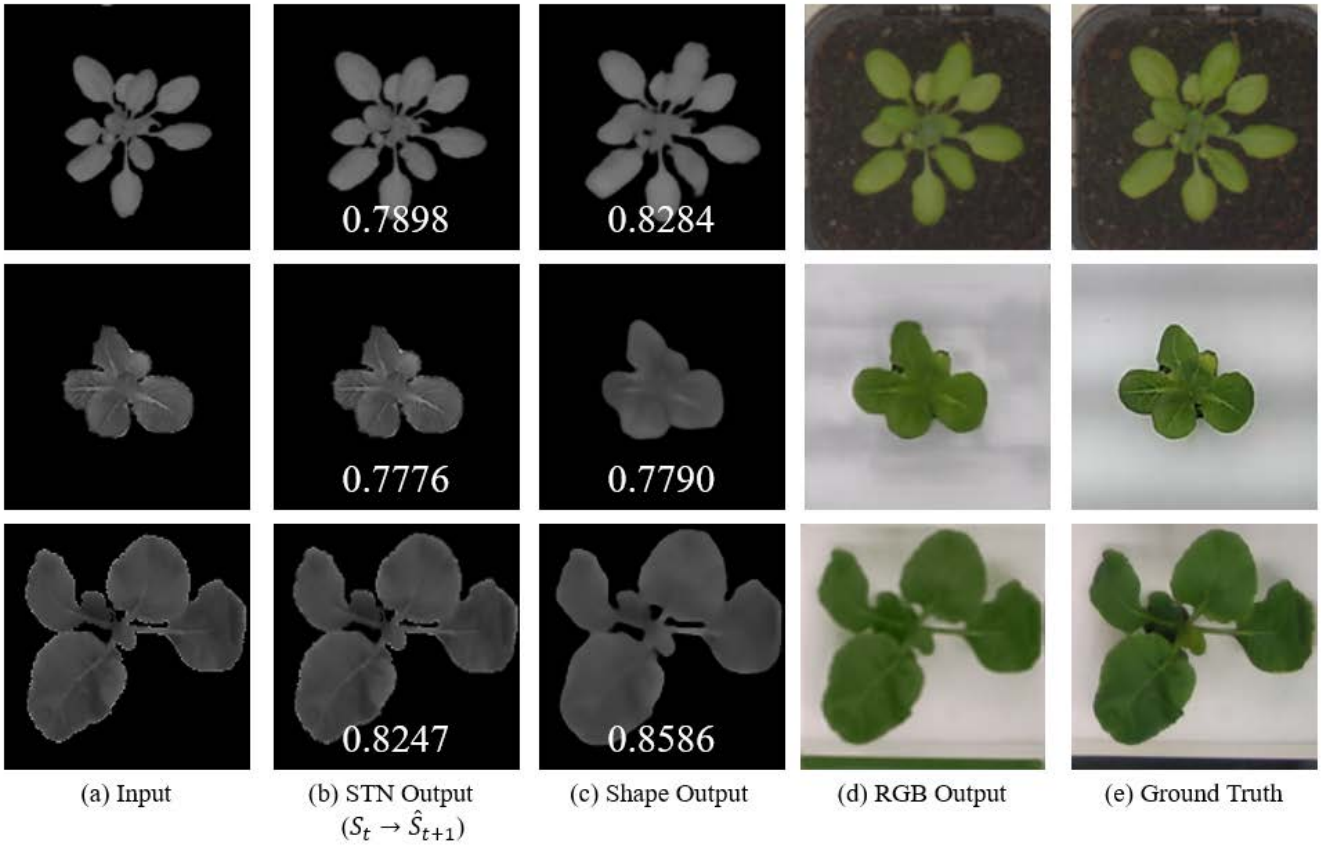

FIGURE 10. Intermediate outputs of the proposed network structure. (b) shows the outputs of the fourth STN from the top, (c) the outputs of the shape estimation subnet, and (d) the final RGB outputs of the network. Note that the numbers in (b) and (c) indicate the IoU.

TABLE 2. PSNR and SSIM comparisons of U-Net, U-Net-LSTM, MC-Net, HP-Net and the proposed network.

Dataset	Metric	Algorithm				
		U-Net [3]	U-Net-LSTM [3]	MC-Net [12]	HP-Net [13]	**Proposed**
Aberystwyth [19]	PSNR(dB)	30.08	30.29	30.36	30.31	**30.55**
	SSIM	0.8283	0.8372	0.8348	0.8316	**0.8425**
Butterhead	PSNR(dB)	21.61	22.04	22.56	22.34	**22.95**
	SSIM	0.7607	**0.7910**	0.7830	0.7769	0.7862
Komatsuna [20]	PSNR(dB)	25.35	25.22	25.02	24.66	**25.95**
	SSIM	0.8937	0.8945	0.8995	0.8904	**0.9042**

fusion, consequently leading to inaccurate prediction of the subsequent leaf.

Replacing concatenation with LSTM for fusion improves the growth prediction of the individual leaves, as shown in Fig. 9 (c). Compared with U-Net, we can see that U-Net-LSTM estimates the shape of individual leaves more accurately. The importance of LSTM in time series data is also confirmed in the other methods. Note that all the methods except for Fig. 9 (b) adopt one or more LSTMs in the network architecture.

The first result image in Fig. 9 (d) is similar to the ground truth. Without leaf artifact, the leaf boundary is less distorted and the individual leaves are well generated to the right size. However, the second shows more artifacts than the first. At the beginning of plant growth, the amount of growth over time can be clearly recognized at a glance. However, as the growth is matured, its growth rate decreases. For this reason, we can see that the leaf shape prediction is poor in the second result compared to the first. MC-Net is originally divided into motion and content encoders as our proposed network structure is divided into shape estimation and color reconstruction subnets. For the comparative experiments, we use motion encoders to predict the leaf shape and content encoders to replenish the leaf color. Fig. 9 (d) shows that the adoption

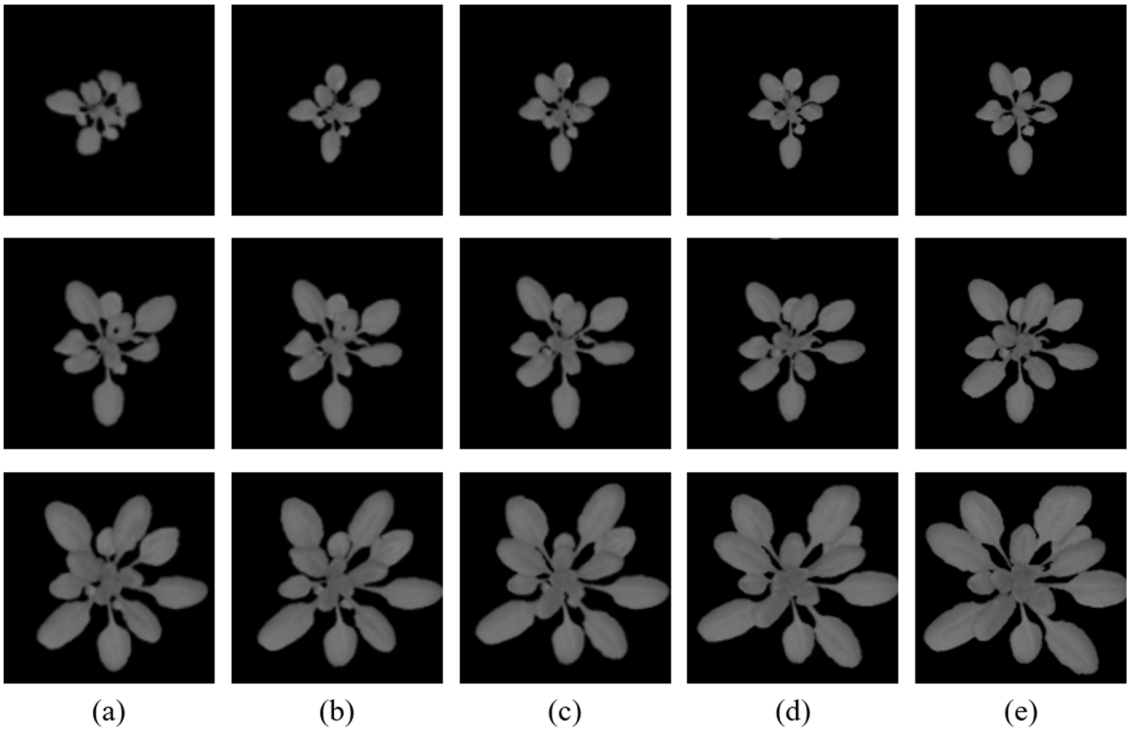

FIGURE 11. The aligned images with STN. (a) is the output of $S_{t-3} \rightarrow \hat{S}_{t+1}$, (b) is the output of $S_{t-2} \rightarrow \hat{S}_{t+1}$, (c) is the output of $S_{t-1} \rightarrow \hat{S}_{t+1}$, and (d) is the output of $S_t \rightarrow \hat{S}_{t+1}$.

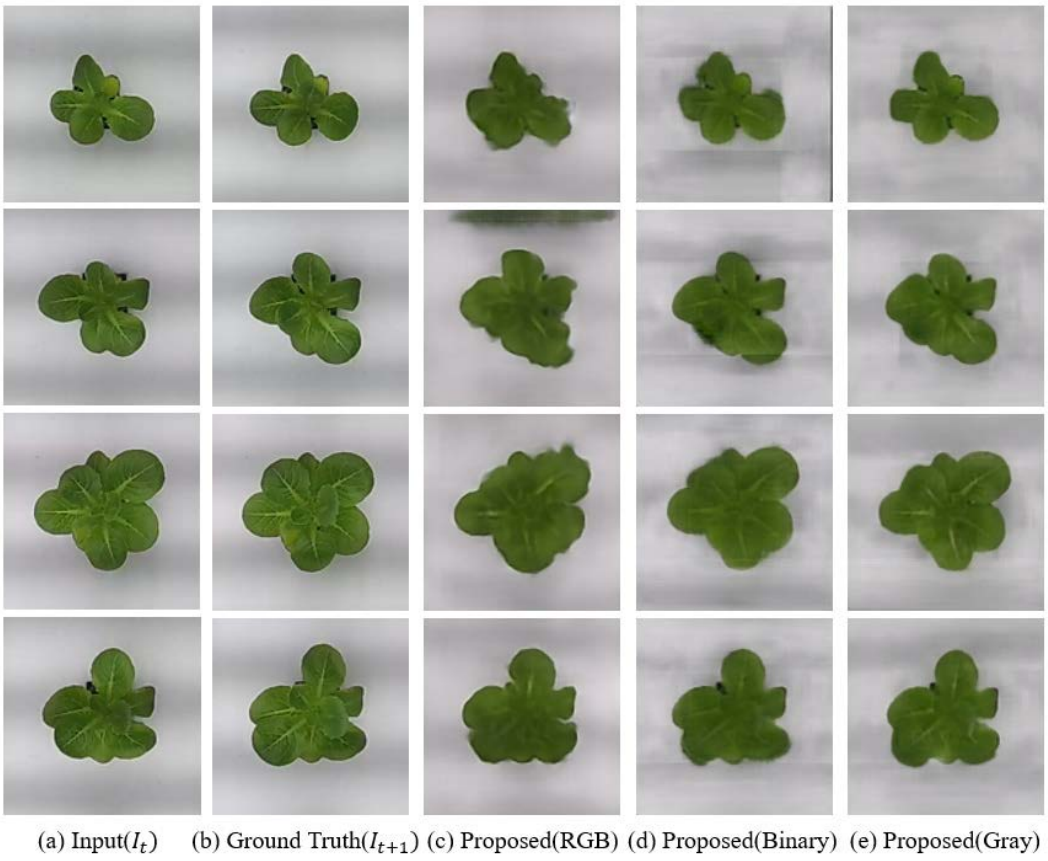

FIGURE 12. Comparison of different shape image types.

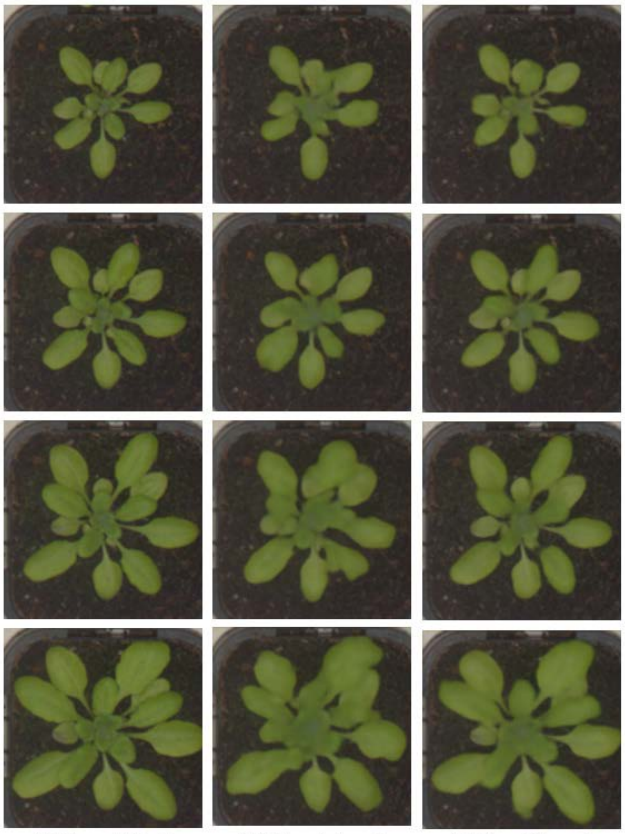

FIGURE 13. Comparison for different time intervals.

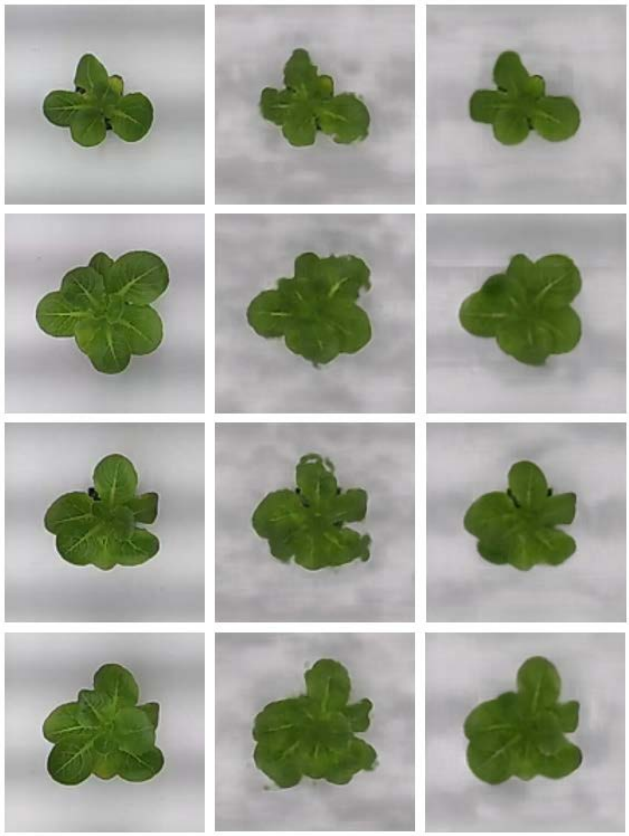

FIGURE 14. Comparison between results obtained with and without shape output.

of content encoders in MC-Net improves the textures of the plant leaves compared to Fig. 9 (b) and Fig. 9 (c), for which no content reinforcement modules are used.

Fig. 9 (e) achieves a similar performance with less artifact for the first and second images, regardless of the degree of plant growth. However, if individual leaf is observed closely, they grow less, so the leaves are either comparatively smaller or shorter. And the overall shape of the leaves is distorted compared to the ground truth. HP-net originally consists of pose estimation and image generation. Similar to MC-Net, we predict the shape of plant leaves with pose estimation for the comparative experiments and enhance the texture inside the plant leaves using image generation. Three convolutional LSTMs are used for the pose estimation. Because of the poor posture estimation, the leaves do not grow properly and the shapes are distorted. The image generation subnet creates the final image by concatenating the difference between the t-th and $(t+n)$-th pose with the t-th RGB image. Look at the leaf texture in Fig. 9 (e), the texture is blurred. These experimental results show that the image generation subnet performs poorly in plant growth prediction.

The proposed method in Fig. 9 (f) shows highly accurate prediction of plant leaves. It is superior to the existing methods in terms of plant boundary, artifact, and blur. In particular, the proposed method performs better in predicting the plant leaf shapes than the existing methods. This is because the STN first aligns the plant leaves, and then they are fused with LSTM for leaf shape prediction. Our experimental results demonstrate that the blur is reduced while the texture of the leaves is enhanced. In addition, if the execution time is compared, the conventional U-Net-LSTM [3] takes 3.298 seconds while the proposed method takes 3.351 seconds. The existing method predicts a future plant image directly from RGB inputs. On the other hand, the proposed method consists of two subnets (shape estimation and color reconstruction). So, the network is more complex and takes longer to go through STN. However, the usage scenario of the proposed method does not require real-time processing. The proposed prediction is used for the control of plant cultivation environment whose interval is relatively long (one-day at least). In other words, the prediction is made every day, and based on the prediction, the environment factors of the plant factory are controlled.

Fig. 10 shows the intermediate outputs of the proposed network shown in Fig. 2. Initially, the shape inputs are aligned to the target, S_{t+1} by passing through the STN. Fig. 10 (b) shows the STN output for the input S_t. The aligned outputs are then fused with the U-Net with two LSTMs to estimate the shape output, which is shown in Fig. 10 (c). If the IoU

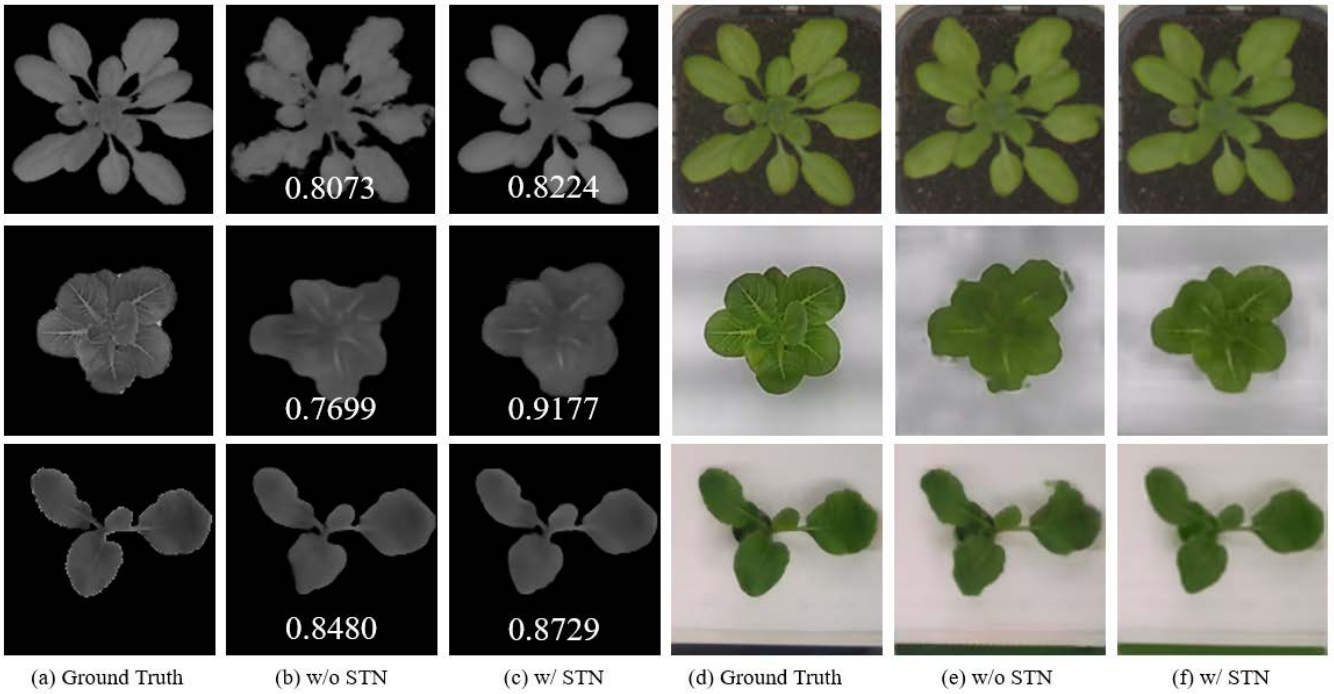

FIGURE 15. Comparison between results obtained with and without STN. (a) gray ground truth images, (b) the outputs of the shape estimation subnet without STN, (c) the outputs of the shape estimation subnet with STN, (d) RGB ground truth images, (e) the final RGB outputs of the network without STN, and (f) the final RGB outputs of the network with STN. The numbers in (b) and (c) represent the IoU. The results for the Averystwyth[19], Butterhead, and Komatsuna[20] datasets are shown from the top to bottom rows.

values (the measure of binary shape estimation) are compared between Figs. 10 (b) and (c), the accuracy of the shape output is higher than that of the STN output, as expected. The result confirms that the shape estimation subnet can generate a shape image that is more accurate than the affine-transformed version. The final RGB output in Fig. 10 (d) shows that the leaf texture is restored through the RGB reconstruction subnet.

Table 2 shows the SSIM and PSNR comparisons of the results from U-Net, U-Net-LSTM, MC-Net, HP-Net, and the proposed network. Please, note that SSIM and PSNR are quantitative evaluation metrics for digital images. The former measures scene structure similarity while the latter does the error of image signals. The formulas for SSIM and PSNR are as follows.

$$SSIM(\mathrm{A}, \mathrm{B}) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\mu_A^2 + \mu_B^2 + C_2)} \quad (9)$$

where μ is the mean, σ is the standard deviation, C is a predefined constant

$$PSNR = 10 \log \frac{S^2}{MSE} \quad (10)$$

where s is the maximum value in the image and MSE indicates mean square error. Detailed explanations are provided in [41], [42].

Overall, U-Net-LSTM and MC-Net achieve better results than HP-Net and U-Net among the conventional networks, and our proposed network achieves the best quantitative results.

Fig. 11 shows the intermediate results of the proposed network. The $(t-3)$-th $\sim$ t-th input images are aligned to the $(t+1)$-th image with STN as a preprocessing. As can be seen in Fig. 11, the accuracy of STN is higher if the time interval between the current image and the target image is short. The STN output of $S_t \rightarrow \hat{S}_{t+1}$ with the shortest time interval is the most similar to ground truth. This is straightforward as expected. By aligning each input to the target as a preprocessing, the shape estimation subnet efficiently works, and it can generate better accurate shape.

D. EFFECT OF SHAPE IMAGE TYPES

The proposed shape estimation subnet runs in the gray domain. Its performance is compared for the different image types of binary, gray, and RGB images. As shown in Fig. 12, prediction with binary images causes blur and shape distortion in the final RGB output, resulting in poor visual quality and inaccurate shape estimation. RGB inputs are poor at predicting the shapes of the leaves, leading to shape boundary artifacts. Note that for RGB inputs, the color reconstruction subnet is not needed, and the shape estimation subnet works only on RGB. The proposed gray inputs accomplish less blur and better texture restoration for each leaf as confirmed in Fig. 12 (e). Comparing Fig. 12 (c) with Fig. 12 (d) and Fig. 12 (e), we can see the importance of the separate shape estimation. In addition, gray images, which are simpler than RGB images and contain more information than binary images, are the most suitable for the shape estimation subnet. Recall that for RGB inputs, the shape estimation subnet

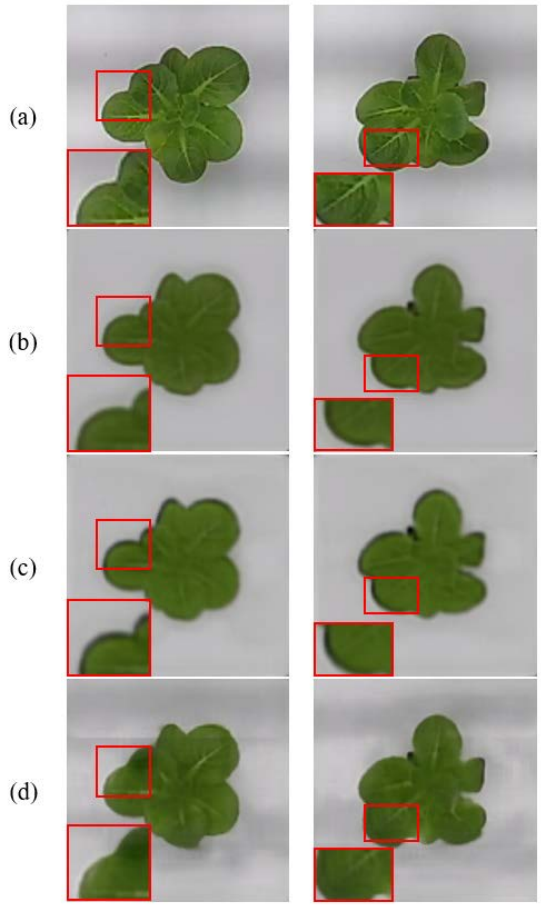

FIGURE 16. Comparison between shape estimation subnet structures. (a) Ground Truth (b) Proposed (one LSTM) (c) Proposed (two Concatenations) (d) Proposed (two LSTMs).

generates the final RGB output without the color reconstruction subnet, unlike the binary and gray inputs.

E. EFFECT OF TIME INTERVAL

When time-series plant images are fed into the network, the time interval between inputs is increased to identify its effect on the plant growth prediction accuracy. The proposed network is evaluated for the two time intervals, of one-day and two-days. As shown in Fig. 13, the capability of plant growth prediction is superior for the one-day interval rather than the two-days interval as expected.

F. ABLATION STUDIES

The proposed network is evaluated by changing its structure to demonstrate its superior performance. First, the shape output in the shape estimation subnet is removed to find its importance as a constraint. Fig. 14 shows the results of the experiment with and without the shape output. Note that 'with shape output' means adding a shape loss at the output of the shape estimation subnet. As shown in Fig. 14 (b), the leaf boundary is not restored properly for the case of 'without shape output'. The results are closer to input rather than ground truth. For the case of 'with shape output' in Fig. 14 (c), plant leaf shape prediction is better without artifact, and unlike 'without shape output', it is closer to the ground truth than the input. This confirms the significance of the shape constraint.

Next, our network is compared with the network without the STN preprocessing to know the effectiveness of the STN. If there is no STN preprocessing, the boundaries of the leaves are not clear and their shape distortion is severe. On average, the IoU value is lower when STN preprocessing is not performed. In particular, the second row of Fig. 15 (b) shows a significant difference between with and without STN preprocessing in IoU values. As a result, we can see that the alignment of the plant shape before the shape estimation subnet can contribute to clear and less distorted shape.

Finally, we evaluate the performance of different shape estimation subnet structures. The proposed subnet is compared to two structures. One is to use a single LSTM, and the other is to fuse multiple inputs by concatenation instead of the proposed LSTM fusion. As confirmed in the red box in Fig. 16 (b), the size of the predicted leaf is still small. This means that 'one LSTM' fails to predict the plant growth properly. For concatenation in the red box of Fig. 16 (c), we observe a gray boundary artifact in the outer area of the leaf. This is an error caused by the poor plant shape prediction. The proposed method shown in Fig.16 (d) is the most similar to the ground truth. This demonstrates that fusing with two LSTMs produces the best results.

V. CONCLUSION

In this paper, we proposed a new plant growth prediction network, which consists of two subnets; shape estimation and color reconstruction. In the former, the shape of the plant leaves is estimated with gray images to increase the accuracy of affine transformation by the STN. Then, the latter performs color reconstruction by the hierarchical fusion of shape and RGB images with auto-encoder. Unlike existing networks, we first align four shape inputs with a future target as a preprocessing, and this achieves better shape prediction. We evaluate three different kinds of plant datasets where the Butterhead dataset is created by growing plants ourselves in plant factory. Diverse experiments demonstrate that the proposed network shows the superior prediction performance for plant growth, compared with the existing plant growth prediction method and video frame generation methods.

In the future, we will study the prediction of plant growth using other information (e.g., illumination on leaves) as well as plant image.

ACKNOWLEDGMENT

(Joo-Yeon Jung and Sang-Ho Lee contributed equally to this work.)

REFERENCES

[1] J. I. L. Morison and D. W. Lawlor, "Interactions between increasing CO_2 concentration and temperature on plant growth," *Plant, Cell Environ.*, vol. 22, no. 6, pp. 659–682, 1999.

[2] S. Nesteruk, D. Shadrin, V. Kovalenko, A. Rodriguez-Sanchez, and A. Somov, "Plant growth prediction through intelligent embedded sensing," in *Proc. IEEE 29th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2020, pp. 411–416.

[3] T. Hamamoto, H. Uchiyama, S. Atsushi, and R. I. Taniguchi, "3D plant growth prediction via image-to-image translation," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl. (VISIGRAPP)*, 2020, pp. 153–161.

[4] A. Rizkiana, A. P. Nugroho, N. M. Salma, S. Afif, R. E. Masithoh, L. Sutiarso, and T. Okayasu, "Plant growth prediction model for lettuce *(Lactuca sativa.)* in plant factories using artificial neural network," in *Proc. IOP Conf.: Earth Environ. Sci.*, 2021, vol. 733, no. 1, Art. no. 012027.

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

[6] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2017–2025.

[7] A. Verma, M. Sharma, R. Hebbalaguppe, E. Hassan, and L. Vig, "Automatic container code recognition via spatial transformer networks and connected component region proposals," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 728–733, doi: 10.1109/ICMLA.2016.0130.

[8] C.-H. Lin and S. Lucey, "Inverse compositional spatial transformer networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2568–2576.

[9] S. K. Sønderby, C. Kaae Sønderby, L. Maaløe, and O. Winther, "Recurrent spatial transformer networks," 2015, *arXiv:1509.05329*.

[10] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion GAN for future-flow embedded video prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1744–1752.

[11] J. van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," 2017, *arXiv:1701.08435*.

[12] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," 2017, *arXiv:1706.08033*.

[13] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3560–3569.

[14] S. H. I. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[15] S. Sakurai, H. Uchiyama, A. Shimada, and R.-I. Taniguchi, "Plant growth prediction using convolutional LSTM," in *Proc. VISIGRAPP*, 2019, pp. 1–9.

[16] *PyTorch*. Accessed: Oct. 2016. [Online]. Available: http://pytorch.org

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[18] J. Bell and H. Dee, "Aberystwyth leaf evaluation dataset," 2016, p. 2, doi: 10.5281/zenodo 168158.17-36.

[19] H. Uchiyama, S. Sakurai, M. Mishima, D. Arita, T. Okayasu, A. Shimada, and R.-I. Taniguchi, "An easy-to-setup 3D phenotyping platform for KOMATSUNA dataset," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2038–2045.

[20] A. Khmag, S. A. R. Al Haddad, R. A. Ramlee, N. Kamarudin, and F. L. Malallah, "Natural image noise removal using nonlocal means and hidden Markov models in transform domain," *Vis. Comput.*, vol. 34, no. 12, pp. 1661–1675, Dec. 2018.

[21] A. Khmag, S. A. R. Al-Haddad, and N. Kamarudin, "Recognition system for leaf images based on its leaf contour and centroid," in *Proc. IEEE 15th Student Conf. Res. Develop. (SCOReD)*, Dec. 2017, pp. 467–472.

[22] Z. Wang, H. Li, Y. Zhu, and T. Xu, "Review of plant identification based on image processing," *Arch. Comput. Methods Eng.*, vol. 24, no. 3, pp. 637–654, Jul. 2017.

[23] S. Nagano, S. Moriyuki, K. Wakamori, H. Mineno, and H. Fukuda, "Leaf-movement-based growth prediction model using optical flow analysis and machine learning in plant factory," *Frontiers Plant Sci.*, vol. 10, p. 227, Mar. 2019.

[24] R. Akula and G. A. Ravishankar, "Influence of abiotic stress signals on secondary metabolites in plants," *Plant Signaling Behav.*, vol. 6, no. 11, pp. 1720–1731, Nov. 2011.

[25] J. Fuhrer, "Agroecosystem responses to combinations of elevated CO_2, ozone, and global climate change," *Agricult., Ecosyst. Environ.*, vol. 97, nos. 1–3, pp. 1–20, Jul. 2003.

[26] D. A. Way and R. Oren, "Differential responses to changes in growth temperature between trees from different functional groups and biomes: A review and synthesis of data," *Tree Physiol.*, vol. 30, no. 6, pp. 669–688, Jun. 2010.

[27] R. J. Norby and Y. Luo, "Evaluating ecosystem responses to rising atmospheric CO_2 and global warming in a multi-factor world," *New Phytologist*, vol. 162, no. 2, pp. 281–293, 2004.

[28] J. Monteiro and J. Barata, "Artificial intelligence in extended agri-food supply chain: A short review based on bibliometric analysis," *Proc. Comput. Sci.*, vol. 192, pp. 3020–3029, Jan. 2021.

[29] R. Pahuja, H. K. Verma, and M. Uddin, "A wireless sensor network for greenhouse climate control," *IEEE Pervasive Comput.*, vol. 12, no. 2, pp. 49–58, Apr. 2013.

[30] K. O. Flores, I. M. Butaslac, J. E. M. Gonzales, S. M. G. Dumlao, and R. S. J. Reyes, "Precision agriculture monitoring system using wireless sensor network and raspberry Pi local server," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2016, pp. 3018–3021.

[31] J. Oliveira, J. Boaventura-Cunha, and P. M. Oliveira, "Automation and control in greenhouses: State-of-the-art and future trends," in *Proc. CONTROLO*. Cham, Switzerland: Springer, 2017, pp. 597–606.

[32] Z. Li, R. Guo, M. Li, Y. Chen, and G. Li, "A review of computer vision technologies for plant phenotyping," *Comput. Electron. Agricult.*, vol. 176, Sep. 2020, Art. no. 105672.

[33] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer, and D. Rueckert, "Joint learning of motion estimation and segmentation for cardiac MR image sequences," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 472–480.

[34] V. Madhu Babu, K. Das, A. Majumdar, and S. Kumar, "UnDEMoN: Unsupervised deep network for depth and ego-motion estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1082–1088.

[35] J.-T. Hsieh, B. Liu, D.-A. Huang, L. Fei-Fei, and J. Carlos Niebles, "Learning to decompose and disentangle representations for video prediction," 2018, *arXiv:1806.04166*.

[36] J. S. Amthor, "Effects of atmospheric CO_2 concentration on wheat yield: Review of results from experiments using various approaches to control CO_2 concentration," *Field Crops Res.*, vol. 73, no. 1, pp. 1–34, Oct. 2001.

[37] F. N. Tubiello and F. Ewert, "Simulating the effects of elevated CO_2 on crops: Approaches and applications for climate change," *Eur. J. Agronomy*, vol. 18, nos. 1–2, pp. 57–74, Dec. 2002.

[38] D. Shadrin, A. Menshchikov, D. Ermilov, and A. Somov, "Designing future precision agriculture: Detection of seeds germination using artificial intelligence on a low-power embedded system," *IEEE Sensors J.*, vol. 19, no. 23, pp. 11573–11582, Dec. 2019.

[39] C. Lu, M. Hirsch, and B. Scholkopf, "Flexible spatio-temporal networks for video prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6523–6531.

[40] H. Yu, S. Sun, H. Yu, X. Chen, H. Shi, T. S. Huang, and T. Chen, "FOAL: Fast online adaptive learning for cardiac motion estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4313–4323.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[42] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.

JOO-YEON JUNG received the B.S. degree from the Department of Computer Science, Sookmyung Women's University, Seoul, South Korea, in 2021. She is currently pursuing the M.S. degree in electrical engineering with Korea University, Seoul. Her current research interests include image processing, image prediction, and deep learning.

SANG-HO LEE received the B.S. degree from the School of Electrical Engineering, Korea University, Seoul, South Korea, in 2015, where he is currently pursuing the integrated M.S. and Ph.D. degree in electrical engineering. His current research interests include image generation, color constancy, and visible light communication.

TAE-HYEON KIM received the B.S. degree from the School of Electrical Engineering, Sejong University, Seoul, South Korea, in 2021. He is currently pursuing the M.S. degree in electrical engineering with Korea University, Seoul. His current research interests include image generation, image prediction, and deep learning.

MYUNG-MIN OH received the B.S. and M.S. degrees in horticulture from Seoul National University, Seoul, Republic of Korea, in 1999 and 2003, respectively, and the Ph.D. degree in horticulture from Kansas State University, Manhattan, KS, USA, in 2008. From 2008 to 2009, he was a Postdoctoral Researcher with Kansas State University. From 2009 to 2010, he was a Postdoctoral Researcher with the Children's Nutrition Research Center, Baylor College of Medicine, and the United States Department of Agriculture, Houston, TX, USA. In 2010, he started to serve as an Assistant Professor for the Horticultural Science Department, Chungbuk National University, Cheongju, Republic of Korea, where he has been a Professor, since 2018. He is the author of three books, more than 80 articles, and more than ten patents. His research interests include environmental control of plant growth in greenhouses and vertical farms. He is an Associate Editor of the journal *Horticulture, Environment*, and *Biotechnology*.

JONG-OK KIM (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Korea University, Seoul, South Korea, in 1994 and 2000, respectively, and the Ph.D. degree in information networking from Osaka University, Osaka, Japan, in 2006. From 1995 to 1998, he served as an Officer for the Korea Air Force. From 2000 to 2003, he was with SK Telecom Research and Development Center and Mcubeworks Inc., South Korea, where he was involved in the research and development on mobile multimedia systems. From 2006 to 2009, he was a Researcher with the Advanced Telecommunication Research Institute International (ATR), Kyoto, Japan. He joined Korea University, in 2009, where he is currently a Professor. His current research interests include image processing, computer vision, and intelligent media systems. He was a recipient of the Japanese Government Scholarship, from 2003 to 2006.