mathematics

MDPI

Article

Image Region Prediction from Thermal Videos Based on Image Prediction Generative Adversarial Network

Ganbayar Batchuluun, Ja Hyung Koo, Yu Hwan Kim and Kang Ryoung Park *

Division of Electronics and Electrical Engineering, Dongguk University, 30 Pildong-ro, 1-gil, Jung-gu, Seoul 04620, Korea; ganabata87@dongguk.edu (G.B.); koo6190@dongguk.edu (J.H.K.); taekkuon@naver.com (Y.H.K.)

* Correspondence: parkgr@dgu.edu

Abstract: Various studies have been conducted on object detection, tracking, and action recognition based on thermal images. However, errors occur during object detection, tracking, and action recognition when a moving object leaves the field of view (FOV) of a camera and part of the object becomes invisible. However, no studies have examined this issue so far. Therefore, this article proposes a method for widening the FOV of the current image by predicting images outside the FOV of the camera using the current image and previous sequential images. In the proposed method, the original one-channel thermal image is converted into a three-channel thermal image to perform image prediction using an image prediction generative adversarial network. When image prediction and object detection experiments were conducted using the marathon sub-dataset of the Boston University-thermal infrared video (BU-TIV) benchmark open dataset, we confirmed that the proposed method showed the higher accuracies of image prediction (structural similarity index measure (SSIM) of 0.9839) and object detection (F1 score (F1) of 0.882, accuracy (ACC) of 0.983, and intersection over union (IoU) of 0.791) than the state-of-the-art methods.

Keywords: image prediction; thermal videos; deep learning; generative adversarial network

Citation: Batchuluun, G.; Koo, J.H.; Kim, Y.H.; Park, K.R. Image Region Prediction from Thermal Videos Based on Image Prediction Generative Adversarial Network. *Mathematics* **2021**, *9*, 1053. https://doi.org/10.3390/math9091053

Academic Editor: Akemi Galvez Tomida

Received: 22 April 2021
Accepted: 4 May 2021
Published: 7 May 2021

1. Introduction

Various studies have been conducted on object detection [1–4], tracking [5–9], action recognition [10–12] using a camera-based video surveillance system in addition to depth, ego-motion, and optical flow estimation [13]. However, when a walking or running object leaves the field of view (FOV) of the camera, part of the object's body becomes invisible, which leads to a failure in human detection and tracking, thus inducing errors in action recognition. However, no studies have considered this issue so far. To solve this problem, this study conducted an experiment for the first time for predicting the region outside the FOV that is not included in the current image (t), as shown in the image (t'), which is illustrated in Figure 1, to restore the part of the object's body that is invisible.

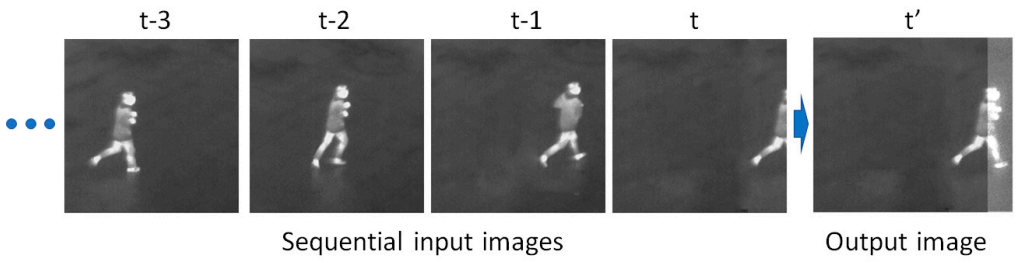


Figure 1. Example of thermal image prediction.

The proposed method widens the FOV of the current image using the current image, previous sequential images, and an image prediction generative adversarial network

(IPGAN)-based method. Furthermore, the original one-channel thermal image is converted into a three-channel thermal image to be used as an input in the IPGAN. In this study, various experiments were conducted using the marathon sub-dataset of the Boston University-thermal infrared video (BU-TIV) benchmark open dataset [14]. Existing studies related to the proposed method are explained in Section 2.

2. Related Works

The following studies attempted to predict the next image based on previous sequential images. In studies [15–19], image prediction methods were proposed for creating a future frame using a current frame and previous sequential frames. In [15], image prediction was performed using an encoder and decoder model based on long short-term memory (LSTM) and a 3D convolution layer. In [16], image prediction was performed using PhyDNet based on LSTM and the newly suggested PhyCell. In [17], image prediction was performed using LSTM and a convolutional neural network (CNN). In [18], image prediction was performed using an encoder and decoder model. In [19], image prediction was performed using a stochastic variational video prediction (SV2P) method.

Instead of predicting the current image based on previous sequential images, image inpainting methods were proposed in [20–24] where the deleted information is restored from a current image. In [20], image inpainting was performed using a fine deep-generative-model-based approach with a novel coherent semantic attention (CSA) layer. In [21], image inpainting was performed based on gated convolution and SN-PatchGAN. In [22], image inpainting was performed based on a parallel extended-decoder path for semantic inpainting network (PEPSI). In [23], image inpainting was performed using a context encoder method based on a channel-wise fully connected layer. In [24], image inpainting was performed using edge prediction and image completion based on the predicted edge map.

Furthermore, the following review and survey studies have been conducted. In a review paper [25], the datasets created between 2004 and 2019 that were used in image prediction were compared with the image prediction models created between 2014 and 2020. In a survey paper [26], papers and datasets based on image prediction were described. In another review paper [27], sequential-based, CNN-based, and generative adversarial network (GAN)-based image inpainting methods and the datasets used in image inpainting were described.

As explained, studies have been extensively conducted on image inpainting and the prediction of the next image based on previous sequential images. However, no study has examined an image prediction method for generating an image region outside the FOV, which is proposed in this article. In addition, no previous study on image prediction and image inpainting adopted thermal images. Table 1 presents a summary of the comparisons between the present and previous studies. This study is novel in the following four ways compared with the previous works:

- This study performed image prediction using thermal videos for the first time.
- This study designed an image prediction method that generates an image region outside the FOV for the first time.
- A new IPGAN for performing image prediction is proposed herein.
- The IPGAN model proposed herein is disclosed for a fair performance assessment [28] to other researchers.

The remainder of the paper is organized as follows. A detailed explanation of the proposed method is provided in Section 3. The experiment results and analysis with discussions are provided in Section 4. Finally, the discussion and the conclusion are presented in Sections 5 and 6.

Table 1. Comparison between the present and previous studies.

Category		Methods	Advantages	Disadvantages
Using visible-light images	Future image prediction	Encoder–decoder model [15,18], PhyDNet [16], CNN + LSTM [17], SV2P [19], and review and survey [25,26]	High performance of future image prediction based on a current frame and previous frames	Do not consider the image prediction out of FOV Do not use thermal image of low resolution and low image quality
	Image inpainting	CSA layer [20], gated convolution + SN-PatchGAN [21], PEPSI [22], context encoder [23], edge prediction and image completion [24], and review [27]	High performance of image inpainting based on a current frame	
Using thermal images	Image region prediction out of FOV	Three-channel thermal image and IPGAN (proposed method)	Consider the image prediction out of FOV Use thermal image of low resolution and low image quality	The predicted image out of FOV has a size limit

3. Materials and Methods

3.1. Overall Procedure of Proposed Method

In this section, the method proposed in this study is described in detail. The proposed method performs the image region prediction based on sequential three-channel thermal images using preprocessing, IPGAN, and postprocessing. In Sections 3.2–3.5. preprocessing, the IPGAN architecture, postprocessing, and the dataset for image prediction, respectively, are described in detail. Figure 2 shows the overall flowchart of the proposed method. The length of the sequential input images is 20 frames ($t-0, t-1, \ldots, t-19$), the size of each image is 85×170 pixels, and the size of the output image is 105×170 pixels. Specifically, the output image is created by combining a generated image region (an image outside the FOV) and the current image (an image inside the FOV).

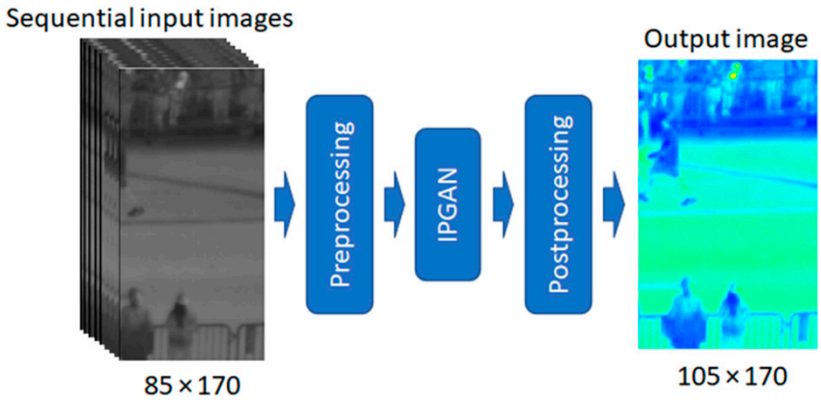


Figure 2. Overall flowchart of the proposed method.

3.2. Preprocessing

The preprocessing step is described in detail in this subsection. For thermal images captured with a thermal camera, a one-channel thermal image is converted into a three-channel thermal image using a colormap function. The jet colormap array [29] is used for performing color conversion. The jet colormap array is a mapping function that expresses heat in the most appropriate color compared with other colormaps. It maps a one-channel image into a three-channel image for 256 pixel values from 0 and 255. For example, the hottest part of a one-channel image has a pixel value of 255 (white), whereas the coldest part has a pixel value of 0 (black). Conversely, the pixel value of the hottest part of a three-channel (red, green, blue) image is [255,0,0] (red color), whereas that of the coldest

part is [0,0,255] (blue color). A color conversion example is shown in Figure 3. A one-channel thermal image is converted into a three-channel thermal image because several studies have shown that performing object detection, recognition, and classification using color visible light images results in a better performance than using grayscale visible light images [30–32]. Furthermore, for making the input and output sizes of the IPGAN structure identical, the region being predicted (the black area of 85 × 170 pixels) in the input image is created through the zero padding, thus changing the size of the input image from 85 × 170 pixels to 170 × 170 pixels.

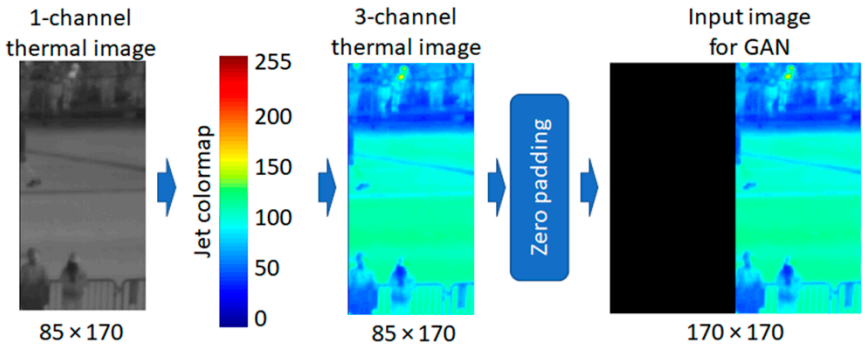


Figure 3. Procedure of preprocessing.

3.3. Proposed IPGAN Model

The three-channel image (170 × 170 pixels) obtained through preprocessing, as shown in Figure 3, is used as an input for the IPGAN proposed in this study. The structure of the IPGAN is illustrated in Figure 4. The generator shown in Figure 4 includes a concatenate layer (L1), convolution blocks (L2 and L7), residual blocks (L3–L5 and L8–L11), and convolution layers (L12 and L13) in that order. The discriminator includes convolution blocks (L1–L6) and a fully connected layer (L7) in that order.

In addition, the details of the IPGAN structure are presented in Tables 2–6. In Tables 2–4, the filter size, stride, and padding are (3 × 3), (1 × 1), and (1 × 1), respectively. In Table 2, two different numbers of filters, 128 and 64, are used for conv_block_1 and conv_block_2. In Table 5, the filter size, stride, and padding in conv_block_1–conv_block_3 are (3 × 3), (1 × 1), and (0 × 0), respectively, whereas in conv_block_4–conv_block_6, the filter size, stride, and padding are (3 × 3), (2 × 2), and (0 × 0), respectively. Prelu, lrelu, tanh, res_block, conv2d, add, conv_block, dense, and sigmoid represent the parametric rectified linear unit (relu), leaky relu, hyperbolic tangent activation function, residual block, two-dimensional convolution layer, addition operation, convolution block, fully connected layer, and sigmoid activation function, respectively. In Table 2, 20 sequential three-channel thermal images (170 × 170 × 3) are used as input as shown in Figure 4, whereas the output image is an image of size (170 × 170 × 3).

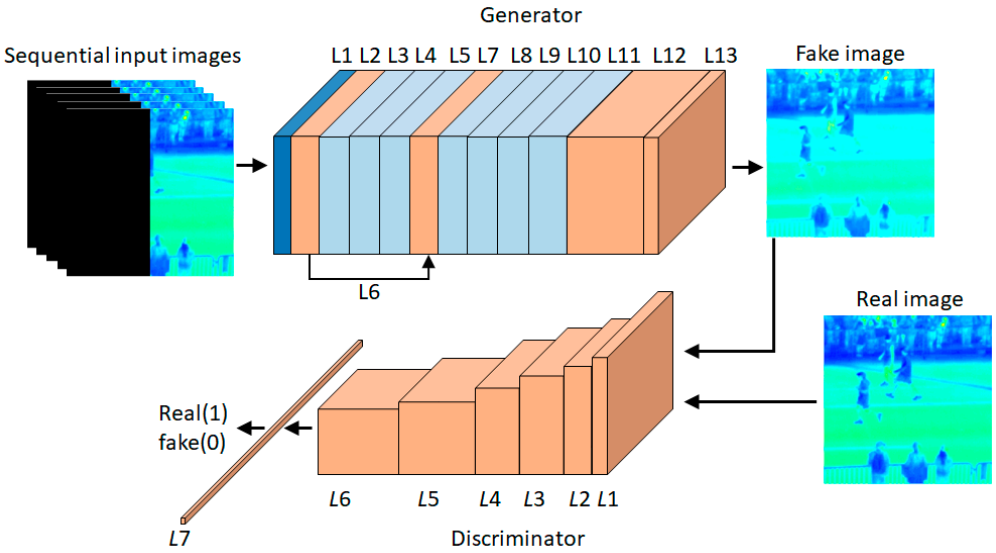


Figure 4. Example of the structure of the proposed IPGAN.

Table 2. Description of the generator of the proposed IPGAN.

Layer Number	Layer Type	Number of Filters	Number of Parameters	Layer Connection (Connected to)
0	input_layers_1–20	0	0	input_1–20
1	concat	0	0	input_layers_1–20
2	conv_block_1	128/64	143,232	concat
3	res_block_1	64	73,920	conv_block_1
4	res_block_2	64	73,920	res_block_1
5	res_block_3	64	73,920	res_block_2
6	add	64	0	res_block_3 & conv_block_1
7	conv_block_2	64	147,840	add
8	res_block_4	64	73,920	conv_block_2
9	res_block_5	64	73,920	res_block_4
10	res_block_6	64	73,920	res_block_5
11	res_block_7	64	73,920	res_block_6
12	conv2d_1	256	147,712	res_block_7
13	conv2d_2	3	6915	conv2d_1
14	tanh		0	conv2d_2
Total number of trainable parameters: 963,139				

Table 3. Description of a convolution block of the generator.

Layer Number	Layer Type	Number of Filters	Layer Connection (Connected to)
1	conv2d_1	128	input
2	prelu_1		conv2d_1
3	conv2d_2	64	prelu_1
4	prelu_2		conv2d_2

Table 4. Description of a residual block of the generator.

Layer Number	Layer Type	Number of Filters	Layer Connection (Connected to)
1	conv2d_1	64	input
2	prelu		conv2d_1
3	conv2d_2	64	prelu
4	add		conv2d_2 & input

Table 5. Description of the discriminator of the proposed IPGAN.

Layer Number	Layer Type	Number of Filters	Number of Parameters	Layer Connection (Connected to)
0	input layer	0	0	input
1	conv_block_1	32	896	input layer
2	conv_block_2	64	18,496	conv_block_1
3	conv_block_3	128	73,856	conv_block_2
4	conv_block_4	128	147,584	conv_block_3
5	conv_block_5	256	295,168	conv_block_4
6	conv_block_6	256	590,080	conv_block_5
7	dense		92,417	conv_block_6
8	sigmoid		0	dense
Total number of trainable parameters: 1,218,497				

Table 6. Description of a convolution block of the discriminator.

Layer Number	Layer Type	Layer Connection (Connected to)
1	conv2d	input
2	lrelu	conv2d

3.4. Postprocessing

During postprocessing, the final output is acquired from the RGB output image obtained using the IPGAN as shown in Figure 5. The region predicted in the output image obtained using the IPGAN is cropped as illustrated in Figure 5. The cropped region is combined with the original three-channel image (t − 0) to acquire the final output. The reasons for the smaller predicted region and the poor prediction of the remaining region are explained in Section 4.2 (ablation study) based on the experimental results.

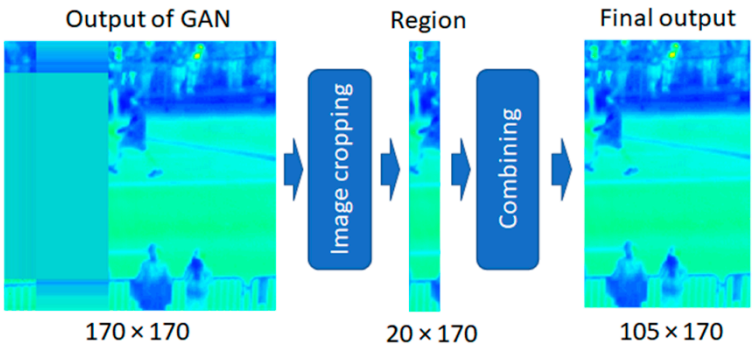


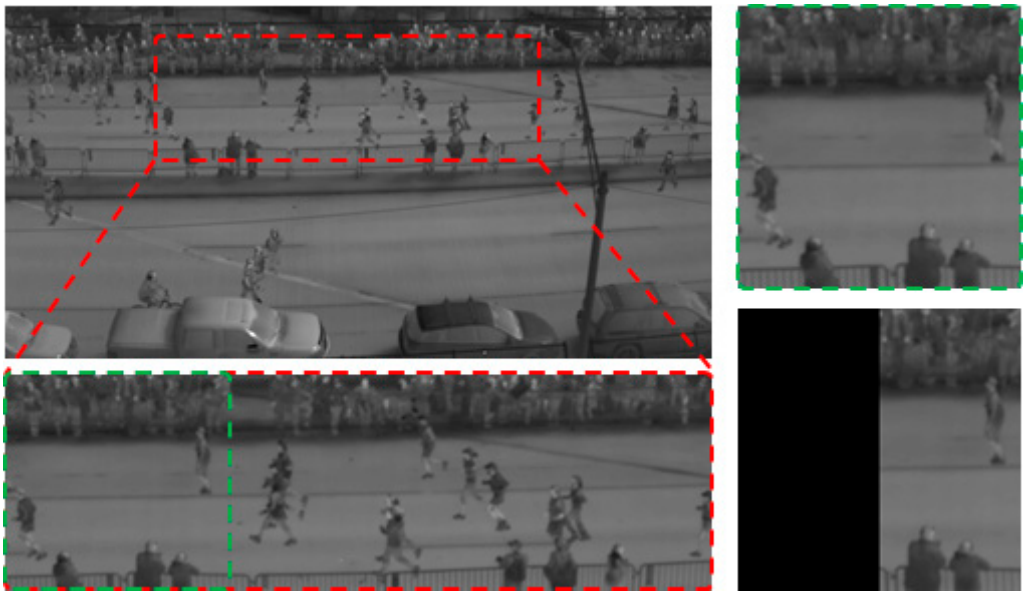
Figure 5. Example of the postprocessing.

3.5. Dataset and Experimental Setup

The experiment in this study was conducted using the marathon sub-dataset [14] of the BU-TIV benchmark open thermal dataset. The task of the marathon dataset was for multi-object tracking. The dataset has included various objects, namely, pedestrians, cars, motorcycles, bicycles, etc. The dataset consists of four videos (image sequences) with different sizes. The total number of images used in this experiment is 6552. Moreover, the size of an image in the marathon sub-dataset is 1024 × 512 × 1, and the pixel depth is 16 bits. The pixel value ranges between 3000 and 7000 units of uncalibrated temperature [14]. Images in the dataset are provided in portable network graphics (PNG) format. The four sequences were provided with annotations for the object detection. The dataset was collected using FLIR SC800 cameras (FLIR Systems, Inc., Wilsonville, OR, USA) [14]. We cropped all images into 170 × 170 × 1 and converted the image depth into 8 bits in this study.

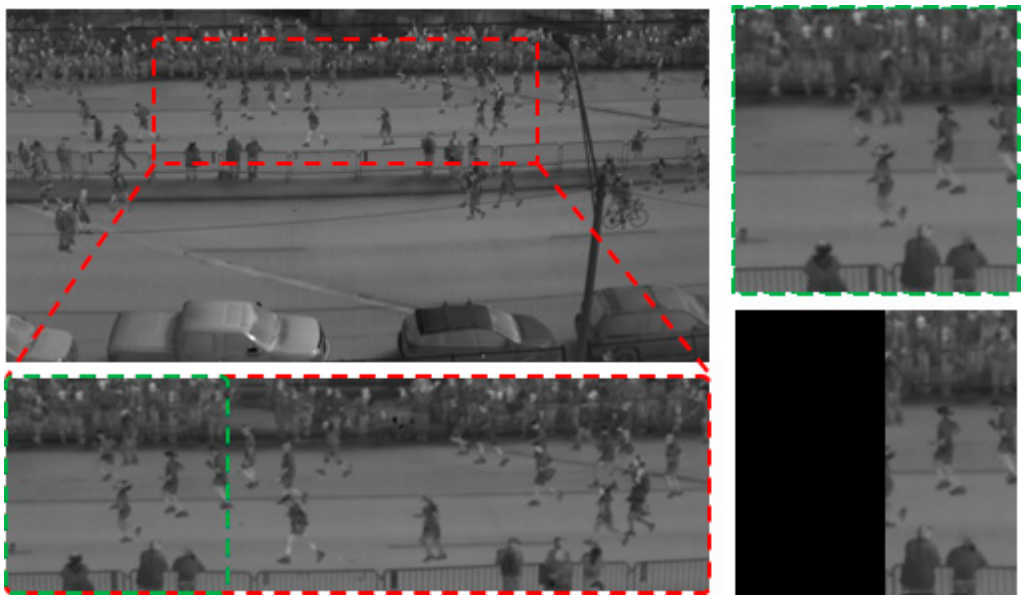
The experiment was conducted in two-fold cross validation. In other words, half of the total data were used for training, the other half for testing, and the average value of the two testing accuracies (obtained by repeating the same process after swapping the training and testing data) was set as the final accuracy. In this study, the region was cropped with respect to the road on which people are running (the region of interest (ROI) of the red dashed box in Figure 6) in the original image. Ground-truth images (green dashed box) and input images (an image with zero paddings) were generated by cropping the ROI images into images of size 170 × 170. The process of creating the dataset used in this study is shown in Figure 6.

The training and testing of the algorithm proposed in this study were conducted using a desktop computer equipped with Intel Core i7-6700 CPU @ 3.40 GHz (Intel Corp., Santa Clara, CA, USA), Nvidia GeForce GTX TITAN X graphic processing unit (GPU) card [33] (Nvidia Corp., Santa Clara, CA, USA), and a random-access memory (RAM) of 32 GB. The model and algorithm proposed in this study were implemented using the OpenCV library (version 4.3.0) [34] (Intel Corp., Santa Clara, CA, USA), Python (version 3.5.4) (Python Software Foundation, Wilmington, NC, USA), and Keras application programming interface (API) (version 2.1.6-tf) (MIT, Boston, MA, USA) with the TensorFlow backend engine (version 1.9.0) [35] (Google LLC, Mountain View, CA, USA).

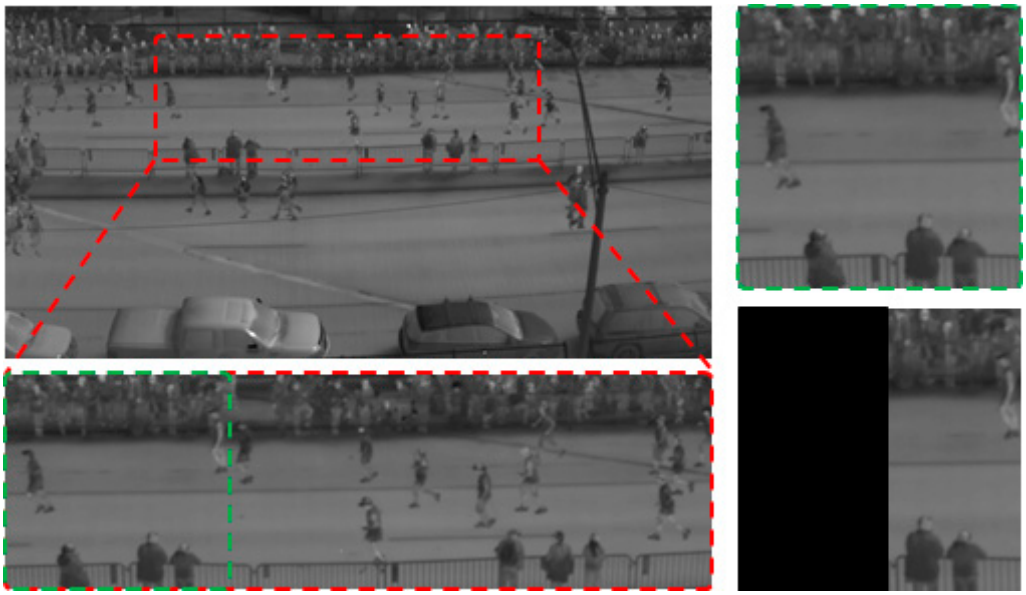


(**a**)

Figure 6. *Cont.*



(**b**)



(**c**)

Figure 6. Examples of dataset preparation. In (**a**–**c**), on the left: from top to bottom, an original thermal image and an ROI image. In (**a**–**c**), on the right: from top to bottom, a ground-truth image and an input image.

4. Results

The experiment conducted in this study was a two-fold cross validation. In other words, half of the data were used for training, whereas the remaining half were used for testing; then, the training data and testing data were switched to repeat the process, and the average of the two testing accuracies was determined as the final accuracy. In this section, the experimental results for the training, testing, and comparison are described in three separate subsections. In the training section, the hyperparameters and training loss used for training are described. In the testing section, the results obtained through ablation studies are compared. Finally, in the comparison section, the results obtained using the proposed method and the state-of-the-art methods are compared.

4.1. Training

The IPGAN structure proposed in this study was trained as follows. The batch size, training iterations, and learning rate of the IPGAN were set to 1, 800,000, and 0.0001, respectively. Furthermore, for both the generator and discriminator losses, we used the binary cross-entropy loss, and adaptive moment estimation (Adam) optimizer [36] was used as optimizer. Twenty sequential images of size 170 × 170 pixels were used in all the methods for both training and testing. Figure 7 shows the training loss curves of the IPGAN by iteration. In Table 7, detailed information of the hyperparameter tuning is presented. The remaining hyperparameters were determined according to the default values by Keras API [35].

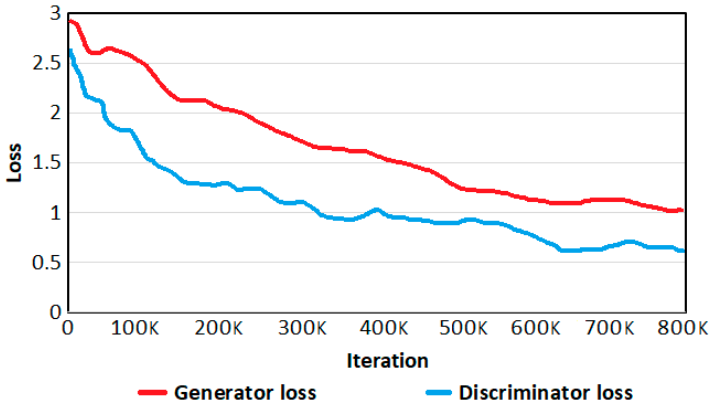


Figure 7. Training loss curves of GAN.

Table 7. Detailed information of hyperparameter tuning.

Parameters	Search Space	Selected Value
Weight decay (Weight regularization L2)	[0.001, 0.01, 0.1]	0.01
Loss	'mse', 'VGG-19 loss'	'mse'
Kernel initializer	'glorot uniform'	'glorot uniform'
Bias initializer	'zeros'	'zeros'
Optimizer	'SGD', 'adam'	'adam'
Learning rate	[0.0001, 0.001, 0.01, 0.1]	0.0001
Beta_1	[0.7, 0.8, 0.9]	0.9
Beta_2	[0.8, 0.9, 0.999]	0.999
Epsilon	$[1 \times 10^{-9}, 1 \times 10^{-8}, 1 \times 10^{-7}]$	1×10^{-8}
Iterations	[1~1638 K]	723,996
Batch size	[1, 4, 8]	1

4.2. Testing (Ablation Study)

In this section, the results of ablation studies for the proposed method are presented. The experiments were conducted using the same dataset and two types of GAN structures. For measuring the image prediction accuracy, the image region cropped in the resulting image (Figure 5) was compared with respect to the ground-truth region based on similarity. The accuracy was measured using three types of metrics shown in Equations (1)–(3).

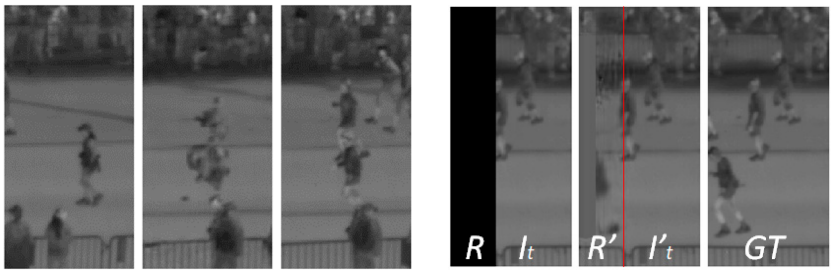
$$\mathrm{MSE} = \frac{\left(\sqrt{\sum_{y=1}^{H}\sum_{x=1}^{W}(T(x,y)-O(x,y))^2}\right)^2}{MN} \tag{1}$$

$$\mathrm{PSNR} = 10log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right) \tag{2}$$

$$\mathrm{SSIM} = \frac{(2\mu_O\mu_T + R1)(2\sigma_{OT} + R2)}{({\mu_O}^2 + {\mu_T}^2 + R1)({\sigma_O}^2 + {\sigma_T}^2 + R2)} \tag{3}$$

MSE represents the mean squared error [37] in Equation (1). W and H represent the image width and height, respectively, in Equation (1). Furthermore, in Equations (1) and (3), O and T represent the output image and target image (ground-truth image), respectively. PSNR represents the peak signal-to-noise ratio [38] in Equation (2). In Equation (3), the structural similarity index measure (SSIM) [39] is presented, in which μ_T and σ_T represent the mean and standard deviation of the pixel values of the ground-truth image, respectively, and μ_O and σ_O represent the mean and standard deviation of the pixel values of the output image, respectively. σ_{OT} represents the covariance of the two images. $R1$ and $R2$ are positive constants so as not to make the denominator zero.

In this section, seven different experiments were conducted. In Figures 8–10 the I_t image, the tth image of the 20 sequential images, is shown as the input image (far left). In Figure 8a, the target image (ground-truth (GT) image) (right image) is the subsequent image of the I_t image (left image), where GT and I_t do not include the same region. Specifically, the entire GT image that does not include any region of I_t was predicted in Method 1. However, the results obtained using this method varied significantly from the ground-truth image as shown by the output image (middle image) in Figure 8a. Accordingly, only the region R (zero padded black area of 30 × 170 pixels) of the image was predicted as shown in Figure 8b (Method 2). This method aims to predict the spatial information R which is not included in I_t when I_t is included in GT. However, gray noise is generated within the R region being predicted in the output image obtained using this method.



(**a**) (**b**)

Figure 8. Examples of result images obtained using Methods 1 and 2. From left to right, the input, output, and ground-truth images, respectively, obtained using (**a**) Method 1 and (**b**) Method 2. The size of the input, output, and ground-truth images is 80 × 170 pixels.

For improving the accuracy, unlike Methods 1 and 2 in Figure 8, which used the size of the input, output, and ground-truth images as 80 × 170 pixels, Methods 3 and 4 in Figure 9 set the size of the input, output, and ground-truth images to 170 × 170 pixels in order to use the spatial information that is wider in the horizontal direction. In addition, the experiment was conducted by setting the region R being predicted to be larger for Method 4 in Figure 9b (in Methods 3 and 4, the sizes of R were 17 × 170 pixels). However, the gray noise generated in R' became larger in Figure 9b. As the width of R increased, the gray noise also became larger in this experiment. Therefore, the region can only be predicted between the red and yellow lines of R' in Figure 9b, and it was difficult to predict the region to the left of the yellow line in this experiment. Therefore, as shown in Figure 5, the predicted region was cropped to a fixed size (20 × 170 pixels).

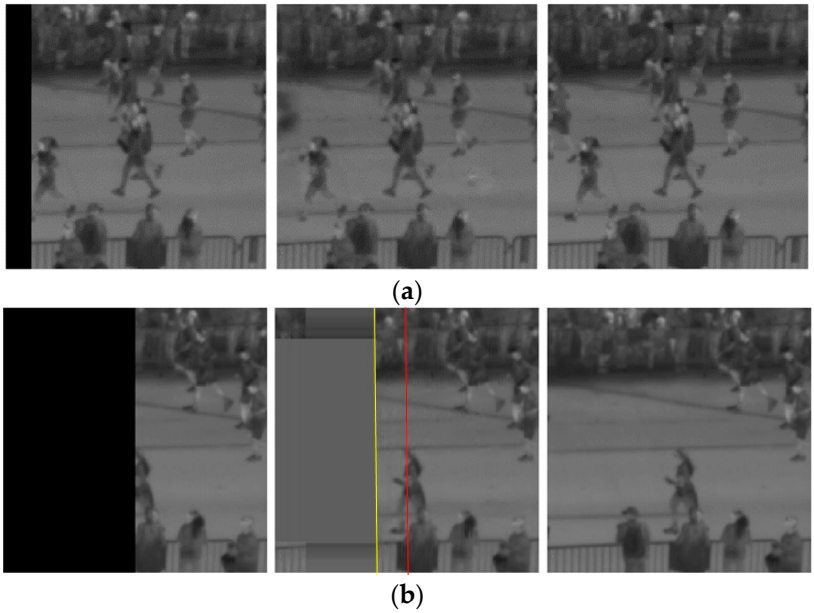


Figure 9. Examples of result images obtained using Methods 3 and 4. From left to right, the input, output, and ground-truth images, respectively, obtained using (**a**) Method 3 and (**b**) Method 4.

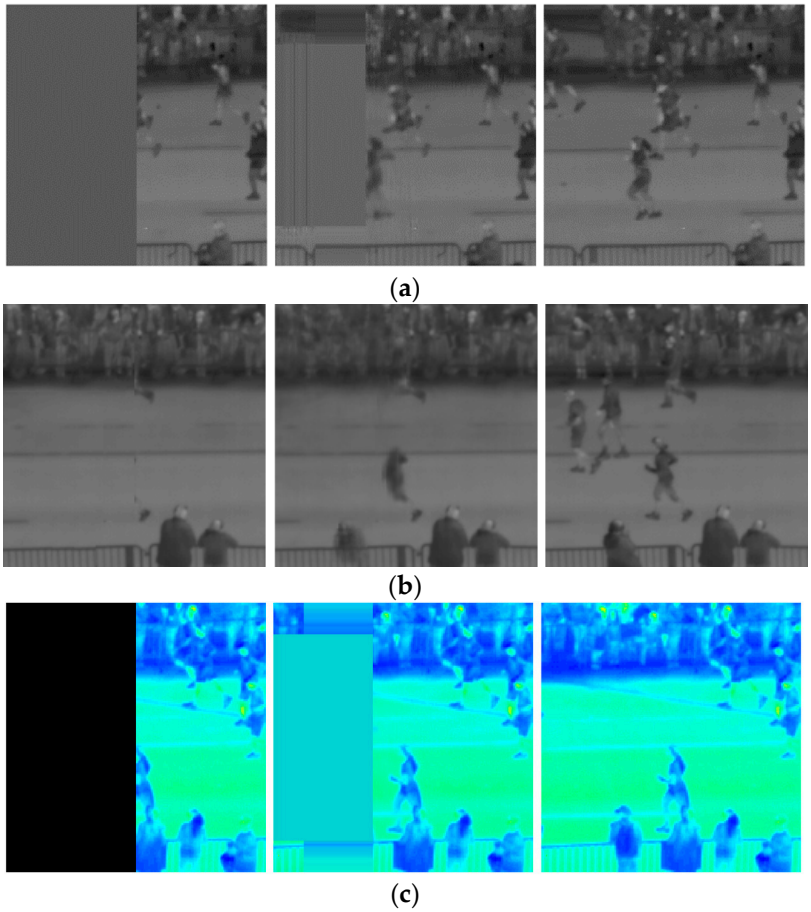


Figure 10. Examples of result images obtained using Methods 5–7. From left to right, the input, output, and ground-truth images, respectively, obtained using (**a**) Method 5, (**b**) Method 6, and (**c**) Method 7.

Moreover, the experiment was conducted as shown in Figure 10a by paddings with the average value of I_t to examine the effects of zero padding. In Method 6, the padding was performed using an empty background as in an input image shown in Figure 10b. The empty background was selected manually from marathon thermal images in order to examine the effects of zero padding. Moreover, in Figure 10, the size of the input, output, and ground-truth images was set to 170 × 170 pixels in order to use wider spatial information in the horizontal direction. However, the result obtained through zero paddings (Method 4) as shown in Figure 9b demonstrated the best performance among the results thus far. Finally, as shown in Figure 10c, the experiment was conducted using the converted three-channel color image (Method 7), and the accuracy was compared. A comparison of all the experimental results is presented in Table 8. The results of using a one-channel color image (Method 4) and a three-channel color image (Method 7) were compared in the images in Figure 11. As shown in Figures 8–11 and Table 8, Method 7 exhibited the best image prediction performance. In Table 8, Method 4 exhibited a better performance than Method 7 in terms of PSNR; however, it has been reported that the PSNR is a poor measure for evaluating the difference and similarity in the human visual-image quality [40,41]. SSIM can better evaluate the similarity in the image quality [39]. Thus, Method 7 demonstrated the highest accuracy. Figure 12 shows the examples of the output images obtained using the proposed method.

Table 8. Comparison of various region prediction methods.

Methods	PSNR	SSIM
Method 1	10.468	0.6157
Method 2	13.214	0.7817
Method 3	12.565	0.7423
Method 4	20.320	0.9131
Method 5	17.181	0.8814
Method 6	15.001	0.8303
Method 7	18.813	0.9535

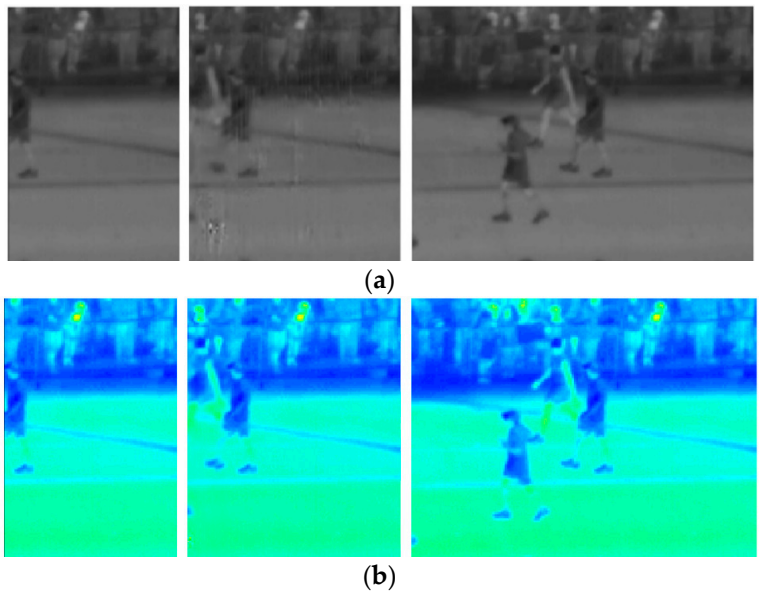


Figure 11. Examples of result images obtained using Methods 4 and 7. From left to right, the input, output, and ground-truth images, respectively, obtained using (**a**) Method 4 and (**b**) Method 7.

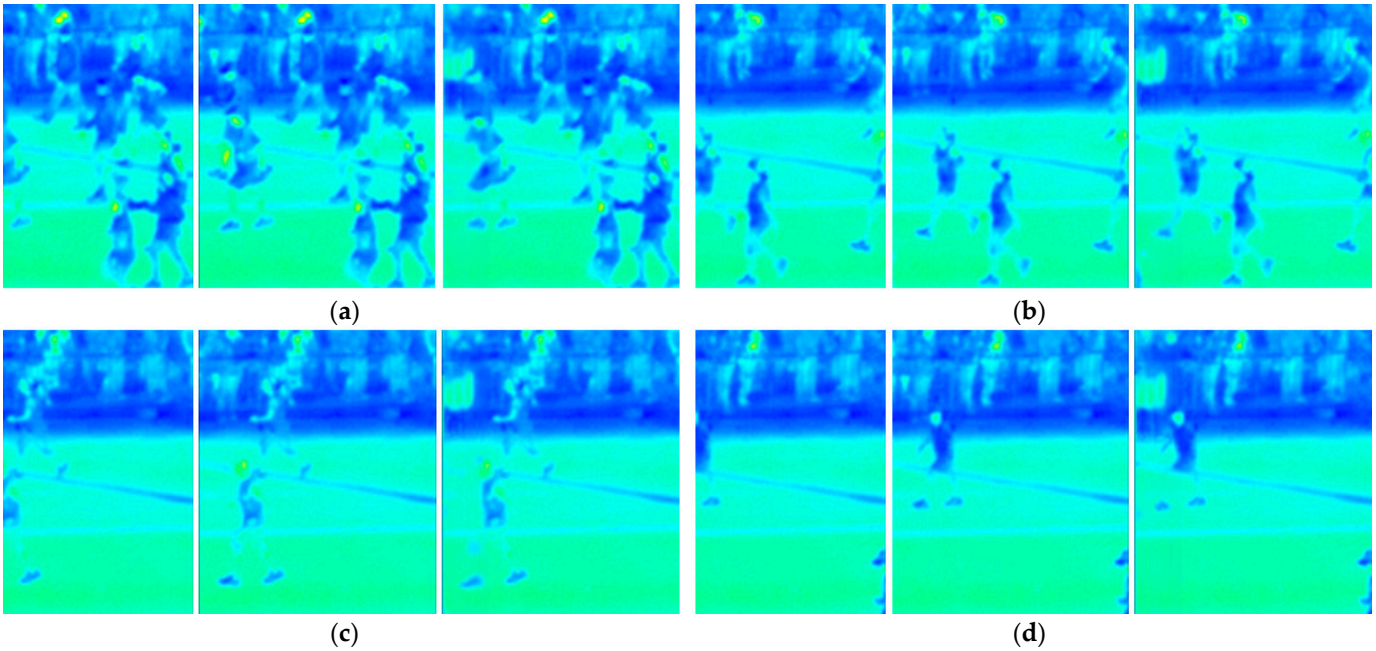


Figure 12. Examples of result images obtained using the proposed method. In (**a**–**d**), from left to right, the original, ground-truth, and predicted (output) images, respectively.

For inspecting the efficiency of the proposed method, the results of detecting humans in the original input and ground-truth images were compared with the result of detecting humans in the predicted image using the proposed method. Mask R-CNN [42] was used for conducting the experiment on human detection. Figure 13 shows the result of detecting humans using Mask R-CNN as mask images.

As shown in Figure 13, the result of human detection in the ground-truth image is similar to the result of human detection in the image predicted by the IPGAN for which a three-channel color image is input. Furthermore, the detection result from the predicted image is closer to the detection result from the ground-truth image than that from the original input image.

Additionally, the detection (detection 1) accuracy was measured between the results obtained with the original input images and the results obtained with the ground-truth images. The detection (detection 2) accuracy was also measured between the results obtained with the images predicted using our method and the results obtained with the ground-truth images. These detection results (detection 1 and detection 2) were compared in Table 9. To this end, the true positive rate (TPR) (#TP/(#TP + #FN)) and positive predictive value (PPV) (#TP/(#TP + #FP)) [43], as well as the accuracy (ACC) [43], F1 score (F1) [44], and intersection over union (IoU) [43], which are expressed in Equations (4)–(6), respectively, were used to measure the accuracy for a comparison. Here, TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative, respectively. Positive and negative in this experiment indicate the pixels detected in the ground-truth image (white pixel in Figure 13) and those not detected (black pixel in Figure 13), respectively. More specifically, TP refers to the case when positive pixels are detected correctly, whereas TN refers to the case when negative pixels are not detected correctly. FP refers to the case when negative pixels are incorrectly detected as positive pixels, whereas FN refers to the case when positive pixels are incorrectly detected as negative pixels. Here, "#" denotes "the number of."

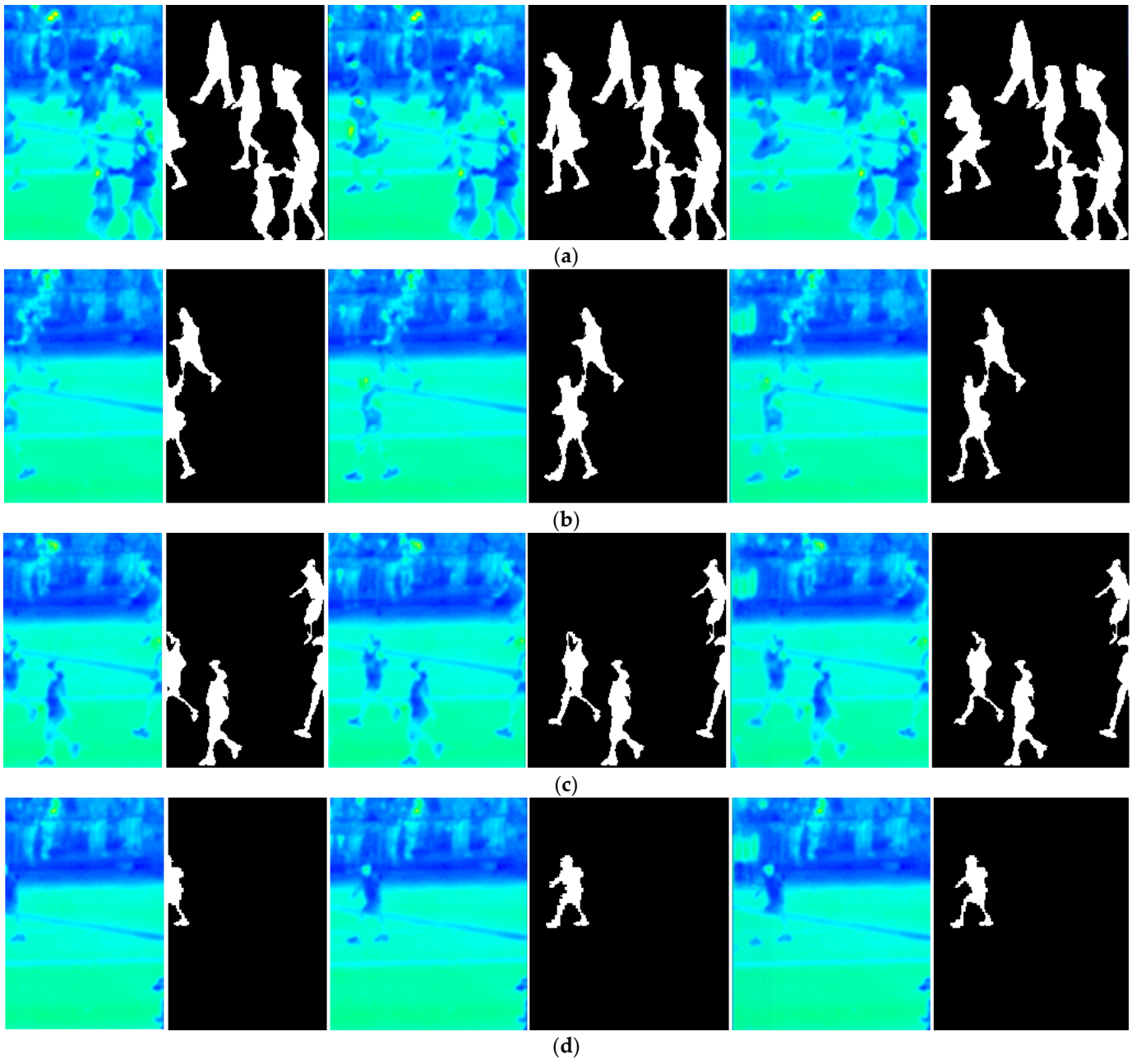


Figure 13. Examples of detection results before and after image prediction. In (**a**–**d**), from left to right, the original input images, results with original input images, ground-truth images, results with ground-truth images, images predicted using our method, and results with predicted images, respectively.

Table 9. Comparisons of object detection accuracies by detections 1 and 2.

Methods	TPR	PPV	F1	ACC	IoU
Detection 1	0.82	0.81	0.815	0.941	0.713
Detection 2	0.901	0.864	0.882	0.983	0.791

As shown in Table 9, detection 2 was more accurate than detection 1, which indicates that using the image predicted with our method produced the detection results closer to the results of using the ground-truth image than using the original input image.

$$\mathrm{ACC} = \frac{\#\mathrm{TP} + \#\mathrm{TN}}{\#\mathrm{TP} + \#\mathrm{TN} + \#\mathrm{FP} + \#\mathrm{FN}} \tag{4}$$

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} \tag{5}$$

$$\mathrm{IoU}(\mathrm{X}, \mathrm{Y}) = \frac{|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X} \cup \mathrm{Y}|} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP} + \#\mathrm{FN}} \tag{6}$$

4.3. Comparisons of Proposed Method with the State-of-the-Art Methods

In this section, the results of comparing the proposed method and the state-of-the-art methods are presented. For measuring the image prediction accuracy, the entire image (including R' and I'_t as in Figure 8b) obtained using the proposed method was compared with respect to the ground-truth based on similarity. In Table 10, the existing image prediction [15] method and inpainting [20,22,24] methods are compared with the IPGAN-based image region prediction method proposed in this study. In Figure 14, the result images obtained using all the methods are compared. The methods that originally used a single image [20,22,24] are made to use sequential images as inputs, as in our method, for a fair performance evaluation; the input layer of these methods [20,22,24] was changed to layers 0 and 1 of Table 2, as in the proposed method. Moreover, a three-channel color image was used as the input and output of all the methods, as in our method, for a fair comparison and evaluation. As shown in Table 10 and Figure 14, the proposed method exhibited a better performance than the state-of-the-art methods.

For the next experiment, all the methods were compared using Mask-R-CNN-based human detection. Table 11 and Figure 15 show the accuracy of the detection results as well as the output images. The experiment showed that the proposed method demonstrated the best performance.

Table 10. Comparison of the image prediction methods.

Methods	PSNR	SSIM
Haziq et al.'s [15]	23.185	0.9523
Liu et al.'s [20]	22.210	0.9310
Shin et al.'s [22]	22.813	0.9451
Nazeri et al.'s [24]	22.742	0.9131
Proposed method	23.243	0.9839

Table 11. Comparisons of object detection accuracies obtained using our method with those of the state-of-the-art methods based on Mask R-CNN.

Methods	TPR	PPV	F1	ACC	IoU
Haziq et al.'s [15]	0.825	0.684	0.747	0.957	0.589
Liu et al.'s [20]	0.652	0.687	0.669	0.961	0.491
Shin et al.'s [22]	0.739	0.676	0.706	0.959	0.558
Nazeri et al.'s [24]	0.71	0.662	0.685	0.931	0.522
Proposed method	0.901	0.864	0.882	0.983	0.791

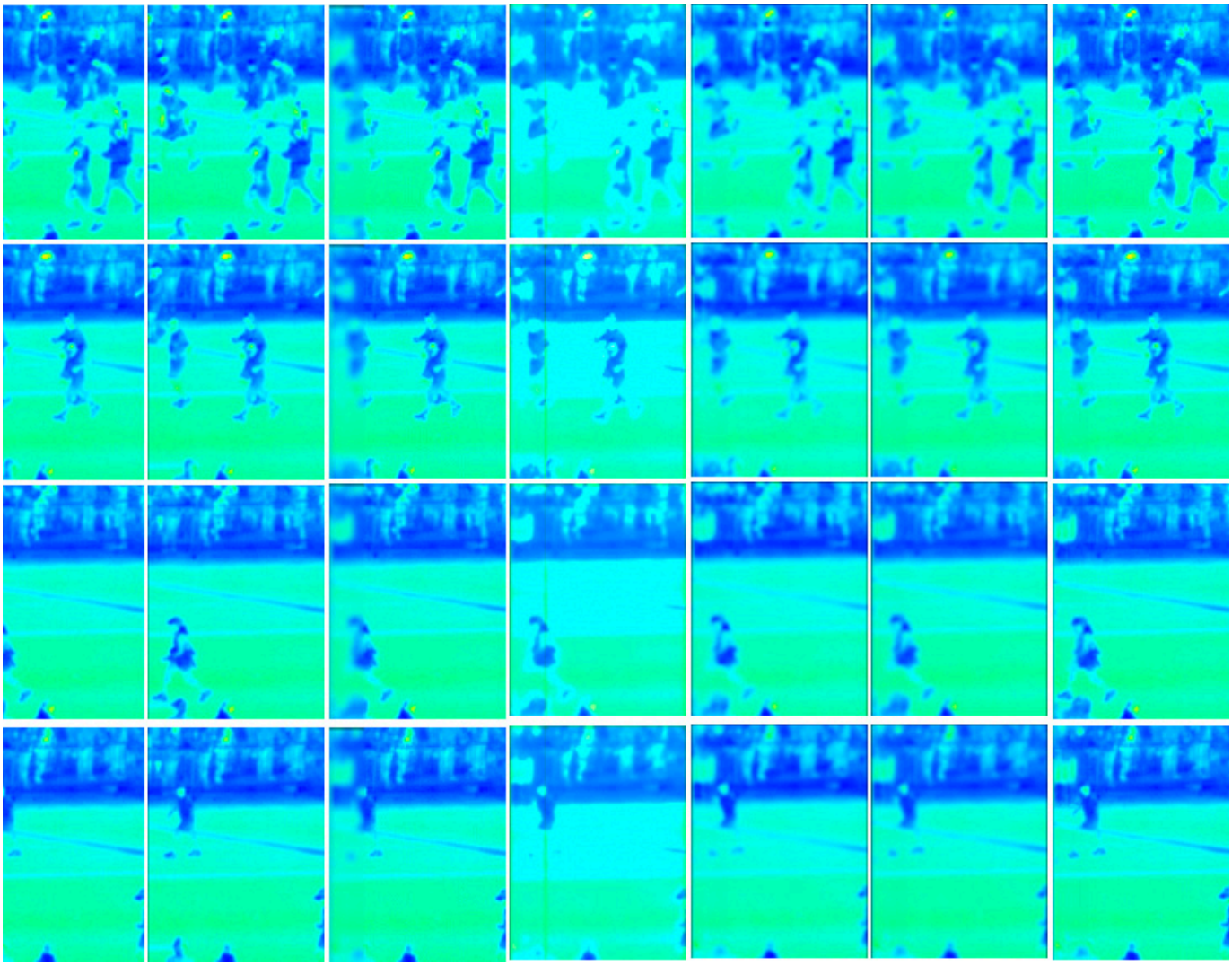


Figure 14. Comparisons of the original images, ground-truth images, and prediction results obtained using the state-of-the-art methods and our method: (**a**) original images; (**b**) ground-truth images. Images predicted using: (**c**) Haziq et al.'s method; (**d**) Liu et al.'s method; (**e**) Shin et al.'s method; (**f**) Nazeri et al.'s method; (**g**) the proposed method.

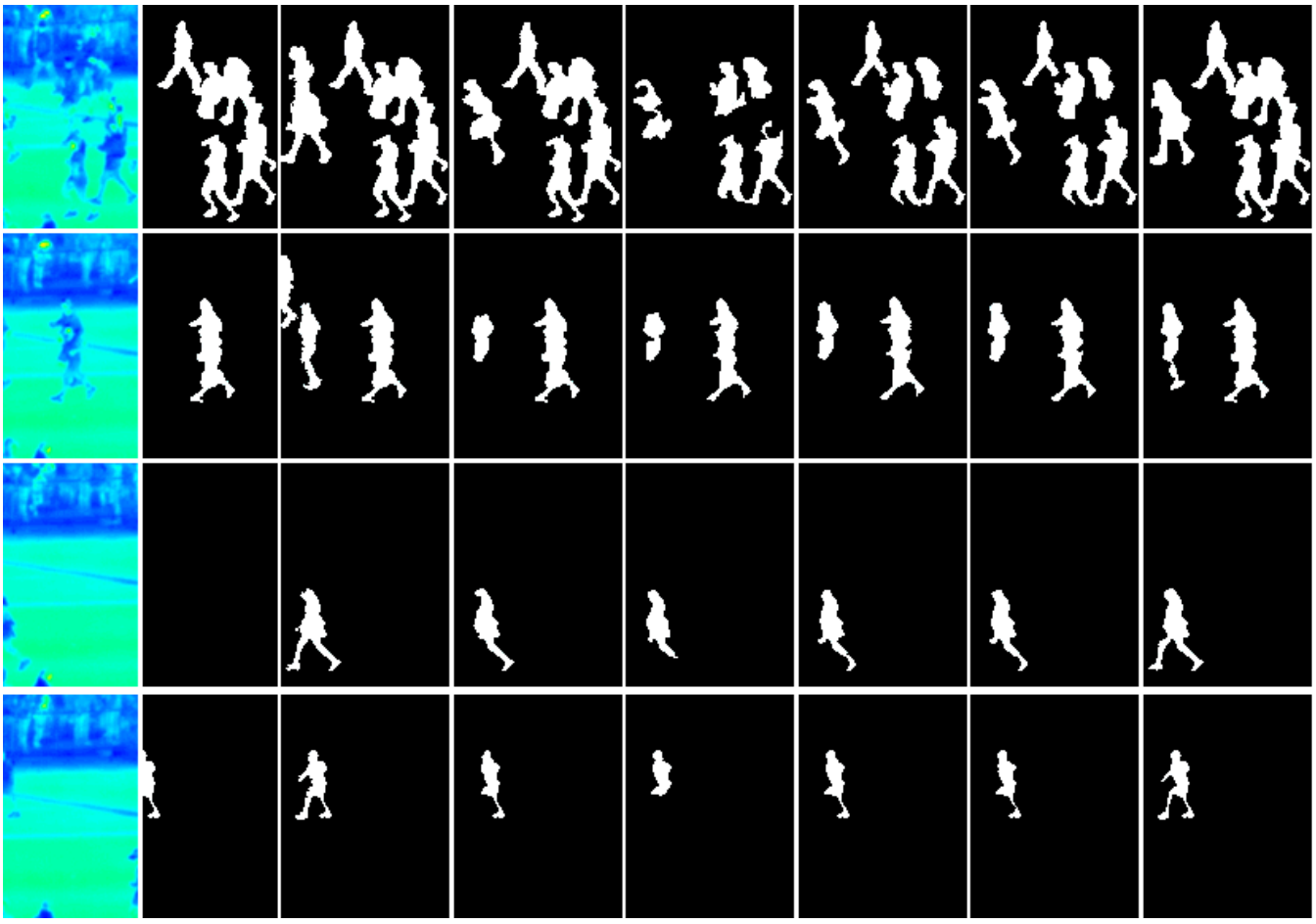


Figure 15. Comparisons of detection results using the original images, ground-truth images, and the predicted images obtained using the state-of-the-art methods and our method. (**a**) Original images. Detection results using the (**b**) original images, (**c**) ground-truth images, (**d**) images predicted using Haziq et al.'s method, (**e**) images predicted using Liu et al.'s method, (**f**) images predicted using Shin et al.'s method, (**g**) images predicted using Nazeri et al.'s method, and (**h**) images predicted using our method.

4.4. Processing Time

In Table 12, the processing time of each sub-part of the proposed method (Figure 2) is presented. The processing time was measured in the environments described in Section 3.5. As shown in Table 12, the processing time of the Mask R-CNN is higher than other sub-parts. The frame rate of the proposed prediction method is about 23.4 frames per second (1000/(9.97 + 32.8 + 0.01)), and the total frame rate including image prediction and detection method is about 10.6 frames per second (1000/94). Thus, the processing time of the proposed method to perform both image prediction and object detection is sufficiently short.

Table 12. Processing time of the proposed method per image (unit: ms).

Sub-Part	Processing Time
Preprocessing	9.97
Image prediction by IPGAN	32.8
Postprocessing	0.01
Object detection by Mask R-CNN	51.22
Total	94

5. Discussion

In this study, a method was proposed for predicting the image outside the FOV of a camera. The proposed method was studied for accurately detecting humans who are leaving the FOV of a camera, by which the object detection error, due to a part of a human body being invisible in the input image, can be reduced. As shown in the result images of Figures 13–15, the invisible body parts of humans leaving the FOV of a camera became visible in the images by the proposed image prediction method. Therefore, it is confirmed that the proposed method is efficient to predict missing parts of a human body as well as is sufficient to increase the accuracy of a human detection.

However, it is confirmed that the size of region being predicted is limited when the images outside of the FOV of a camera are predicted as shown in Figure 9b. In addition, the gray noises are generated in the *R* region (Figure 8b). As the width of *R* increased, the gray noise also became larger in this experiment, and the consequent size of the predicted region became limited. Therefore, our method can be used for the applications where the region of limited size is predicted for human detection in thermal videos.

6. Conclusions

In this study, a method was proposed for predicting the image outside the FOV of a camera using a one-channel thermal image converted into a three-channel thermal image as an input of the IPGAN. Various ablation studies based on different image size and image channels were conducted and compared in this study. The method based on a three-channel thermal image showed a higher SSIM (0.9535) value compared to one-channel thermal image-based methods. Moreover, it was confirmed that the image prediction method increased the accuracy of object detection as shown in Table 9. For example, the TPR = 0.82, PPV = 0.81, F1 score = 0.815, ACC = 0.941, and IoU = 0.713 were increased to TPR = 0.901, PPV = 0.864, F1 score = 0.882, ACC = 0.983, and IoU = 0.791. In addition, the proposed method was compared with the state-of-the-art methods, and our method showed higher PSNR = 23.243 and SSIM = 0.9839 values than the state-of-the-art methods as shown in Table 10. Our method was also compared with the state-of-the-art methods in terms of human detection, and the proposed method showed the TPR = 0.901, PPV = 0.864, F1 score = 0.882, ACC = 0.983, and IoU = 0.791 which were higher than the state-of-the-art methods as shown in Table 11.

In future work, the methods for predicting a wider region will be studied. Furthermore, an image prediction method in which the front viewing angle of a vehicle's visible-light camera is expanded in horizontal directions will be investigated by expanding the scope of this study.

Author Contributions: Methodology, G.B.; conceptualization, J.H.K.; validation, Y.H.K.; supervision, K.R.P.; writing—original draft, G.B.; writing—review and editing, K.R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) through the Basic Science Research Program (NRF-2019R1F1A1041123), in part by the NRF funded by the MSIT through the Basic Science Research Program (NRF-2020R1A2C1006179), and in part by the MSIT, Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jeon, E.S.; Kim, J.H.; Hong, H.G.; Batchuluun, G.; Park, K.R. Human detection based on the generation of a background image and fuzzy system by using a thermal camera. *Sensors* **2016**, *16*, 453. [CrossRef]
2. Batchuluun, G.; Kang, J.K.; Nguyen, D.T.; Pham, T.D.; Muhammad, A.; Park, K.R. Deep learning-based thermal image reconstruction and object detection. *IEEE Access* **2021**, *9*, 5951–5971. [CrossRef]
3. Batchuluun, G.; Yoon, H.S.; Nguyen, D.T.; Pham, T.D.; Park, K.R. A study on the elimination of thermal reflections. *IEEE Access* **2019**, *7*, 174597–174611. [CrossRef]
4. Batchuluun, G.; Baek, N.R.; Nguyen, D.T.; Pham, T.D.; Park, K.R. Region-based removal of thermal reflection using pruned fully convolutional network. *IEEE Access* **2020**, *8*, 75741–75760. [CrossRef]
5. Liu, Q.; Li, X.; He, Z.; Fan, N.; Yuan, D.; Wang, H. Learning deep multi-level similarity for thermal infrared object tracking. *IEEE Trans. Multimedia* 2020. [CrossRef]
6. Zulkifley, M.A. Two streams multiple-model object tracker for thermal infrared video. *IEEE Access* **2019**, *7*, 32383–32392. [CrossRef]
7. Zulkifley, M.A.; Trigoni, N. Multiple-model fully convolutional neural networks for single object tracking on thermal infrared video. *IEEE Access* **2018**, *6*, 42790–42799. [CrossRef]
8. Stojanović, M.; Vlahović, N.; Stanković, M.; Stanković, S. Object tracking in thermal imaging using kernelized correlation filters. In Proceedings of the 17th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 21–23 March 2018.
9. Asha, C.S.; Narasimhadhan, A.V. Experimental evaluation of feature channels for object tracking in RGB and thermal imagery using correlation filter. In Proceedings of the Twenty-third National Conference on Communications (NCC), Chennai, India, 2–4 March 2017.
10. Batchuluun, G.; Kim, Y.G.; Kim, J.H.; Hong, H.G.; Park, K.R. Robust behavior recognition in intelligent surveillance environments. *Sensors* **2016**, *16*, 1010. [CrossRef]
11. Batchuluun, G.; Kim, J.H.; Hong, H.G.; Kang, J.K.; Park, K.R. Fuzzy system based human behavior recognition by combining behavior prediction and recognition. *Expert Syst. Appl.* **2017**, *81*, 108–133. [CrossRef]
12. Batchuluun, G.; Nguyen, D.T.; Pham, T.D.; Park, C.; Park, K.R. Action recognition from thermal videos. *IEEE Access* **2019**, *7*, 103893–103917. [CrossRef]
13. Mun, J.-H.; Jeon, M.; Lee, B.-G. Unsupervised learning for depth, ego-motion, and optical flow estimation using coupled consistency conditions. *Sensors* **2019**, *19*, 2459. [CrossRef]
14. Wu, Z.; Fuller, N.; Theriault, D.; Betke, M. A thermal infrared video benchmark for visual analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014.
15. Haziq, R.; Basura, F. A log-likelihood regularized KL divergence for video prediction with a 3D convolutional variational recurrent network. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, Waikola, HI, USA, 5–9 January 2021.
16. Guen, V.L.; Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
17. Finn, C.; Goodfellow, I.; Levine, S. Unsupervised learning for physical interaction through video prediction. In Proceedings of the Advances in Neural Information Processing Systems 29, Barcelona, Spain, 5–10 December 2016.
18. Xu, J.; Xu, H.; Ni, B.; Yang, X.; Darrell, T. Video prediction via example guidance. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020.
19. Babaeizadeh, M.; Finn, C.; Erhan, D.; Campbell, R.H.; Levine, S. Stochastic variational video prediction. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
20. Liu, H.; Jiang, B.; Xiao, Y.; Yang, C. Coherent semantic attention for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea, 27 October–2 November 2019.
21. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea, 27 October–2 November 2019.
22. Shin, Y.-G.; Sagong, M.-C.; Yeo, Y.-J.; Kim, S.-W.; Ko, S.-J. PEPSI++: Fast and lightweight network for image inpainting. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 252–265. [CrossRef]
23. Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
24. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Structure guided image inpainting using edge prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea, 27 October–2 November 2019.
25. Oprea, S.; Martinez-Gonzalez, P.; Garcia-Garcia, A.; Castro-Vargas, J.A.; Orts-Escolano, S.; Garcia-Rodriguez, J.; Argyros, A. A review on deep learning techniques for video prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020. [CrossRef]
26. Rasouli, A. Deep learning for vision-based prediction: A survey. *arXiv* **2020**, arXiv:2007.00095v2.
27. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Akbari, Y. Image inpainting: A review. *Neural Process. Lett.* **2020**, *51*, 2007–2028. [CrossRef]
28. Image Prediction Generative Adversarial Network (IPGAN). Available online: http://dm.dgu.edu/link.html (accessed on 25 March 2021).

29. MathWorks. Available online: https://www.mathworks.com/help/matlab/ref/jet.html (accessed on 25 March 2021).
30. Batchuluun, G.; Lee, Y.W.; Nguyen, D.T.; Pham, T.D.; Park, K.R. Thermal image reconstruction using deep learning. *IEEE Access* **2020**, *8*, 126839–126858. [CrossRef]
31. Batchuluun, G.; Kang, J.K.; Nguyen, D.T.; Pham, T.D.; Arsalan, M.; Park, K.R. Action recognition from thermal videos using joint and skeleton information. *IEEE Access* **2021**, *9*, 11716–11733. [CrossRef]
32. Funt, B.; Zhu, L. Does colour really matter? Evaluation via object classification. In Proceedings of the 26th Color and Imaging Conference Final Program and Proceedings, Vancouver, BC, Canada, 12–16 November 2018.
33. NVIDIA Corporation. Available online: https://www.nvidia.com/en-us/geforce/products/10series/titan-x-pascal/ (accessed on 25 March 2021).
34. OpenCV. Available online: http://opencv.org/ (accessed on 25 March 2021).
35. Keras. Available online: https://keras.io/ (accessed on 25 March 2021).
36. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
37. Mean Squared Error. Available online: https://en.wikipedia.org/wiki/Mean_squared_error (accessed on 29 April 2021).
38. Peak Signal-to-Noise Ratio. Available online: https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio (accessed on 29 April 2021).
39. Structural Similarity. Available online: https://en.wikipedia.org/wiki/Structural_similarity (accessed on 29 April 2021).
40. Huynh-Thu, Q.; Ghanbari, M. The Accuracy of PSNR in predicting video quality for different video scenes and frame rates. *Telecommun. Syst.* **2012**, *49*, 35–48. [CrossRef]
41. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **2008**, *44*, 800–801. [CrossRef]
42. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
43. Powers, D.M.W. Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Mach. Learn. Technol.* **2011**, *2*, 37–63.
44. Derczynski, L. Complementarity, f-score, and NLP evaluation. In Proceedings of the International Conference on Language Resources and Evaluation, Portorož, Slovenia, 23–28 May 2016.