IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 22, 2022, accepted March 30, 2022, date of publication April 6, 2022, date of current version April 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165211

A Novel Shape Based Plant Growth Prediction Algorithm Using Deep Learning and Spatial Transformation

TAEHYEON KIM, SANG-HO LEE, AND JONG-OK KIM, (Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

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

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (2020R1A4A4079705).

ABSTRACT Plant growth prediction is challenging, as the growth rate varies depending on environmental factors. It is an essential task for efficient cultivation in controlled environments, such as in plant factories. In this paper, we propose a novel deep learning network for predicting future plant images from a number of past and current images. In particular, our focus is on the estimation of leaf shape in a plant, because the amount of plant growth is commonly quantified based on the leaf area. A spatial transform is applied to a sequence of plant images within the network, and the growth behavior is measured using a set of affine transform parameters. Instead of conventional sequential image fusion, the affine transform parameters for all pairs of successive images are fused together to predict the shape of the future plant image. Then, an RGB reconstruction subnet divides the plants into multiple patches to make global and local growth predictions based on hierarchical auto-encoders. A variety of experimental results show that the proposed network is robust to dynamic plant movements and can accurately predict the shapes of future plant images.

INDEX TERMS Plant growth prediction, sequential image, shape estimation, spatial transformer network, hierarchical network.

I. INTRODUCTION

In the agricultural industry, plants have recently been cultivated in closed environments such as plant factories, where the light, humidity, temperature, and CO2 concentration are controlled to improve the plant harvest. In this closed plant system, it is important to understand how the environmental conditions affect plant growth, so as to provide efficient plant cultivation [1]–[5]. Plant growth models for effectively managing plant cultivation have been studied in the field of agricultural research [6], [7]. However, plant growth prediction is challenging, as the growth rate varies greatly depending on environmental factors. Nevertheless, it is an essential task for efficient cultivation in controlled environments.

This study attempts to predict the future dynamic growth behaviors of plants from a sequence of plant images via deep learning. Given a time-series of past and current plant images, as illustrated in Fig. 1, we aim to predict a future plant image in by concentrating on the overall shapes of leaves. As a plant grows over time, the area for the leaves gradually increases, and their shapes change geometrically. Estimating the complex motions of the leaves is essential for growth predictions, and conceptually, is a similar problem to the prediction of future video frames in the field of computer vision [8]–[10].

There are very few previous works in the literature specializing in plant growth prediction [11], [12]. They were mainly based on adopting an auto-encoder [13] with ConvLSTM [14] as the backbone of the network. And recently, the authors in [16] proposed a plant growth prediction method which adopts the spatial transformer network (STN) [15] in the U-Net with ConvLSTM structure. We observe that the overall shape of plant leaves is beneficial because the area and weight of leaves are popular quantitative factors to measure plant growth. Motivated by this observation, the task of plant growth prediction is divided into two processes of shape prediction and RGB reconstruction. For the shape prediction, we propose to leverage the spatial transform, and

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

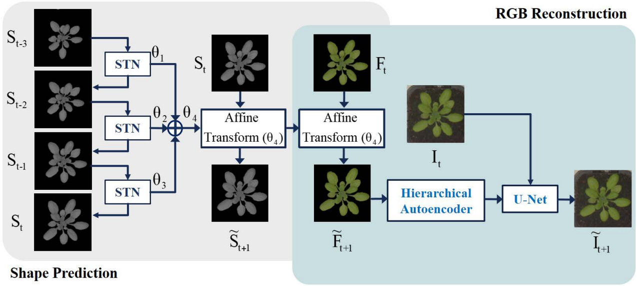

FIGURE 1. **Future plant image prediction on spatial transform parameter domain. *S* is the gray shape image, *F* is the RGB foreground image, and *I* is the RGB image.**

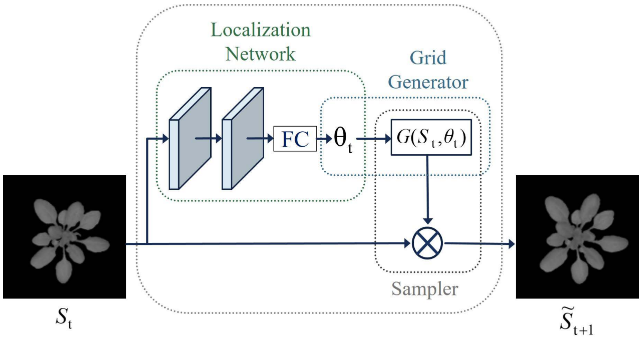

FIGURE 2. **Spatial transformer network on shape domain. Figure inspired by [15].**

this is inspired by STN [15], which has also been applied for future video frame generation [17]–[19].

In this paper, we propose a novel deep network for predicting the future of plant growth from a sequence of plant images. The network aims to generate a plant image at a future time, from which we can easily and quantitatively measure the degree of growth. The multiple leaves within a plant exhibit distinct shapes, sizes, and orientations, and can vary dynamically over time. Also, even though a sequence of plant images is captured at a fixed time interval, the growth rate of a plant may not be constant, and its growth behaviors (e.g., shape, and orientation) can be diverse in each plant sample. Thus, we need a sophisticated algorithm to address these challenges. In our framework, we attempt to model the growth behaviors of the leaves with the affine transform which comprises rotation, scaling and translation. To perform the affine transform in for convolutional neural networks (CNNs), we use the STN, which creates affine transform regression parameters. For each pair of adjacent images, we determine the affine transform parameter set (θ in Fig. 1) which indicates the quantification of plant growth. Then, we estimate an affine transform parameter set by combining past parameters to generate a future plant image as illustrated in Fig. 1. In other words, the shape prediction is made on transform parameter domain. As the rate of growth of leaves varies,the sizes of leaves change as they continue to grow. To accurately predict plant growth, it is necessary to accurately predict the growth of both small and large leaves. As a result, the RGB reconstruction subnet divides the plant into several patches, and predicts local growth with hierarchical auto-encoders. A variety of experimental results show that the proposed network is robust to dynamic plant movement, and can accurately predict the shape of the future plant image.

II. RELATED WORKS

A fundamental limitation of CNN is the lack of spatial invariance to the input data. It is difficult to cope with various spatial variabilities with 2 × 2 pixel unit operation. The spatial transformer network (STN) [15] was proposed to learn invariance to spatial changes such as translation, scale, and rotation. And, it has been popularly applied to motion estimation [17], [19] and future frame prediction [18], [20], [21].

One study [17] proposed a dual adversarial training method in which the network is largely divided into two branches. The first predicts a future frame directly while the other does a future flow. The outputs of the two branches are fused to predict a future frame. In [18], a future frame was predicted in the transformation space. The affine transform for the frame prediction was learned by fusing a sequence of consecutive affine transform parameters from the input video. The network generates the affine transform parameters for a future frame against the affine inputs. Although this seems to be similar to our work, they are quite different in terms of network architecture, separate estimations of shape and content, and the network input and output.

As an example of video prediction studies without the STN, [22] proposed a network for predicting future frames by separating motion and content information into two branches. [23] proposed a network comprising of two processes: pose estimation and image generation. After estimating the poses of the input images, a future frame was predicted on the pose domain, and was then reconstructed through image generation.

In recent years, there have been studied a few works to predict a future plant image using deep learning networks. Research on plant growth prediction is at an initial stage and its approach primarily comes from future video generation. The conventional works of plant growth prediction commonly employ an auto-encoder structure combined with ConvLSTM [11], [12] which has been already studied for video prediction [22], [23], multiple auto-encoders (correspond to multi-inputs) are fused through ConvLSTM. The network accepts label images (the segmentation of leaves) as well as plant RGB ones, and it generates both the label and RGB images in the future. Reference [12] adds GAN to an auto-encoder with ConvLSTM to ensure image generation. Also, by extending [11], multiple auto-encoders are hierarchically fused in 1/2 and 1/8 resolutions through ConvLSTM. The LSTM [24] based fusion has also been replaced

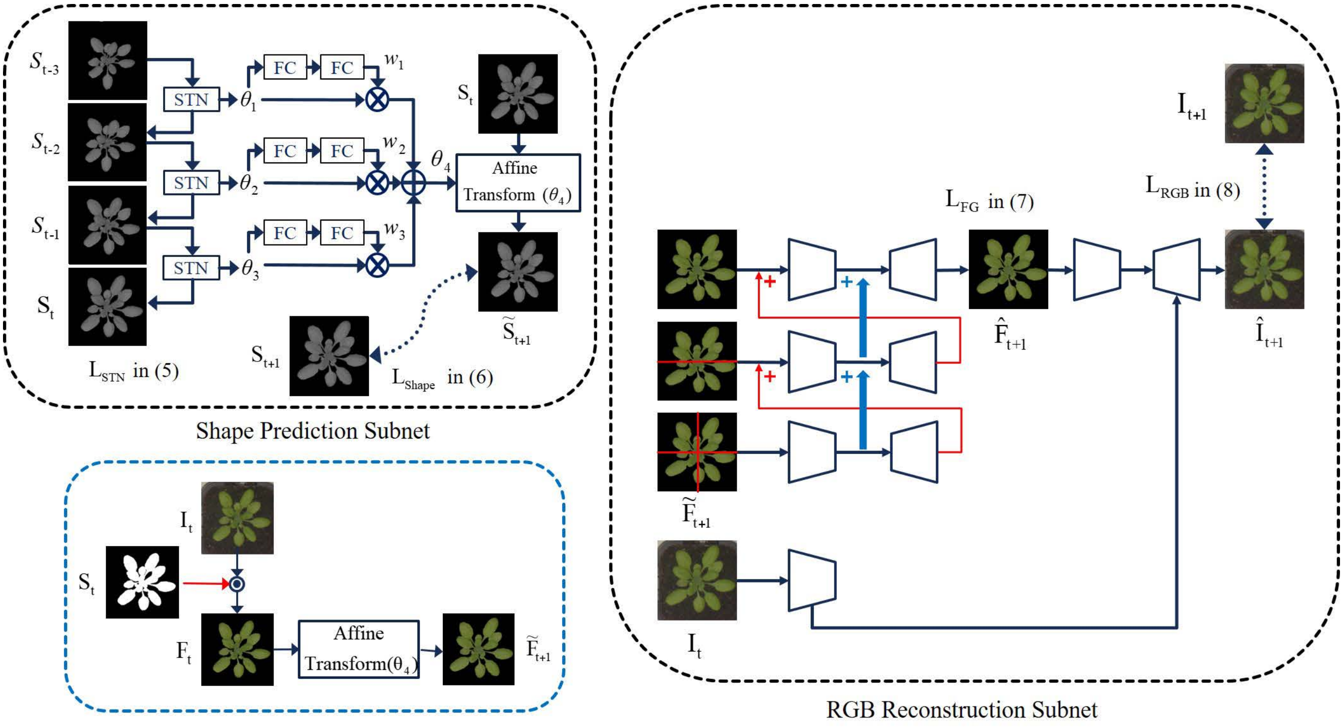

FIGURE 3. Overall architecture of the proposed network for plant growth prediction.

with a simple concatenation of CNN-based channels. In [16], at first, time series gray images of plants are globally matched through STN, and then shape prediction is performed through U-Net with ConvLSTM. After that, the final prediction result is obtained by combining the RGB image.

Unlike the conventional works for plant growth prediction, the proposed network first predicts the shape of a future plant image on spatial transform domain and then, an RGB image is reconstructed from the shape. Also, the existing methods were evaluated with a small dataset with a short time interval, but we conducted vast experiments with three different datasets with distinct plants.

III. SPATIAL TRANSFORMER NETWORK

A spatial transformer module was introduced to provide spatial transformation capabilities with a neural network architecture [15]. It consists of three parts: localization net, grid generator, and sampler as illustrated in Fig. 2. The affine transformation matrix denoted as θ_t in Fig. 2 is estimated from the t-th shape frame S_t by the localization net which is composed of convolution and fully connected layers, and θ_t is given as follows:

$$\theta_t = f_{local}(S_t) = \begin{bmatrix} \theta_{t1} & \theta_{t2} & \theta_{t3} \\ \theta_{t4} & \theta_{t5} & \theta_{t6} \end{bmatrix} \tag{1}$$

According to the parameters from the localization net, the grid generator calculates a sampling grid that determines the points to be sampled on the input feature map.

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = \theta_t \begin{pmatrix} x_i^{t+1} \\ y_i^{t+1} \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{t1} & \theta_{t2} & \theta_{t3} \\ \theta_{t4} & \theta_{t5} & \theta_{t6} \end{bmatrix} \begin{pmatrix} x_i^{t+1} \\ y_i^{t+1} \\ 1 \end{pmatrix} \tag{2}$$

In the above, (x_i^{t+1}, y_i^{t+1}) are the target coordinates of the regular grid in the $(t+1)$-th shape frame, (x_i^t, y_i^t) are the source coordinates in the t-th frame for defining the sample points, and θ_t is the affine transformation matrix. Finally, the sampler applies bilinear sampling to the input t-th frame to create an output $(t+1)$-th frame by using the sampling grid. Further details are provided in [15].

IV. THE PROPOSED METHOD

In this section, we describe the overall architecture of the proposed deep network for plant growth prediction. As illustrated in Fig. 3, it consists of two subnets for estimating the shape and RGB images of a plant in the future. Our work is primarily motivated by the development of technologies for controlling environmental factors (e.g., LED light) in closed cultivation environments, and ultimately aims to maximize plant harvest. To adaptively control these factors, we first attempt to identify the past and current growth behaviors, and to estimate the growth rate in the near future. Then, the factors are configured accordingly based on the sequential change in growth over time. For example, the wavelength and intensity of LED lights can be determined optimally based on the history of plant growth. Therefore, it is essential to accurately predict future plant growth, given the current environmental conditions.

A. PROPOSED NETWORK ARCHITECTURE

In this paper, we propose a novel deep learning algorithm for predicting a future plant image from a number of time-series plant images at past and present times which are captured at a top view. In particular, our focus is placed on the estimation of leaf shape in a plant because the amount of plant growth is commonly quantified by the leaf area. For this reason, we designed a shape-based growth prediction network, as illustrated in Fig. 3.

From the estimation, the degree of growth can be quantitatively measured. The shape difference between adjacent plant images corresponds to the amount of growth, and can be described by the spatial transform parameters (rotation, scaling and translation) when the growth is modeled by the STN. The STN is applied as a single transformation to the whole leaves of a plant, as the overlap between leaves makes it difficult to segment whole leaf boundaries.

Our work is highly inspired by motion estimation and video frame generation, which have been widely studied in the field of computer vision. Although both are quite similar to each other, the problem of plant growth prediction is specifically characterized by the certain aspects. First, the time interval between neighboring plant images is significantly longer than the conventional video prediction task. Whereas a video frame interval is typically within tens of milliseconds, the interval between plant images ranges from several hours to even a day. Thus, leaf movements are more dynamic. Second, a single plant contains several leaves whose shapes, sizes, and orientations change over time. As a plant grows, neighboring leaves overlap, and the orientation of a leaf can randomly vary. These types of motion are distinct from those of objects frequently observed in natural videos. Finally, it is difficult to find a motion vector in the plant images, because all leaves are similar to each other in terms of color, texture, and shape. There are far fewer feature points than common natural images, and it would be undesirable to apply the existing motion estimation techniques (such as optical flow) to plant images. This motivated us to consider spatial transform-based alignment in this paper.

Recently, spatial transform (e.g., affine transform) has been implemented within a neural network. A spatial transform consists of scaling, rotation, translation, and non-rigid deformation. The transformation is performed globally done on the entire image, unlike the conventional feature map. More specifically, the spatial transform is applied to a pair of successive plant images at different times, and find a set of parameters, θ (which collects several affine transform parameters in a matrix form actually) to describe the degree of growth. In other words, the amount of growth is quantified by a set of affine transform parameters. The parameter, θ is learned for every pair of neighboring images, and it is multiplied by its importance map as learned in the network. Then, the multiple θ's are combined to estimate the next spatial transform parameter for the output. Next, the learned transform parameter is applied to a current image as shown in Fig. 3, and the result becomes a future shape image. The process of finding a set of affine transform parameters is important, because it determines the degree of growth prediction. In practical, there are several obstacles that can degrade the accuracy of image alignment. For example, if the transform is performed on RGB domain, the extent of growth detection is adversely affected by background signals, leading to inaccurate spatial transform. This is why we perform the transform on shape domain.

The learned transform parameter is applied to both the shape image and its RGB at the current time, thereby producing their predictions for the next time. A shape image is generated only for calculating the loss on shape domain. The transformed RGB image is close to the ground truth, but it still shows some incorrect estimations particularly in local regions. The leaves of a plant exhibit heterogeneous growth behaviors. In other words, each leaf in a plant may grow in diverse directions, and the growth rate can be different for each leaf. These heterogeneous shape changes and movements of the leaves make it challenging to predict a future plant image using only the STN.

The global plant growth is predicted using the shape estimation subnet. In contrast, the RGB reconstruction subnet focuses on the growth of the local leaves of the plant. Plants produce leaves of different sizes, and all leaves grow at different growth rates. Thus, we employ the RGB reconstruction subnet after the STN module to estimate the local growth of the leaves. In the RGB reconstruction subnet, a plant image is divided into multiple patches, these patches are passed through a hierarchical auto-encoder.

The input to the RGB reconstruction subnet is obtained by transforming the plant foreground image using the set of affine transform parameters already learned from the shape prediction subnet. The affine-transformed RGB foreground is then divided into four patches that are passed through the encoder. After encoding, the four feature maps are concatenated into two features, which are summed to the encoded features of the upper parent layer. In addition, the output (two subimages) of the current layer is summed to the input in the next parent layer. This process is repeated until the top layer. Finally, to create a complete plant image (including the background), the current plant image I_{t+1} is fused with the reconstructed plant foreground $\hat{F}_{t+1}$ to replenish color and image details sufficiently. The proposed method can predict both the large and small leaves of the plant.

B. LOSS FUNCTION

When the proposed network is trained, the L1 norm, also called the Least Absolute Deviation (LAD), is adopted as the loss function. With L1 norm, we can obtain the output RGB image which is less blurred than Mean Square Error (MSE)

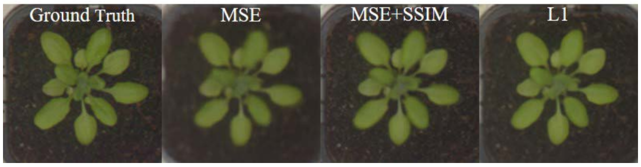

FIGURE 4. Performance comparison according to loss function (L1, MSE and SSIM).

and SSIM. Fig. 4 shows the result images for L1, MSE, and SSIM. The formula for the L1 norm is as follows:

$$L1(X, Y) = \sum_{i=1}^{p} |Y_i - X_i| \tag{3}$$

Here, X_i and Y_i are the i-th pixel values of the images X and Y, respectively, and p is the total number of pixels.

The overall loss function consists of three sub-losses, as follows.

$$L = L_{STN} + L_{Shape} + L_{FG} + L_{RGB} \tag{4}$$

The first term, L_{STN} in (4) is the loss measured in the STN module of the shape prediction subnet. For each pair of adjacent inputs as shown in Fig. 3, the STN is applied to quantify the growth during a time interval. The prediction performance of the STN is reflected to L_{STN}, and the L1 loss between the prediction of STN ($\widetilde{S}_{t-2}$, $\widetilde{S}_{t-1}$, $\widetilde{S}_t$ in Fig. 3) and its ground truth is calculated as

$$L_{STN} = \sum_{k=0}^{2} L1(\widetilde{S}_{t-k}, S_{t-k}) \tag{5}$$

The second term, L_{shape} in (4) is the loss measured for the output of the shape prediction subnet. It is the $L1$ loss between a ground truth shape and the shape prediction subnet output, and is determined as follows:

$$L_{Shape} = L_1(\widetilde{S}_{t+1}, S_{t+1}) \tag{6}$$

Next, L_{FG} in (4) is the loss measured for the output of the hierarchical RGB foreground reconstruction net. It is the $L1$ loss between the ground truth foreground and RGB foreground prediction subnet output, and is determined as follows:

$$L_{FG} = L1(\hat{F}_{t+1}, S_{t+1} \odot I_{t+1}) \tag{7}$$

In the above, $\odot$ means the element-wise multiplication

Finally, L_{RGB} is the loss for the final RGB image and it is given as follows:

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

V. EXPERIMENTAL RESULTS

Our network was implemented using the PyTorch framework on a PC with a NVIDIA RTX 3090 GPU. For loss optimization, we adopted an Adam optimizer with a batch size of eight [25]. The initial learning rate is 0.0001 and is divided by 8 for every 20,000 iterations.

FIGURE 5. Acquisition of our Butterhead dataset in a plant factory.

Three datasets were used to evaluate the performance of the proposed method and to compare it with the existing methods. The resolution of the image is 128 x 128. The training images were rotated by 90,180 and 270 degrees and were reversed left and right for data augmentation. First, The Aberystwyth leaf evaluation dataset [26] was used for experiments. It is composed of time-series image sequences of Arabidopsis Thaliana plants [27]. There are four sets of 20 Arabidopsis Thaliana plants which have been grown in trays. As the plants grow, the leaves are overlapped each other, allowing 10 out of 20 plants to be used in the experiment. For the experiments, the plant data are divided into a training (nine plants) and test (one plant) datasets. Each frame is taken in a 15-minute time lapse sequence. In order to observe the growth of plants more dynamically, the time interval between input images was increased to 1 days. A subset of these images have been hand-annotated to provide the ground truth of a plant.

Second, we conducted the evaluation with another dataset, named as Komatsuna [28]. The dataset consists of five Komatsuna plants, four among which are assigned to training, and one is to test. The images in the dataset are taken from a top view as in the previous dataset. Originally, the dataset has a time interval of one hour between adjacent plant images, but the time interval is changed to a three-hour interval to increase the amount of growth between input images. This makes plant growth prediction more challenging.

Finally our own dataset was created in this study for further evaluation, and it is named as Butterhead. Fairly butterhead lettuce [29] has grown in a plant factory for 13 days as shown in Fig. 5. It includes the image sequences of ten Butterheads with one day interval. For the experiments, the plant data are divided into a training (eight plants) and a test (two plants) datasets. For our network, the leaves are segmented from an RGB image to generate the shape image.

A. THE SHAPE PREDICTION SUBNET

The shape prediction subnet was evaluated to explore the extent to which it could learn plant growth using the STN.

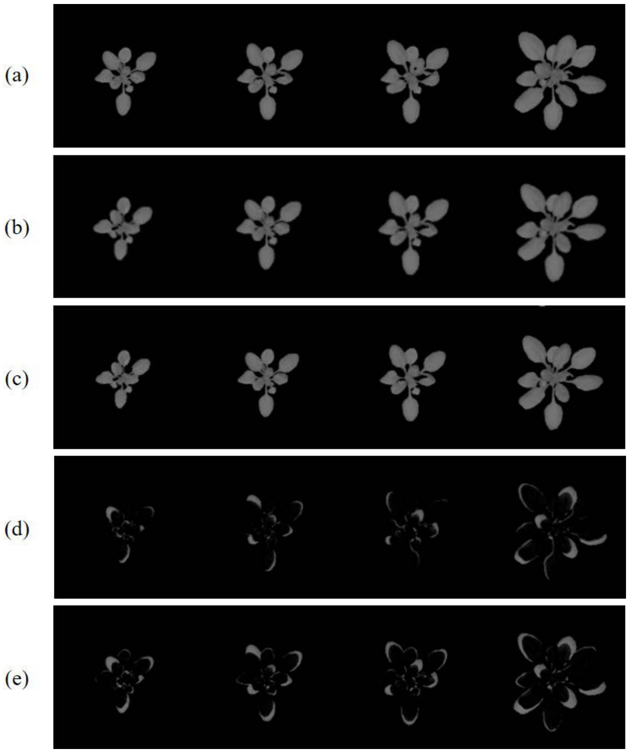

FIGURE 6. Prediction of shape images and their prediction errors. (a) ground truth S_{t+1}, (b) predicted shape $\widetilde{S}_{t+1}$ using STN, (c) the input S_t, (d) difference image between (a) and (b), (e) difference image between (a) and (c).

TABLE 1. Comparison of PSNR, SSIM and CS for the Aberystwyth leaf evaluation dataset [27].

Dataset	*Algorithm*	*PSNR*	*SSIM*	*CS*
Aberystwyth	LSTM [12]	30.29	0.8372	77.06
	Concat [12]	30.08	0.8283	78.08
	HP-Net [23]	30.31	0.8316	79.67
	MC-Net [22]	30.36	0.8348	79.50
	STN-LSTM [16]	30.55	0.8425	79.58
	Proposed	**30.61**	**0.8431**	**80.22**

It was trained using sequential shape image sequences with one-day intervals. It learned the affine transform parameters between the present and future plant shapes by fusing the transform parameters for the past pairs of shape images in the input sequence.

TABLE 2. Comparison of PSNR, SSIM and CS for the Komatsuna dataset [27].

Dataset	*Algorithm*	*PSNR*	*SSIM*	*CS*
Komatsuna	LSTM [12]	25.22	0.8945	78.19
	Concat [12]	25.35	0.8937	78.01
	HP-Net [23]	24.66	0.8904	80.11
	MC-Net [22]	25.02	0.8995	80.51
	STN-LSTM [16]	25.95	0.9042	80.73
	Proposed	**26.55**	**0.9065**	**81.28**

The generated parameters at the shape prediction subnet were used to predict a future RGB image from the current.

TABLE 3. Comparison of PSNR, SSIM and CS for our Butterhead dataset.

Dataset	*Algorithm*	*PSNR*	*SSIM*	*CS*
Butterhead	LSTM [12]	22.04	0.7910	80.18
	Concat [12]	21.61	0.7607	78.23
	HP-Net [23]	22.34	0.7769	79.77
	MC-Net [22]	22.56	0.7830	80.09
	STN-LSTM [16]	22.95	0.7862	80.55
	Proposed	**23.03**	**0.8154**	**81.66**

Fig. 6 (b) shows the shape images generated by the shape prediction subnet. Fig. 6 (d) shows the difference between Fig. 6 (b) and its ground truth, Fig. 6 (a), whereas Fig. 6 (e) shows the difference between the current shape, Fig. 6 (c) and Fig. 6 (a). The former represents the estimation error of the network, and the latter indicates the amount of plant growth from the current time t to the future $t + 1$. Notably, Fig. 6 (b) shows the future shape image as predicted from the current image in Fig. 6 (c).

As can be observed, the prediction of the proposed subnet is much closer to ground truth than the current. Even though the proposed subnet does not predict the future shape perfectly, its prediction is located near to the future. In particular, the errors between large leaves decrease significantly when Fig. 6 (d) is compared to Fig. 6 (e). It means that the proposed shape subnet can predict the overall shape of a plant accurately.

B. RGB PLANT IMAGE RESULTS

The quantitative and qualitative performance comparisons were made using the RGB plant images. In addition, the plant growth predictions were evaluated through existing video prediction networks [22], [23] and plant growth methods [12], [16]. Video predictions are conceptually similar to plant growth predictions, in that they predict object motion. LSTM [12], Concat [12], MC-Net [22], HP-Net [23] and [16] were evaluated for the future plant image generation, and Fig. 7 shows the experimental results. Fig. 7 (g) visually demonstrates the effectiveness of our proposed method, relative to the existing methods in Fig. 7 (b) [12] and Fig. 7 (c) [12]. [12] introduces two methods, which are similar to each other. The only difference is how to connect encoders and decoders. One uses concatenation, and the other does LSTM. Fig. 7 (d) [23] and (e) [22] show the results of HP-Net and MC-Net for video prediction, respectively. Video frame generation is very similar to plant growth predictions in that the motion of an object is predicted over time. In other words, the exercise of objects corresponds to the growth of plants in our work. Fig. 7 (f) shows the results of the recently proposed state-of-the-art plant growth prediction method. The left column of Fig. 7 shows RGB images at different times, and the right column shows the ground truth shapes and shape prediction errors. Fig. 7 (b-g) in the right column are obtained by masking the predicted plant image with its ground truth shape. The green regions in the masked images indicate

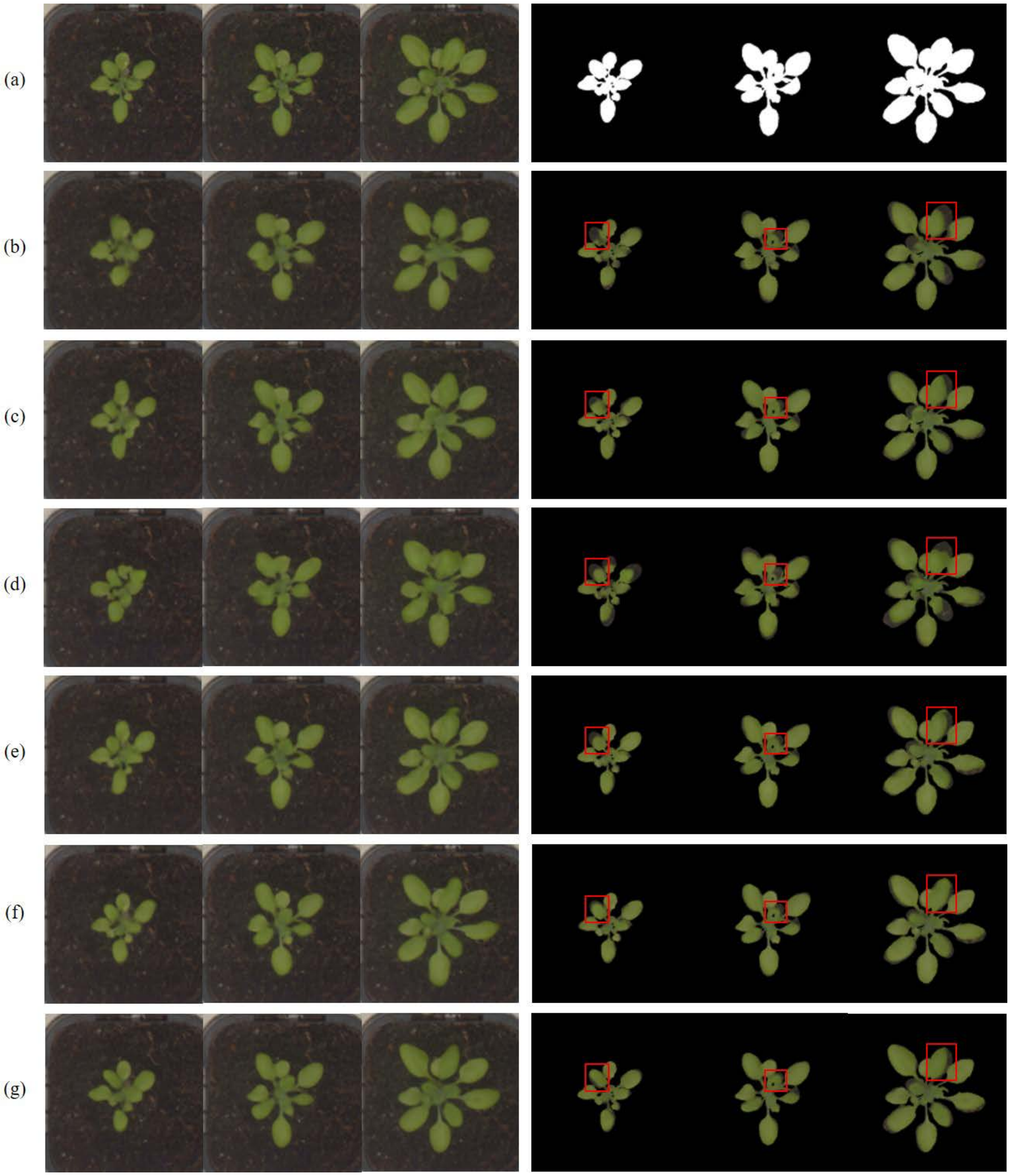

FIGURE 7. Prediction of RGB plant images. (a) ground truth I_{t+1}, (b) LSTM [12], (c) Concat [12], (d) HP-Net [23], (e) MC-Net [22], (f) STN-LSTM [16] and (g) the proposed.

correct prediction while the strong gray regions (originally the color of the background soil in the left column) represent the prediction errors.

TABLE 4. Coverage scores of the shape images in Fig. 11.

	Fig. 11 (a)	Fig. 11 (b)	Fig. 11 (c)	Fig. 11 (d)
CS	80.62	80.71	80.86	**80.91**

As shown in both Fig. 7 (b) and (c), the conventional plant growth methods often fail to correctly predict the shapes of the leaves, and some shape distortions often occur. We observe the distorted shape in Fig. 7 (b) and the disappearance in (c) of the leaf in the red box of the first plant. The conventional video prediction methods, Fig. 7 (d), and (e), show the distorted shapes of the leaves as compared to the ground truth I_{t+1}. It can be observed that the techniques in video prediction are not appropriate for plant growth predictions because they focus only on future frame predictions without preserving the shape of plant leaves. Fig. 7 (f) predicts better than the existing methods, but the prediction performance is

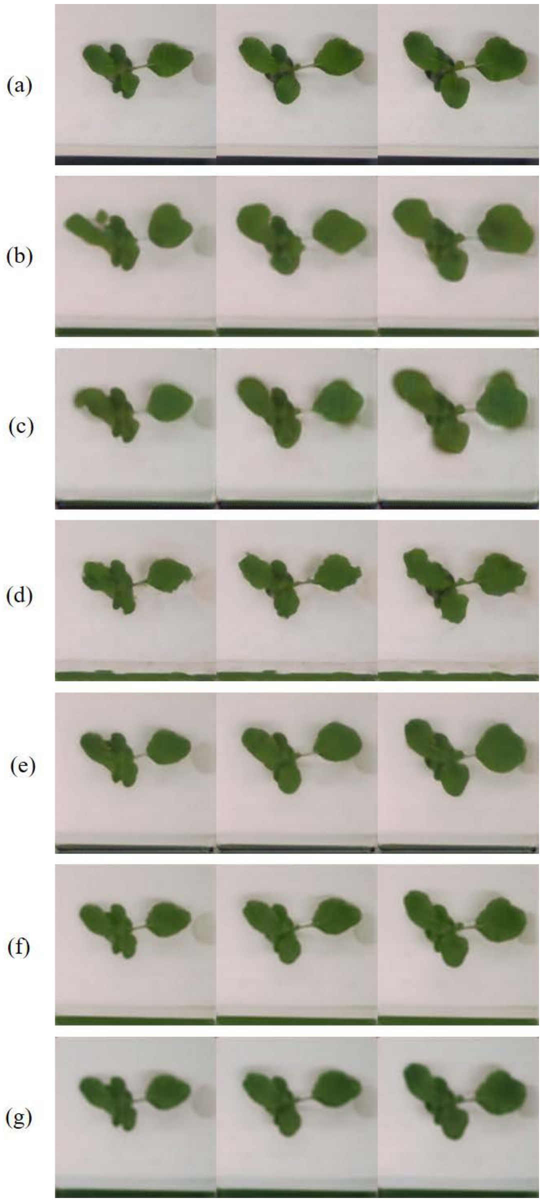

FIGURE 8. Prediction of RGB plant images on the Komatsuna dataset. (a) ground truth I_{t+1}, (b) LSTM [12], (c) Concat [12], (d) HP-Net [23], (e) MC-Net [22], (f) STN-LSTM [16] and (g) the proposed.

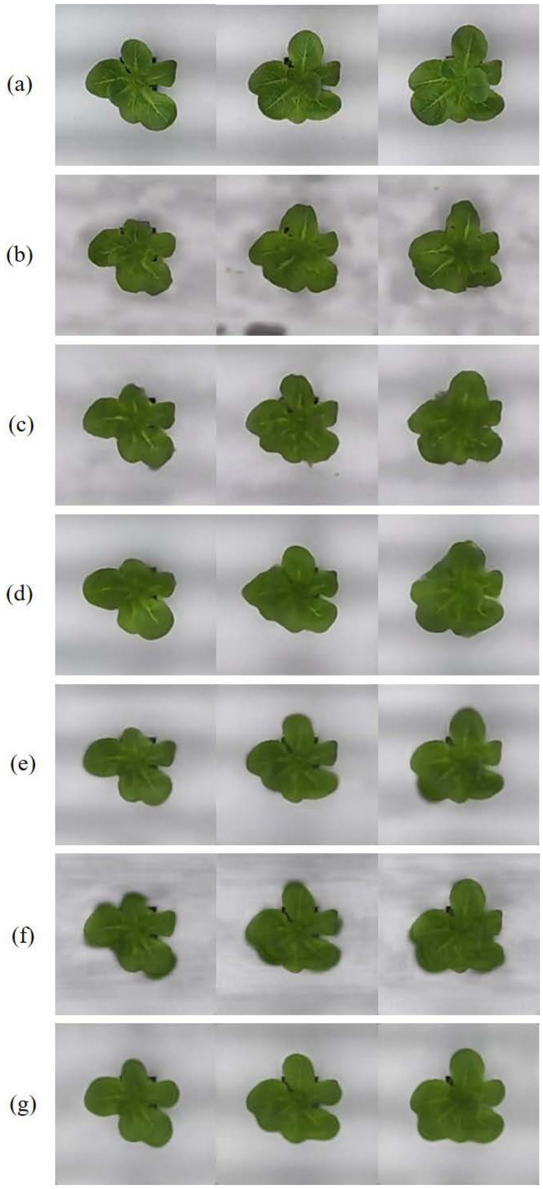

FIGURE 9. Prediction of RGB plant images on our Butterhead dataset. (a) ground truth I_{t+1}, (b) LSTM [12], (c) Concat [12], (d) HP-Net [23], (e) MC-Net [22], (f) STN-LSTM [16] and (g) the proposed.

inferior to that of the proposed method only for certain leaves with large growth changes. In contrast, the proposed method successfully estimates its shape in Fig. 7 (g), compared to the ground truth in Fig. 7 (a). In addition, if the capability of growth tracking is compared, the conventional methods Fig. 7 (b) and (c) fail to track the leaf movement in the red box of the third plant from the left, and the leaf still remains at the current time.

This failure to track the leaf movement produces large errors on the right side of the leaf. However, the proposed method consistently tracks the leaf movements, resulting in fewer errors in the shape prediction, as confirmed in Fig. 7 (g). Table 1 presents the quantitative results with the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [30] for 31 RGB plant images. And the coverage score (CS) [11] is used, which measures the accuracy of shape prediction. The CS is calculated as the ratio of the intersection to the union between two shapes. The CS for two shape images, S_{t+1}, $\widetilde{S}_{t+1}$ as follows:

$$CS(S_{t+1}, \widetilde{S}_{t+1}) = Overlap(S_{t+1}, \widetilde{S}_{t+1}) \tag{9}$$

Here, $Overlap(\cdot)$ is the intersection over union (IoU) between the inputs. As listed in Table 4, the proposed network achieves the highest CS value. As listed in Table 1, the

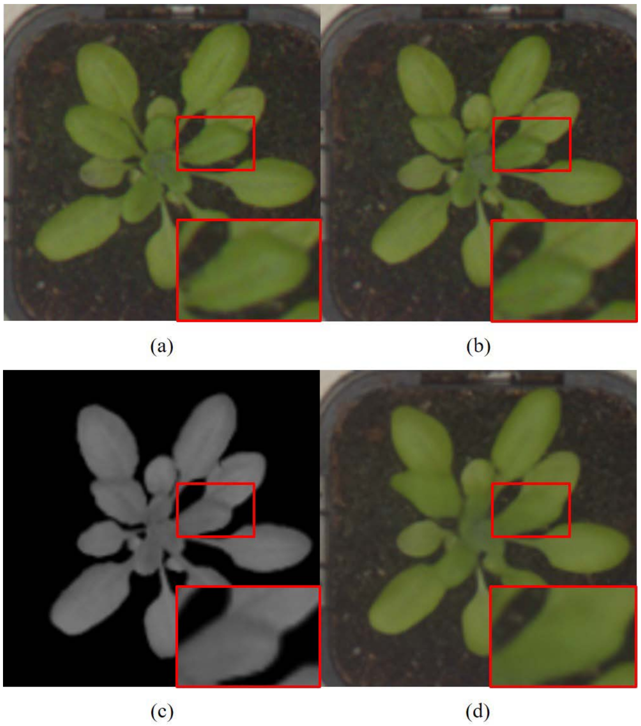

FIGURE 10. Comparison between 'before auto-encoder' and 'after auto-encoder'. (a) ground truth I_{t+1}, (b) the RGB input I_t, (c) the output of the shape prediction subnet $\widetilde{S}_{t+1}$, (d) the final output of the proposed network $\hat{I}_{t+1}$.

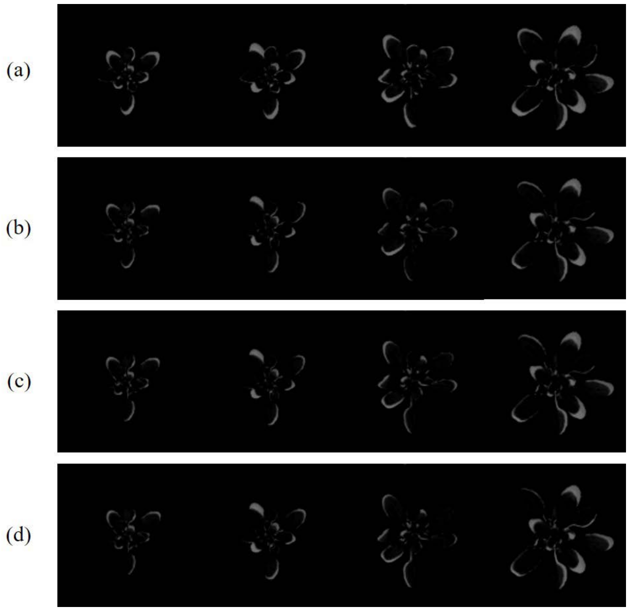

FIGURE 11. Impact of the number of the inputs on the shape prediction performance. (a) three inputs three inputs (I_t, I_{t-2} and I_{t-3}), (b) one input I_t, (c) three inputs (I_t, I_{t-1} and I_{t-2}), (d) four inputs (I_t, I_{t-1}, I_{t-2} and I_{t-3}).

proposed method achieves the best quantitative quality as expected from Fig. 7.

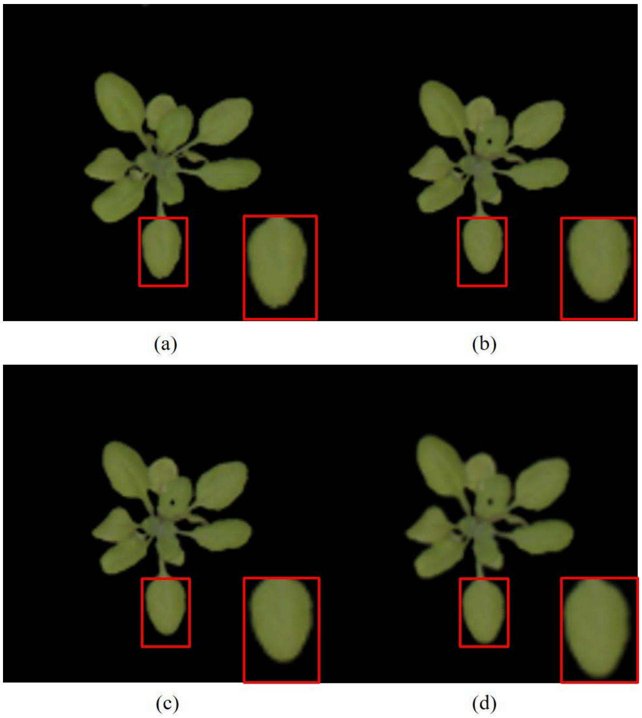

FIGURE 12. STN performance comparison between shape and RGB domains. (a) ground truth I_{t+1}, (b) the RGB input I_t, the predicted image $\widetilde{F}_{t+1}$ on (c) RGB domain and (d) Gray domain.

Fig. 8 shows the experimental results for the Komatsuna dataset. Fig. 8 (b) shows the distorted shapes of the leaves because of the incorrect prediction of their growth direction. In Fig. 8 (c), certain parts of the leaf are overgrown and the edges of the leaf are blurred. The leaf shape is not smooth and severely distorted in Fig. 8 (d). It is observed that plant growth prediction result in Fig. 8 (e) is better than other conventional methods, but the right leaf of the plant was predicted to be larger than the ground truth. FIg. 8 (f), as in Fig. 7, shows slightly inferior prediction performance of leaves with large growth changes. However, the proposed achieves better growth prediction, and the shapes of leaves are well predicted.

Experimental results using our own Butterhead dataset are shown in Fig. 9. As shown in Fig. 9 (b) and (c), the overall shapes of the leaf are very distorted for the existing methods. In addition, Fig. 9 (b) shows the distortion of the background is particularly severe. In the second and third images of Fig. 9 (d), there are some leaves which are not found in the ground truth. Compared to the conventional methods, Fig. 9 (e) was quite predicted well in terms of the overall leaf shape. However, some leaves are predicted to be under-grown and certain leaves are blurred, as shown in the third column of Fig. 9 (e). In Fig. 9 (f), blurring appears at the boundary between the plant and the background, and it can be also observed in some leaves. As confirmed in Fig. 9 (g), the proposed method is generally well-predicted in the overall shape of leaves, and certain leaves that are difficult to predict are also well-predicted. Tables 2 and 3 present the quantitative results of PSNR and SSIM for the Komatsuna and Butterhead datasets, respectively, and the proposed method achieves better quantitative performance.

C. ABLATION STUDIES

In the proposed network, the shape subnet generates spatial transform parameters, which are applied to the present RGB image. The result image corresponds to the future prediction of the present one. Then, it passes through the RGB reconstruction subnet for further enhancement. The output of the shape subnet still lacks full growth prediction. For example, if the final output of the shape prediction subnet, $\widetilde{S}_{t+1}$ is compared to $\hat{I}_{t+1}$ in the red boxes of Fig. 10 (c) and (d), the leaf size is quite different. The STN is capable of predicting the overall structure of leaves in a plant very well, but it sometimes fails to predict the movement of small local leaves. This can be further improved by the auto-encoder as shown in Fig. 10.

Next, we study the impact of the number of the inputs on the performance in the shape prediction subnet. Fig. 11 shows the difference between a ground truth and the predicted shape image. Notably, the number of inputs is fixed to four in the proposed network. As expected, it can be observed from Fig. 11 that additional inputs lead to fewer errors. Fig. 11 (a) and (c) show the results for the three inputs, however their timestamps are different. In Fig. 11 (a), the time interval between the inputs is not uniform, and it shows the largest error, worse than even a single input, (b). Although the network accepts four inputs for simplicity in this paper, it can be easily extended to more inputs. We also studied the performance differences between the shape and RGB domains. For the proposed network, training was conducted on the shape domain rather than RGB to find affine transform parameters for the future plant image prediction. Shape domain is a gray image domain in which only foreground of plants without background exists. The reason for training the network on shape domain is that we aimed to predict the plant growth, as measured by the overall area of leaves in a plant. Spatial Transformation on RGB domain tends to be easily affected by the surrounding environment such as soil, pot and other plants rather than a plant in the image. In addition, in the process of acquiring a top view plant image, even a marginal change in the plant position makes the dataset useless.

As shown in Fig. 12 (d), if training was conducted in the shape domain, the plant predicted for the future would be closer to the growth of the ground truth. In this study, this results was visualized by targeting a single leaf in the plant to clearly observe the difference in growth. As shown in the zoom-in leaf of Fig. 12, the size of the leaf on shape domain is predicted more accurately than that on RGB.

For the ablation study of the RGB reconstruction subnet, we conducted experiments on Arabidopsis Thaliana plants [27]. In the far left column of Fig. 13, there are RGB and shape images of ground truth. The second column shows the result of configuring the RGB reconstruction subnet with a single auto-encoder instead of hierarchical ones. We can observe the poor prediction performance for the local leaves (the blue box) and shape distortion (red box) in Fig. 13, relative to the proposed hierarchical auto-encoders (the fourth column). The third column shows the result of the proposed with background. As noted above, in the proposed method, the image masked with the foreground shape enters the RGB reconstruction subnet in the proposed method. Without foreground segmentation, the original plant image is used in this ablation study. As shown in the blue box, the proposed with background fails to track the growth. Thus, the proposed technique appears to effectively track the growth of the local leaves.

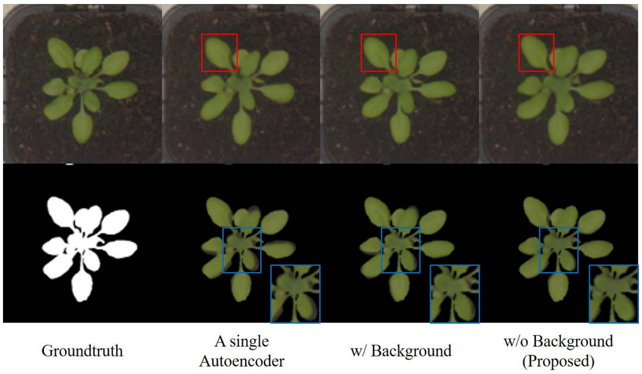

FIGURE 13. Ablation study results of the RGB reconstruction subnet on Arabidopsis dataset [27].

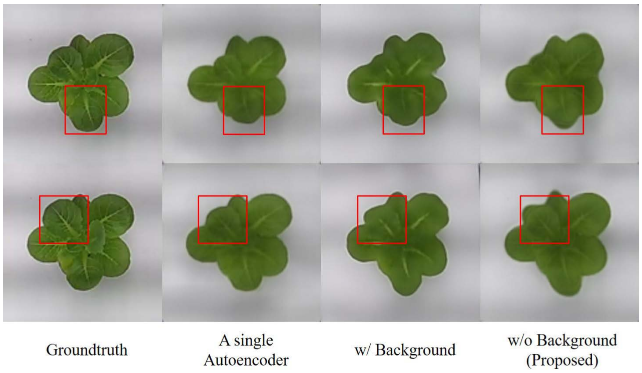

FIGURE 14. Ablation study results of the RGB reconstruction subnet on the Butterhead dataset.

Fig. 14 shows the ablation study results from the RGB reconstruction subnet on our Butterhead dataset. The ablation study scenarios of Fig. 14 are the same as those in Fig. 13, except for the dataset used. Each row in Fig. 14 shows show different plant images at different times. As observed in the red box, we can see that the proposed technique has grown closely to ground truth compared to the other two combinations. In the case of the proposed RGB reconstruction subnet, we can see that it contributes not only to the growth of local leaves but also to the growth of global leaves. Table 5 lists PSNR and SSIM values where the proposed RGB reconstruction subnet achieves a better quantitative performance.

TABLE 5. Ablation study quantitative results of the RGB reconstruction subnet with three datasets.

Dataset	Algorithm	PSNR	SSIM
Aberytwyth	A Single Auto-encoder	30.18	0.8349
	w/Background	30.55	0.8399
	w/o Background(proposed)	**30.61**	**0.8431**
Komatsuna	A Single Auto-encoder	25.77	0.9012
	w/Background	26.29	0.9053
	w/o Background(proposed)	**26.55**	**0.9065**
Butterhead	A Single Auto-encoder	21.81	0.7916
	w/Background	22.38	0.8019
	w/o Background(proposed)	**23.03**	**0.8154**

VI. CONCLUSION

In this study, we attempt to predict the dynamic growth behaviors of leaves in a plant via deep learning, and propose a deep network for predicting a future plant image from past and present images. The shape of a plant image is learned, and its RGB channels are reconstructed. Instead of the traditional sequential image fusion, our framework adopts the affine transform to model the growth behaviors of leaves. The affine transform parameters learned for all pairs of consecutive temporal images are fused together to predict the overall shape of the leaves. Then, we employ the RGB reconstruction subnet after the STN module to estimate the local growth of the leaves additionally. A plant image is divided into multiple patches, and they pass through a hierarchical auto-encoder. The RGB reconstruction subnet reconstructs an RGB image, with particular focus on the prediction of local movements and shape changes.

The proposed network is evaluated using a variety of datasets, including our own Butterhead dataset acquired from a plant factory. The experimental results show that the proposed network is particularly resistant to the dynamic movements of leaves, outperforming the existing methods in both qualitative and quantitative ways.

A one-day future image is predicted in this work, and the time interval can be increased to longer than a day in the future. In addition, we will establish additional datasets including various plant growth behaviors.

ACKNOWLEDGMENT

(Taehyeon Kim and Sang-Ho Lee contributed equally to this work.)

REFERENCES

[1] N. C. Yorio, G. D. Goins, H. R. Kagie, R. M. Wheeler, and J. C. Sager, "Improving spinach, radish, and lettuce growth under red light-emitting diodes (LEDs) with blue light supplementation," *HortScience*, vol. 36, no. 2, pp. 380–383, Apr. 2001.

[2] K.-H. Son and M.-M. Oh, "Growth, photosynthetic and antioxidant parameters of two lettuce cultivars as affected by red, green, and blue light-emitting diodes," *Horticulture, Environ., Biotechnol.*, vol. 56, no. 5, pp. 639–653, Oct. 2015.

[3] A. Amoozgar, A. Mohammadi, and M. R. Sabzalian, "Impact of light-emitting diode irradiation on photosynthesis, phytochemical composition and mineral element content of lettuce cv. Grizzly," *Photosynthetica*, vol. 55, no. 1, pp. 85–95, Mar. 2017.

[4] 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, Jun. 1999.

[5] R. J. Downs, *Environment and the Experimental Control of Plant Growth*, vol. 6. Amsterdam, The Netherlands: Elsevier, 2012.

[6] I. A. Lakhiar, J. Gao, T. N. Syed, F. A. Chandio, and N. A. Buttar, "Modern plant cultivation technologies in agriculture under controlled environment: A review on aeroponics," *J. Plant Interact.*, vol. 13, no. 1, pp. 338–352, Jan. 2018.

[7] G. Tamulaitis, P. Duchovskis, Z. Bliznikas, K. Breivé, R. Ulinskaite, A. Brazaityte, A. Novičkovas, and A. Žukauskas, "High-power lightemitting diode based facility for plant cultivation," *J. Phys. D, Appl. Phys.*, vol. 38, no. 17, p. 3182, 2005.

[8] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros, "A review on deep learning techniques for video prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 15, 2020, doi: 10.1109/TPAMI.2020.3045007.

[9] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[10] Y. Lu, K. M. Kumar, S. S. Nabavi, and Y. Wang, "Future frame prediction using convolutional VRNN for anomaly detection," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[11] S. Sakurai, H. Uchiyama, A. Shimada, and R.-I. Taniguchi, "Plant growth prediction using convolutional LSTM," in *Proc. VISIGRAPP*, 2019, pp. 105–113.

[12] T. Hamamoto, H. Uchiyama, A. Shimada, 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.*, 2020, pp. 153–161.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[14] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," 2015, *arXiv:1506.04214*.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015, *arXiv:1506.02025*.

[16] J.-Y. Jung, S.-H. Lee, T.-H. Kim, M.-M. Oh, and J.-O. Kim, "Shape based deep estimation of future plant images," *IEEE Access*, vol. 10, pp. 4763–4776, 2022.

[17] 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. (ICCV)*, Oct. 2017, pp. 1744–1752.

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

[19] Y. Wu, R. Gao, J. Park, and Q. Chen, "Future video synthesis with object motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2020, pp. 5539–5548.

[20] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," 2015, *arXiv:1511.06309*.

[21] 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.

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

[23] 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. (ICML)*, 2017, pp. 3560–3569.

[24] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

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

[26] J. Bell and H. M. Dee, "Aberystwyth leaf evaluation dataset," Nov. 2016.

[27] E. M. Meyerowitz, "Arabidopsis thaliana," *Annu. Rev. Genet.*, vol. 21, no. 1, pp. 93–111, 1987.

[28] 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.

[29] I. De Vries, "Origin and domestication of Lactuca sativa L," *Genetic Resour. Crop Evol.*, vol. 44, no. 2, pp. 165–174, 1997.

[30] 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.

TAEHYEON 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.

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.

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 with Korea Air Force. From 2000 to 2003, he was with the SK Telecom Research and Development Center and Mcubeworks Inc., South Korea, where he was involved in 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, Seoul, 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.