mathematics

MDPI

Article

Enlargement of the Field of View Based on Image Region Prediction Using Thermal Videos

Ganbayar Batchuluun, Na Rae Baek 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.); naris27@dgu.ac.kr (N.R.B.)
* Correspondence: parkgr@dgu.edu

Abstract: Various studies have been conducted for detecting humans in images. However, there are the cases where a part of human body disappears in the input image and leaves the camera field of view (FOV). Moreover, there are the cases where a pedestrian comes into the FOV as a part of the body slowly appears. In these cases, human detection and tracking fail by existing methods. Therefore, we propose the method for predicting a wider region than the FOV of a thermal camera based on the image prediction generative adversarial network version 2 (IPGAN-2). When an experiment was conducted using the marathon subdataset of the Boston University-thermal infrared video benchmark open dataset, the proposed method showed higher image prediction (structural similarity index measure (SSIM) of 0.9437) and object detection (F1 score of 0.866, accuracy of 0.914, and intersection over union (IoU) of 0.730) accuracies than state-of-the-art methods.

Keywords: image prediction; thermal videos; deep learning; IPGAN-2

Citation: Batchuluun, G.; Baek, N.R.; Park, K.R. Enlargement of the Field of View Based on Image Region Prediction Using Thermal Videos. *Mathematics* **2021**, *9*, 2379. https://doi.org/10.3390/math9192379

Academic Editors: Ezequiel López-Rubio, Esteban Palomo and Enrique Domínguez

Received: 17 August 2021
Accepted: 22 September 2021
Published: 25 September 2021

1. Introduction

Extensive research has been conducted on objection detection [1–4], tracking [5–9], and action recognition [10–13] using conventional camera-based detection systems. However, there are frames where a part of body of a pedestrian disappears because the part of the body of the pedestrian is outside a camera's field of view (FOV) when walking or running. Moreover, there are cases in which a pedestrian comes into the FOV as a part of the body slowly appears. These cases cause the person to be detected or not detected inconsistently. An error also occurs in human tracking and action recognition. In a previous study, the issue in which a part of human body disappears was examined [14], but only a small region within an input image could be predicted. To overcome such an issue, in this study, for the first time, an image restoration was performed, as shown in Figure 1, by predicting the wide region outside the FOV not included in the current image (t) as in image t' for restoring the disappeared part of the body of a pedestrian in a thermal image. The proposed method predicts wider regions on both sides of the FOV in a current image using an image prediction generative adversarial network version 2 (IPGAN-2)-based method, the preceding sequential frame, and the current frame. In this study, various experiments were performed using the marathon subdataset [15] of the Boston University-thermal infrared video (BU-TIV) benchmark open dataset.

In addition, this study is novel in the following four ways compared with the previous studies.

- For thermal camera images, in this study, image prediction was performed in which, for the first time, the occurrence of noise was minimized while wide regions to left and right sides of the FOV in the current image were accurately generated.
- In this study, IPGAN-2 is proposed for performing image prediction.
- For improving the accuracy of image prediction, binary images corresponding to sequential input thermal images were used as input for IPGAN-2.

- The IPGAN-2 model proposed has been disclosed [16] for a fair performance evaluation by other researchers.

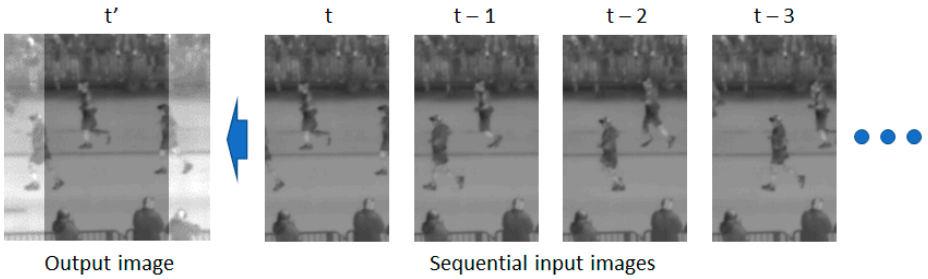


Figure 1. Example of thermal image prediction.

The remainder of this study is organized as follows. In Section 2, previous studies are reviewed. In Section 3, the proposed method is explained in detail. Experimental results and analysis are provided in Section 4. Finally, a discussion and the conclusions are provided in Sections 5 and 6, respectively.

2. Related Works

Previous studies on image prediction that generate the current frame or next frame can be largely divided into five categories, as explained in Sections 2.1 and 2.2.3.

2.1. Not Using Previous Frames but Using Current Frame (Image Inpainting)

Studies (image inpainting) have been conducted on the restoration of part of a current image by using only the current frame [17–22]. In [17], a fine deep-generative-model-based approach with a novel coherent semantic attention (CSA) layer was used to restore a visible light image. In [18], a visible light image was restored based on gated convolution and SN-PatchGAN. In [19], a visible light image was restored based on the parallel extended-decoder path for semantic inpainting network (PEPSI). In [20], a visible light image was restored using a context encoder method based on a channel-wise fully connected layer. A visible light image was restored in [21] using a method based on edge prediction and image completion based on the predicted edge map. Finally, in [22], the sequential-based, convolutional neural network (CNN)-based, and generative adversarial network (GAN)-based image restoration methods and the datasets used were explained.

2.2. Using Current and Previous Frames

2.2.1. Prediction of Next Frame

In some earlier studies [23–25], a next frame was predicted using the current frame and previous sequential frames. A dual-motion GAN model (ConvLSTMGAN) was proposed [23], and image prediction was performed using a visible light image. This method involves encoding sequential input frames using a probabilistic motion encoder (encoder CNN). The encoder CNN consists of four convolutional layers, one intermediate ConvLSTM layer, and two ConvLSTM layers. Furthermore, the next frame and next flow images are generated through future-image and future-flow generators. In [24], a method was proposed for generating the next optical flow image and next frame using a visible light image and encoder and decoder CNN (OptCNN-Hybrid). In this method, the proposed network was trained in a hybrid way using real and synthetic videos. In [25], a method for generating the next frame using a visible light image and ConvLSTM was proposed. In this study, the depth image is predicted using a current image and camera trajectory. Moreover, the next frame is generated using depth information by creating a depth image based on the advantages of camera scene geometry.

2.2.2. Prediction of Next Sequential Frames

In earlier studies [26–30], the next sequential frames were predicted using the current frame and previous sequential frames. In [26], image prediction was performed using a visible light image and the encoder and decoder model based on long short-term memory (LSTM) and a 3D convolution layer. In [27], the image was predicted using a visible light image, the newly proposed PhyCell, and PhyDNet based on LSTM. In [28], image prediction was performed using a visible light image, LSTM, and a CNN. In [29], the image was predicted using a visible light image and the encoder and decoder model. Image prediction was performed in [30] using a visible light image and a stochastic variational video prediction (SV2P) method. In a review study [31], the datasets from 2004 to 2019 used in image prediction were compared with the image prediction models that were released between 2014 and 2020. In the survey in [32], studies on and datasets for image prediction were explained.

2.2.3. Prediction of Small Left Region of Current Frame

In the following study, a region out of the FOV of a current frame was generated using the current frame and previous sequential frames. In [14], image prediction was performed in which a region out of the FOV was generated using a thermal video and GAN. The regions outside the FOV were predicted using the image obtained from a thermal camera that measured the heat of a human body rather than the image obtained from a general visible light camera. However, this method created a wide image by predicting a small region to the left of the FOV. Noise also occurred in the prediction region in the generated image, and the region includes more noise as the size of the region being predicted increased. Therefore, there is a limitation in the size of the region being predicted.

Table 1 provides comparisons between the proposed method and previous studies.

Table 1. Summaries of comparisons between the proposed method and previous image prediction studies.

Category	Not Using Previous Frames but Using Current Frame (Prediction of Removed Part in Current Frame (Image in painting))	Using Current and Previous Frames			
		Prediction of Next Frame	Prediction of Next Sequential Frames	Prediction of Small Left Region of Current Frame	Prediction of Large Right and Left Regions of Current Frame
Methods	CSA layer [17], gated convolution + SN-PatchGAN [18], PEPSI [19], context encoder [20], edge prediction and image completion [21], and review [22]	ConvLSTMGAN [23], OptCNN-Hybrid [24], ConvLSTM [25]	Encoder–decoder model [26,29], PhyDNet [27], CNN + LSTM [28], SV2P [30], and review & survey [31,32]	IPGAN [14]	IPGAN-2 (Proposed method)
Input	High-quality and high-resolution RGB visible light image			Low-quality and low-resolution grayscale thermal image	Low-quality and low-resolution grayscale thermal image and binary image
Output	RGB visible light image			An RGB thermal image	A grayscale thermal image
Advantages	High performance is achieved by restoring the information deleted in the current image by using the remaining information of the current image.	High performance is obtained when generating the next image by using the current image and previous sequential images.		- Considers image prediction besides FOV - Uses low-resolution, low-quality thermal image	- A wide image of left and right of the FOV is generated for image prediction - Noise does not occur in the predicted image outside the FOV
Disadvantages	- Does not consider image prediction outside the FOV - Does not use low-resolution, low-quality thermal image			- The size of predicted image is limited. - Noise occurs in the predicted image - Only the region to the left of the FOV is generated for image prediction.	Low processing speed

3. Materials and Methods

3.1. Overall Procedure of Proposed Method

In this section, the proposed method is explained in detail. Image region prediction is performed in this method based on sequential thermal images using IPGAN-2. In Sections 3.2–3.5, the IPGAN-2 architecture, postprocessing, differences between IPGAN and the proposed IPGAN-2, and dataset with experimental setup for image prediction are explained in detail. Figure 2a shows the overall flowchart of the proposed method and Figure 2b shows the overall procedure of the proposed method with image examples. The length of the sequential input images is 20 frames ($t-0, t-1, \ldots t-19$), the size of each image is 120×160 pixels, and the size the output image is 200×160 pixels. Specifically, the part connecting a disappeared part of a person not in the camera FOV (left region of the FOV) and the background in the current image is generated, while simultaneously generating a disappeared part of a person coming into the camera FOV (right region of the FOV) to generate the output image. As shown in Figure 2b, sequential thermal images for input and the corresponding sequential binary images are used as input for IPGAN-2. Input images for image prediction are horizontally flipped, and IPGAN-2 is applied one more time, during which the same model is used. In Figure 2b, red arrows represent a horizontal concatenate operation of the three images. In Table 2, the detailed procedure of the proposed algorithm is explained step by step.

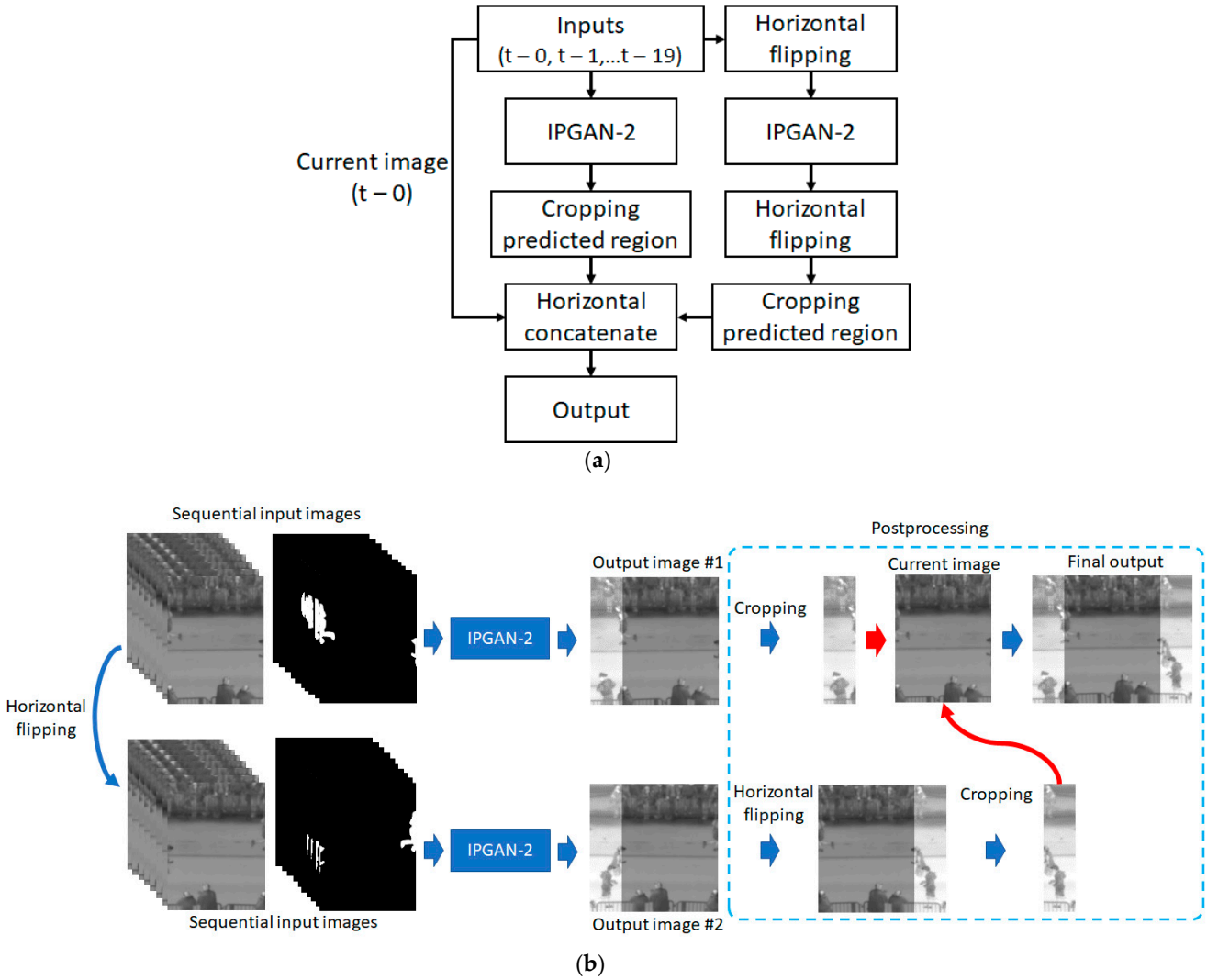


Figure 2. Flowchart of the proposed method. (**a**) Overall flowchart; (**b**) overall procedure with image examples.

Table 2. The proposed method detailed by using pseudo code.

Input thermal image with size of (120 × 160 × 1) = X_{t-0}	
Input binary image with size of (120 × 160 × 1) = Y_{t-0}	
Output thermal image with size of (160 × 160 × 1) = O_{t-0}	
Final output thermal image with size of (200 × 160 × 1) = Z_{t-0}	
IPGAN-2 model = *model*	
Horizontal image flipping = $flip$	
Image cropping = *crop*	
Algorithm procedure	**Output shape/dimension**
$A = [X_{t-0},\ X_{t-1}, \ldots\ X_{t-19},]$	120 × 160 × 20
$B = [Y_{t-0},\ Y_{t-1}, \ldots\ Y_{t-19},]$	120 × 160 × 20
$A' = flip(A)$	120 × 160 × 20
$B' = flip(B)$	120 × 160 × 20
$C = concatenate(A,\ B, axis = channel)$	120 × 160 × 40
$C' = concatenate(A',\ B', axis = channel)$	120 × 160 × 40
$O_{t-0} = model(C)$	160 × 160 × 1
$O'_{t-0} = model(C')$	160 × 160 × 1
$O''_{t-0} = flip(O'_{t-0})$	160 × 160 × 1
$R_{t-0} = crop(O_{t-0},\ [0\ :\ 39,\ 0\ :\ 159])$	40 × 160 × 1
$R'_{t-0} = crop(O''_{t-0},\ [120\ :\ 159,\ 0\ :\ 159])$	40 × 160 × 1
$C'' = concatenate(R_{t-0},\ X_{t-0},\ axis = horizontal)$	200 × 160 × 1
$Z_{t-0} = concatenate(C'',\ R'_{t-0},\ axis = horizontal)$	200 × 160 × 1

3.2. Proposed IPGAN-2 Model

As shown in Figure 2b, sequential thermal images (120 × 160 × 20 pixels) and sequential binary images (120 × 160 × 20 pixels) are used as input for the proposed IPGAN-2. The IPGAN-2 architecture is shown in Figure 3. The generator in Figure 3 includes the concatenate layer (L2 and L31), convolution blocks (L9, L22, and L26), encoder blocks (L3, L4, L5, L14, L15, and L16), residual blocks (L6–L8, L10–L13, L23–25, and L27–30), and convolution layers (L17, L20, L21, L32, and L33) in order. In the concatenate layer (L2), sequential images are applied with depth-wise concatenation to generate a multichannel single image (120 × 160 × 40), while in the concatenate layer (L31), feature maps are combined in the horizontal direction to generate a wide image. Furthermore, the discriminator includes convolution blocks (L1–L6) and a fully connected layer (L7) in order.

Specific details of the IPGAN-2 architecture are presented in Tables 3–8. In Tables 3–6, the filter size, stride, and padding are (3 × 3), (1 × 1), and (1 × 1), respectively. In Table 3, two filter numbers, 128 and 64, are used in conv_block_1–conv_block_3. In Table 7, the filter size, stride, and padding in conv_block_1–conv_block_3 are (3 × 3), (1 × 1), and (0 × 0), while the filter size, stride, and padding in conv_block_4–conv_block_6 are (3 × 3), (2 × 2), and (0 × 0), respectively. The layer types of Tables 3–8 are prelu (parametric rectified linear unit (relu)), lrelu (leaky relu), maxpool (max pooling operation), tanh (hyperbolic tangent activation function), res_block (residual block), encod_block (encoder block), conv2d (two-dimensional convolution layer), add (addition operation), conv_block (convolution block), dense (fully connected layer), concat (concatenate layer), and sigmoid (sigmoid activation function). Furthermore, reshape (L19) of Table 3 is a layer that reshapes input tensors into the given shape. As the input in Table 3, 20 sequential thermal images (120 × 160 × 1) and 20 sequential binary images (120 × 160 × 1) were used, as in Figure 3, and the output image was one image (160 × 160 × 1).

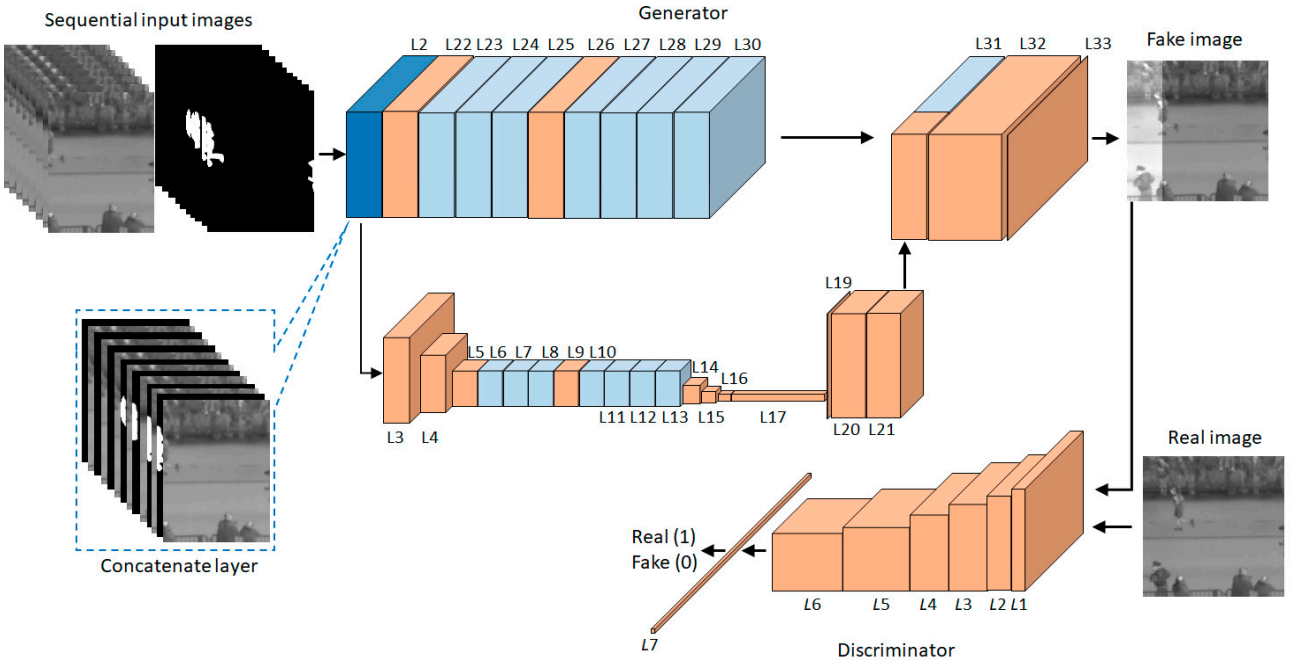


Figure 3. Example of the structure of proposed IPGAN-2.

Table 3. Description of the generator of IPGAN-2.

Layer Number	Layer Type	Number of Filters	Number of Parameters	Layer Connection (Connected to)
0	input_layers_1–20		0	input_1–20
1	input_layers_21–40		0	input_21–40
2	concat_1		0	input_layers_1–20 & input_layers_21–40
3	encod_block_1	64	48,640	concat_1
4	encod_block_2	64	73,984	encod_block_1
5	encod_block_3	64	73,984	encod_block_2
6	res_block_1	64	73,920	encod_block_3
7	res_block_2	64	73,920	res_block_1
8	res_block_3	64	73,920	res_block_2
9	conv_block_1	128/64	147,840	res_block_3
10	res_block_4	64	73,920	conv_block_1
11	res_block_5	64	73,920	res_block_4
12	res_block_6	64	73,920	res_block_5
13	res_block_7	64	73,920	res_block_6
14	encod_block_4	64	73,984	res_block_7
15	encod_block_5	64	73,984	encod_block_4
16	encod_block_6	64	73,984	encod_block_5
17	conv2d_1	3200	1,846,400	encod_block_6
18	prelu_1	3200	3200	
19	reshape		0	conv2d_1
20	conv2d_2	64	640	reshape
21	conv2d_3	64	36,928	conv2d_2
22	conv_block_2	128/64	97,152	concat_1
23	res_block_8	64	73,920	conv_block_2
24	res_block_9	64	73,920	res_block_8
25	res_block_10	64	73,920	res_block_9
26	conv_block_3	128/64	147,840	res_block_10
27	res_block_11	64	73,920	conv_block_3
28	res_block_12	64	73,920	res_block_11

Table 3. *Cont.*

Layer Number	Layer Type	Number of Filters	Number of Parameters	Layer Connection (Connected to)
29	res_block_13	64	73,920	res_block_12
30	res_block_14	64	73,920	res_block_13
31	concat_2		0	conv2d_3 & res_block_14
32	conv2d_4	256	147,712	concat_2
33	conv2d_5	1	2305	conv2d_4
34	tanh		0	conv2d_5
Total number of trainable parameters: 3,883,457				

Table 4. Description of an encoder block of the generator.

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

Table 5. 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	128	conv2d_1
3	conv2d_2	64	prelu_1
4	prelu_2	64	conv2d_2

Table 6. 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	64	conv2d_1
3	conv2d_2	64	prelu
4	add		conv2d_2 & input

Table 7. Description of the discriminator of IPGAN-2.

Layer Number	Layer Type	Number of Filters	Number of Parameters	Layer Connection (Connected to)
0	input layer		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 8. Description of a convolution block of the discriminator.

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

3.3. Postprocessing

The postprocessing method shown in Figure 4 was used in this study. As shown in Figure 3, the final output shown in Figure 4 is obtained by cropping and combining the predicted regions outside the FOV from the first output image obtained using IPGAN-2 and the second output image obtained by horizontally flipping sequential input images and using IPGAN-2. The reason for using the method in Figure 4, instead of performing image prediction for both sides of the FOV of the current image, is explained in Section 4.2 based on experimental results.

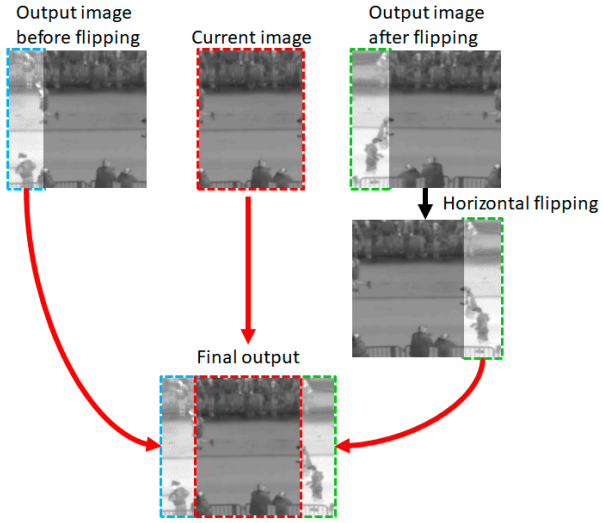


Figure 4. Example of the postprocessing.

3.4. Differences between IPGAN and Proposed IPGAN-2

In this section, the difference between the proposed method and a previous method [14] is explained in detail. These two methods have different architectures overall. In particular, the region in the image being predicted is different, and each step is designed differently. Table 9 shows the overall structure of the two methods in steps. Table 10 presents the advantage and disadvantage of the two methods.

Table 9. Comparison of overall structure of previous method [14] and proposed method.

Steps	Previous Method [14] (IPGAN)	Proposed Method (IPGAN-2)
Input	Original thermal image (85 × 170 × 1)	Original thermal image (120 × 160 × 1)
Preprocessing	Conversion of original thermal image to three-channel color thermal image with zero padding	Image binarization by using background subtraction and horizontal flipping
Network input	Three-channel color thermal image (170 × 170 × 3)	Original thermal image (120 × 160 × 1) and binary image (120 × 160 × 1)
Network output	Three-channel color thermal image (170 × 170 × 3)	One-channel thermal image (160 × 160 × 1)

Table 9. *Cont.*

Steps	Previous Method [14] (IPGAN)	Proposed Method (IPGAN-2)
Postprocessing	Image cropping (crop a small part of predicted region) and combining	Image cropping (crop the entire predicted region), horizontal flipping, and combining
Output	Three-channel color thermal image (105 × 170 × 3)	One-channel thermal image (200 × 160 × 1)

Table 10. Comparison of advantage and disadvantage of previous method [14] and proposed method.

Factors	Previous Method [14] (IPGAN)	Proposed Method (IPGAN-2)
Predicted region	Only left side	Left and right sides
Size of predicted region	Smaller (input of 85 × 170 to output of 105 × 170)	Larger (input of 120 × 160 to output of 200 × 160)
Error	Gray noise occurs over predicted region	No gray noise
Processing speed	Higher	Lower

3.5. Dataset and Experimental Setup

The experiment in this study was conducted using the marathon subdataset of the BU-TIV benchmark open thermal dataset [15]. The marathon subdataset was created for the purpose of multi-object tracking and includes various objects, such as pedestrians, cars, motorcycles, and bicycles. This dataset also consists of four videos (image sequences) with different sizes. Images in the dataset are provided in the image format of portable network graphics (PNG). Annotations for the object detection for the four sequences are provided. An FLIR SC800 camera (FLIR Systems, Inc., Wilsonville, OR, USA) was used to collect the dataset. The pixel value of a thermal image ranges between 3000 and 7000 units of uncalibrated temperature [15]. In this study, image sequences 1 and 2 of the marathon sub-dataset were used, and 3999 images (size of 1024 × 512 × 1, and pixel depth of 16 bits) were used. When training the proposed model, 3999 original images are cropped to create 19,995 images in a dataset (image size of 160 × 160 × 1, and pixel depth of 8 bits) to perform training and testing. The region in which pedestrians are running (region of interest (ROI) of the red boxes in Figure 5) in the original image was cropped. A ground-truth (*GT*) image (green boxes) and input images (blue boxes) were generated, as shown in Figure 5, by cropping the ROI into 160 × 160.

Our network is not aware of the scenes of testing cases. Various scenes have been used in our experiments. In the below Figure 5d,e, example images used in training and testing phases are presented. As shown in this figure, the images used in training phase are completely different from those in testing phase, and they were not cropped from same scene.

The experiment was conducted as two-fold cross validation. More specifically, half the total data were used for training, while the other half were used for testing (10% of the testing data were used as validation data, while the remaining 90% were used as testing data). The two datasets were switched for performing training and testing once again, and the average of the two testing accuracy values was set as the final accuracy.

Training and testing of the proposed algorithm were performed using a desktop computer. The desktop computer was equipped with Intel core i7-6700 CPU@3.40GHz, a Nvidia GeForce GTX TITAN X graphics processing unit (GPU) card [33], and random-access memory (RAM) of 32 GB. The proposed model and algorithm were implemented with OpenCV library (version 4.3.0) [34], Python (version 3.5.4) [35], and the Keras application programming interface (API) (version 2.1.6-tf) with a TensorFlow backend engine (version 1.9.0) [36].

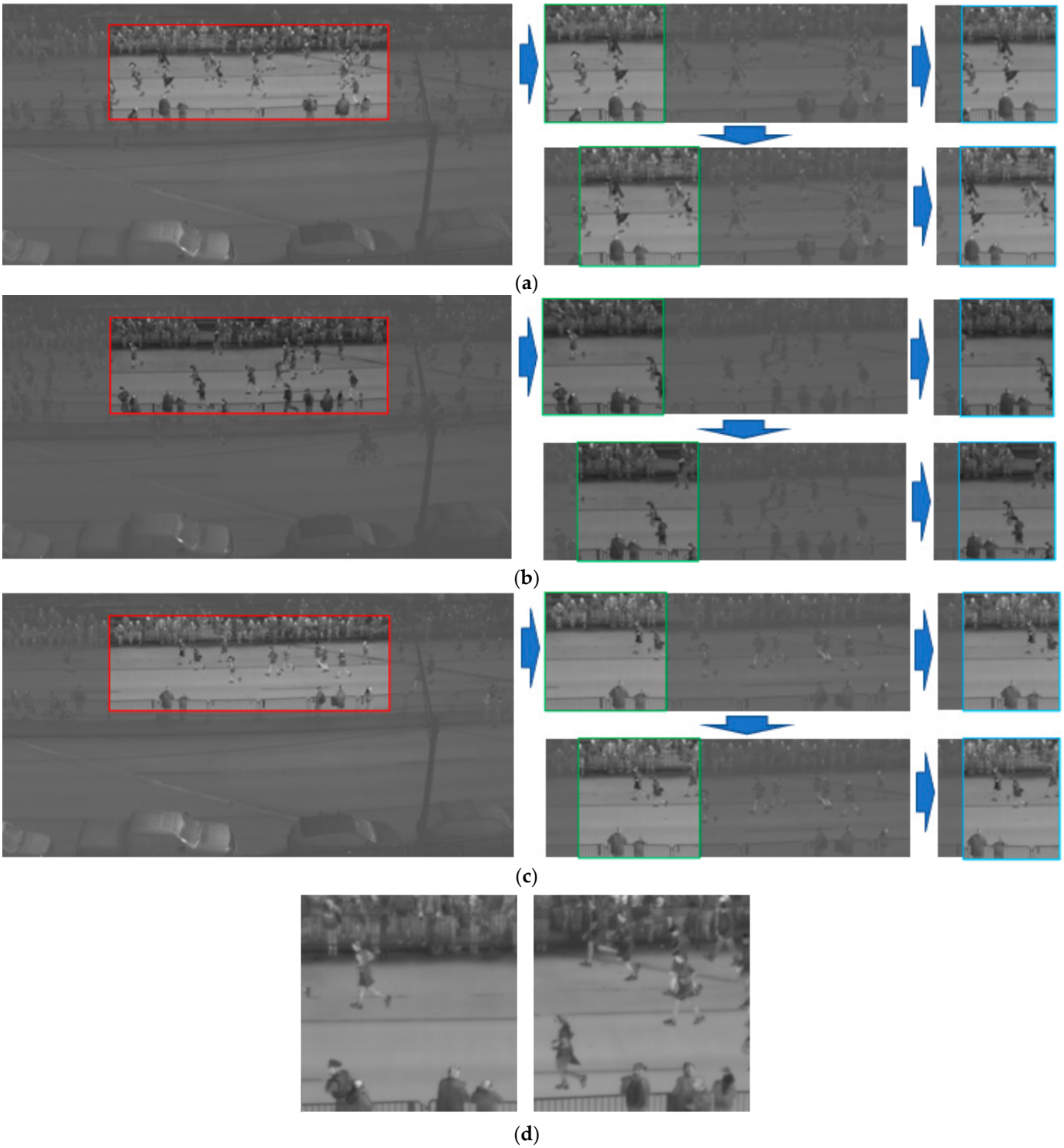
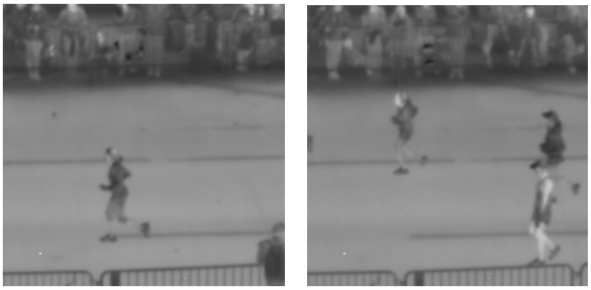


Figure 5. *Cont*.



(**e**)

Figure 5. Example images of dataset: the left images in red boxes show the cropped ROI, the middle images in green boxes represent the *GT* images, and the right images in blue boxes show the input images to model for image prediction. (**a**–**c**) the procedure of making the dataset; (**d**) example images used in training phase; (**e**) example images used in testing phase.

4. Results

This section is divided into four subsections on training, testing, comparisons, and processing time to explain the experimental results. In Section 4.1, the hyperparameters and training loss used in the training step are explained. In Section 4.2, the results obtained through the ablation study are compared. In Section 4.3, the results obtained using the proposed method and the state-of-the-art methods are compared. In Section 4.4, additional experiments using different datasets (Casia thermal image dataset C and BU-TIV marathon thermal image dataset) for training and testing are conducted. Finally, in Section 4.5, the processing time was measured for each component.

4.1. Training

The proposed IPGAN-2 was trained as follows. The batch size, training iteration, and learning rate in IPGAN-2 were 1, 483,581, and 0.001, respectively. Moreover, binary cross-entropy loss was used as generator loss and discriminator loss, and adaptive moment estimation (Adam) [37] was used as an optimizer. More detailed information about the search space and selected hyperparameter values is provided in Table 11. Hyperparameters were selected based on the best accuracies of human segmentation explained in Section 4.3 using the training data. Forty sequential images (20 thermal images and 20 binary images) of 120 × 160 × 1 pixels were used for all training and testing methods. Figure 6a shows the training loss curves of IPGAN-2 per iteration, while Figure 6b shows the validation loss curves of IPGAN-2 per iteration. All results converged as the iterations increased; in particular, Figure 6b shows that IPGAN-2 was sufficiently trained without being overfitted by the training data.

Table 11. Search space and selected values of hyperparameters.

Parameters	Weight Decay (Weight Regularization L2)	Loss	Kernel Initializer	Bias Initializer	Optimizer	Learning Rate	Beta_1	Beta_2	Epsilon	Iterations	Batch Size
Search Space	[0.001, 0.01, 0.1]	[“binary cross-entropy loss”, “mse”,“VGG-19 loss”]	“glorot uniform”	“zeros”	[“SGD”, “adam”]	[0.0001, 0.001, 0.01, 0.1]	[0.7, 0.8, 0.9]	[0.8, 0.9, 0.999]	[1×10^{-9}, 1×10^{-8}, 1×10^{-7}]	[1~500 K]	[1,4,8]
Selected Value	0.01	“binary cross-entropy loss”	“glorot uniform”	“zeros”	“adam”	0.001	0.9	0.999	1×10^{-8}	483,581	1

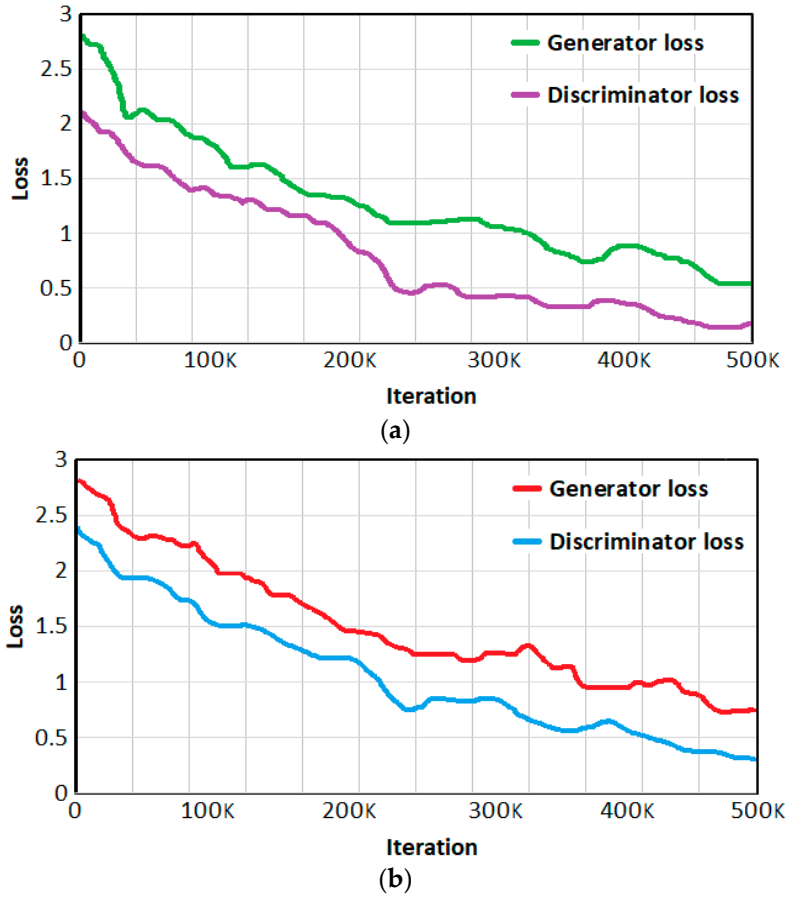


Figure 6. Loss curves of IPGAN-2 with (**a**) training data and (**b**) validation data.

4.2. Testing (Ablation Study)

In this subsection, the results of several ablation studies using the proposed method are explained. Identical datasets and six GAN structures were used for the experiment. The accuracy of image prediction was measured in terms of the similarity between the output image and the *GT* image. The accuracy of image prediction was measured using three types of metric in Equations (1)–(3). In Equation (1), *R* and *C* represent the number of rows (height) and columns (width) of the image matrix, respectively. In Equations (1) and (3), *Res* and *GT* refer to result image and *GT* image, respectively. In Equation (2), PSNR is the peak signal-to-noise ratio [38]. In the structural similarity index measure (SSIM) [39] equation, m_{GT} and S_{GT} are the mean and standard deviation of the pixel values of a *GT* image, respectively, m_{Res} and S_{Res} are the mean and standard deviation of the pixel values of the result image, respectively, S_{ResGT} is the covariance of the two images, and $St1$ and $St2$ represent positive constants to make the denominator nonzero.

$$\mathrm{MSE} = \frac{\left(\sqrt{\sum_{i=1}^{R}\sum_{j=1}^{C}(GT(j,i) - Res(j,i))^2}\right)^2}{RC} \tag{1}$$

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

$$\mathrm{SSIM} = \frac{(2m_{Res}m_{GT} + St1)(2S_{ResGT} + St2)}{({m_{Res}}^2 + {m_{GT}}^2 + St1)({S_{Res}}^2 + {S_{GT}}^2 + St2)} \tag{3}$$

In addition, the accuracy of human detection was measured based on true positive rate (TPR) (#TP/(#TP+#FN)) and positive predictive value (PPV) (#TP/(#TP+#FP)) [40]

and using accuracy (ACC) [40], F1 score (F1) [41], and intersection over union (IoU) [40], which are expressed in Equations (4)–(6). Here, TP, FP, FN, and TN refer to true positive, false positive, false negative, and true negative, respectively. Positive and negative signify the pixels detected in the *GT* image and the pixels not detected in the *GT* image. TP refers to a case where positive pixels were accurately detected, while TN refers to a case where negative pixels were inaccurately detected. FP refers to a case where negative pixels were falsely detected as positive pixels, while FN refers to a case where positive pixels were falsely detected as negative pixels. Here, the symbol # indicates "the number of".

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

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

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

Six methods were comparatively examined through ablation studies. In Figure 7, the t-th image I_t was set as the input image (at the far left) among the 20 sequential input thermal images.

In Figure 7a, the *GT* image (at the far right) included the images on both sides (left and right of the I_t image) of the I_t image (left image). Specifically, the images on both sides ($40 \times 160 \times 1$ and $40 \times 160 \times 1$) of the I_t image ($80 \times 160 \times 1$) were predicted (pred2reg), as shown in Figure 7a. After the images on both sides of the I_t image are predicted, the regions predicted as the last are cropped to be combined with the current image I_t, as shown in Figure 1. In this experiment, sequential original images ($80 \times 160 \times 20$) and sequential binary images ($80 \times 160 \times 20$) were used for prediction. However, the output image obtained thereby ($160 \times 160 \times 1$) differs significantly from the *GT* image as in the output image O_t (middle image) in Figure 7a.

Unlike pred2reg in Figure 7b, the experiment was conducted to predict the entire O_t image ($160 \times 160 \times 1$) (predWholeIm) from sequential input images (thermal image ($80 \times 160 \times 20$) and binary image ($80 \times 160 \times 20$)).

In Figure 7c, after extracting feature maps from sequential binary images ($80 \times 160 \times 20$) and sequential thermal images ($80 \times 160 \times 20$) through a two-channel convolution structure, the feature maps ($160 \times 160 \times 64$ and $160 \times 160 \times 64$) obtained from sequential binary images and original sequential images were combined along the depth axis (2-chanPred) in the last convolution layer.

The following three methods were utilized as follows to improve the accuracy. In Figure 7d, two images ($40 \times 160 \times 1$ and $40 \times 160 \times 1$) on both sides of the current image were predicted (singImPred) using one thermal image ($80 \times 160 \times 1$) and one binary image ($80 \times 160 \times 1$) rather than sequential images. The output image ($160 \times 160 \times 1$) obtained through singImPred had a clearer background than the images obtained through previously mentioned methods but had a poorer performance in human prediction.

In Figure 7e, a method (seq&sing) was utilized where sequential images (thermal images ($80 \times 160 \times 20$) and binary images ($80 \times 160 \times 20$)) were used to predict the image on the left of the current image, while a current image (thermal image ($80 \times 160 \times 1$) and binary image ($80 \times 160 \times 1$)) was used to predict the image on the right of the current image. A two-channel convolution structure was applied in the experiment, but the predicted images on both sides were combined with the current image, as in pred2reg in the last concatenate layer. In this method, the left predicted image (using sequential images) has a higher result than the right predicted image (using a single image).

Figure 7f shows the result obtained through a method using a three-channel image (pred3-chan [14]). The final result of this method has the removed part that was not predicted well in the image generated through the GAN structure. For comparing the output images with the *GT* image, as with other methods, Figure 7f shows the result before

removing the parts that were not predicted well for the comparison. However, as explained in pred3-chan [14], it is difficult to obtain a result similar to the *GT* because of gray noise.

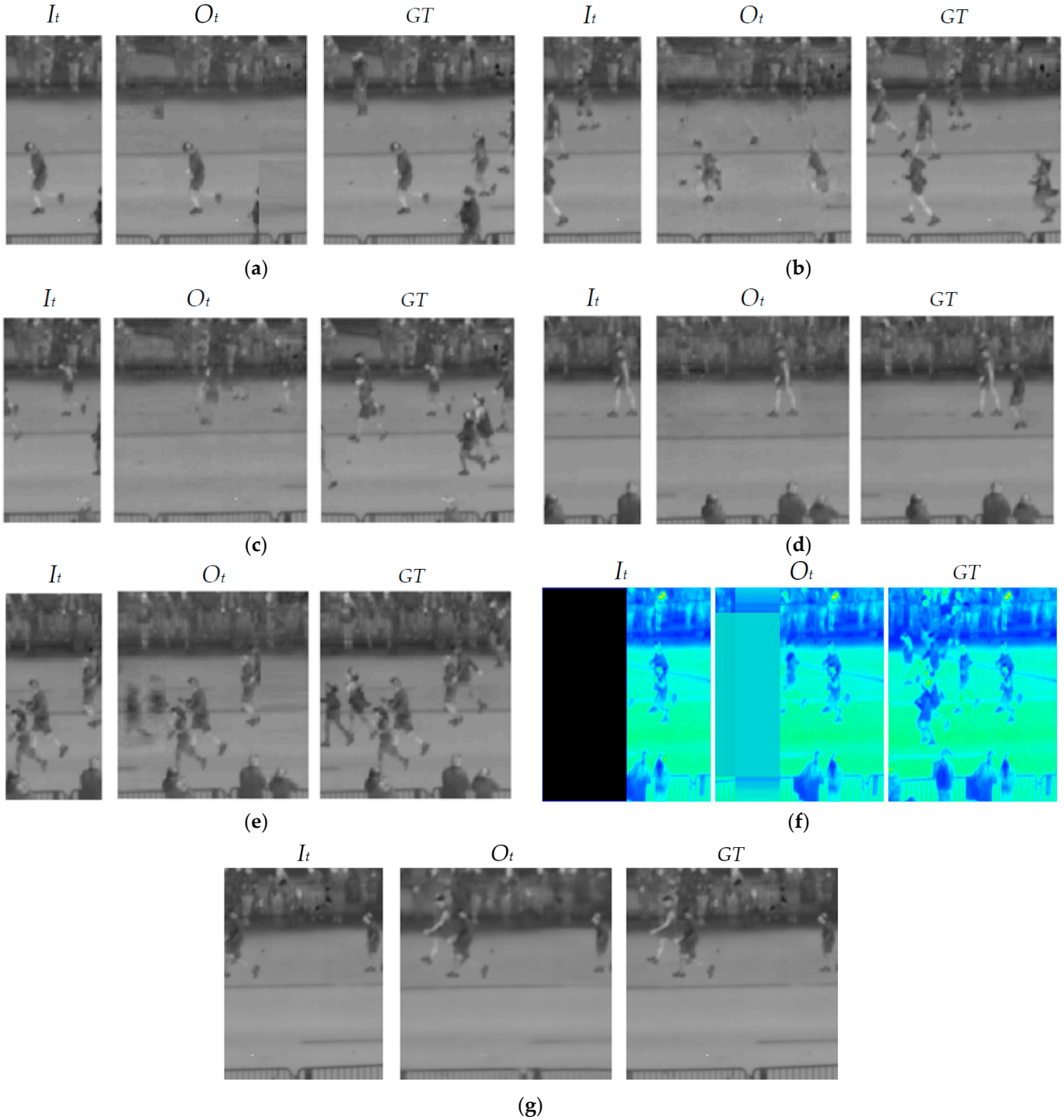


Figure 7. Examples of result images obtained by various methods: from left to right, input, output, and *GT* images obtained by (**a**) pred2reg, (**b**) predWholeIm, (**c**) 2-chanPred, (**d**) singImPred, (**e**) seq&sing, (**f**) pred3-chan [14], and (**g**) predLreg.

Therefore, only the left image (40 × 160 × 1) was predicted (predLreg) among the sequential input images (thermal images (120 × 160 × 20) and binary images (120 × 160 × 20)) in Figure 7g. When conducting this experiment, the feature maps (40 × 160 × 64) that were extracted similarly to L31 in Figure 3 were combined with the feature maps (120 ×

160 × 64) along the horizontal axis to obtain the feature maps (160 × 160 × 128), and the final output image (160 × 160 × 1) is obtained.

In the next experiment, the PSNR and SSIM of the *GT* image and the output image generated by each method were compared, as shown in Table 12. As Figure 7 and Table 12 show, predLreg had the highest PSNR and SSIM accuracy among the methods. Therefore, this study used predLreg to generate the images on both sides of the current image through flipping, cropping, and combining operations, as shown Figures 2 and 4. Figure 8a shows the image generated by predLreg, while Figure 8b shows the image generated by predLRreg (proposed method). Figure 9 shows the examples of various images generated by predLRreg (proposed method).

Table 12. Comparisons of various image prediction methods.

Methods	PSNR	SSIM
pred2reg	19.450	0.8156
predWholeIm	14.501	0.6395
2-chanPred	15.261	0.6121
singImPred	19.214	0.8132
seq&sing	21.340	0.8413
pred3-chan [14]	24.927	0.8403
predLreg	26.592	0.9581

(**a**)

(**b**)

Figure 8. Examples of result images obtained by predLreg and predLRreg (proposed method): from left to right, input, *GT*, and output images obtained by (**a**) predLreg and (**b**) predLRreg (proposed method).

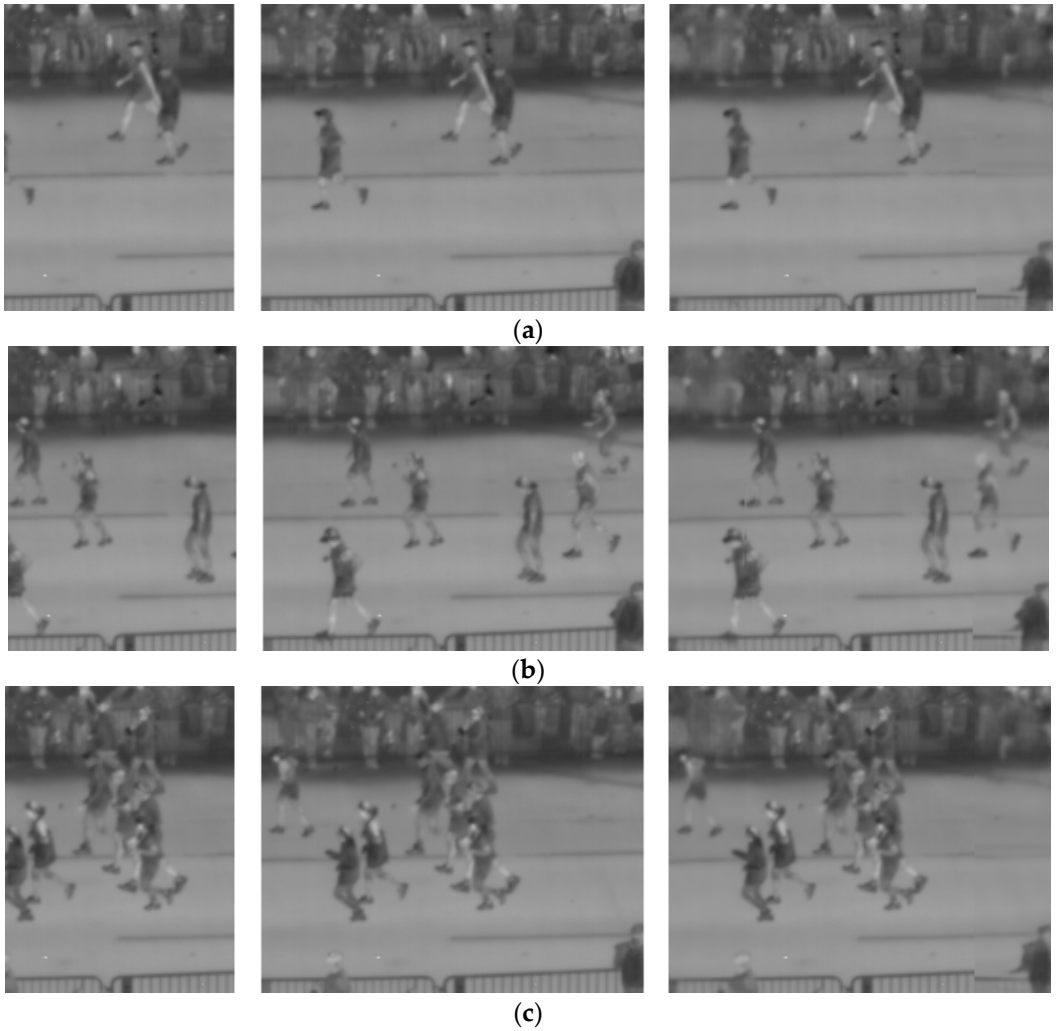


Figure 9. Examples of result images obtained by predLRreg (proposed method): in (**a**–**c**), from left to right, original images, *GT* images, and predicted (output) images.

In the next experiment, the results of detecting humans in the original input image and *GT* image were compared with the result of detecting humans in the image predicted by the proposed method for examining the efficiency of the proposed method. For a fair experiment, an identical Mask R-CNN [42] was used for the two methods during human seg-mentation. Figure 10 shows the result of human segmentation using Mask R-CNN as mask images. As shown in Figure 10, the result of human segmentation in the *GT* image and the result of human segmentation in the image predicted by the proposed method are quite similar. The segmentation result in the predicted image is closer to the segmentation result in the *GT* image than in the original input image. In Table 13, the detection accuracies measured between the result images of object segmentation with original images (or-detect) and the result images of object segmentation with *GT* images are shown. Furthermore, the detection accuracies were measured and compared between the resulting images of object segmentation with images predicted by the proposed method (pred-detect) and the resulting images of object segmentation with *GT* images. As shown in Table 13, pred-detect was more accurate than or-detect, indicating that the result is closer to the segmentation in the *GT* image when the image predicted by the proposed method was used than when the original input image was used.

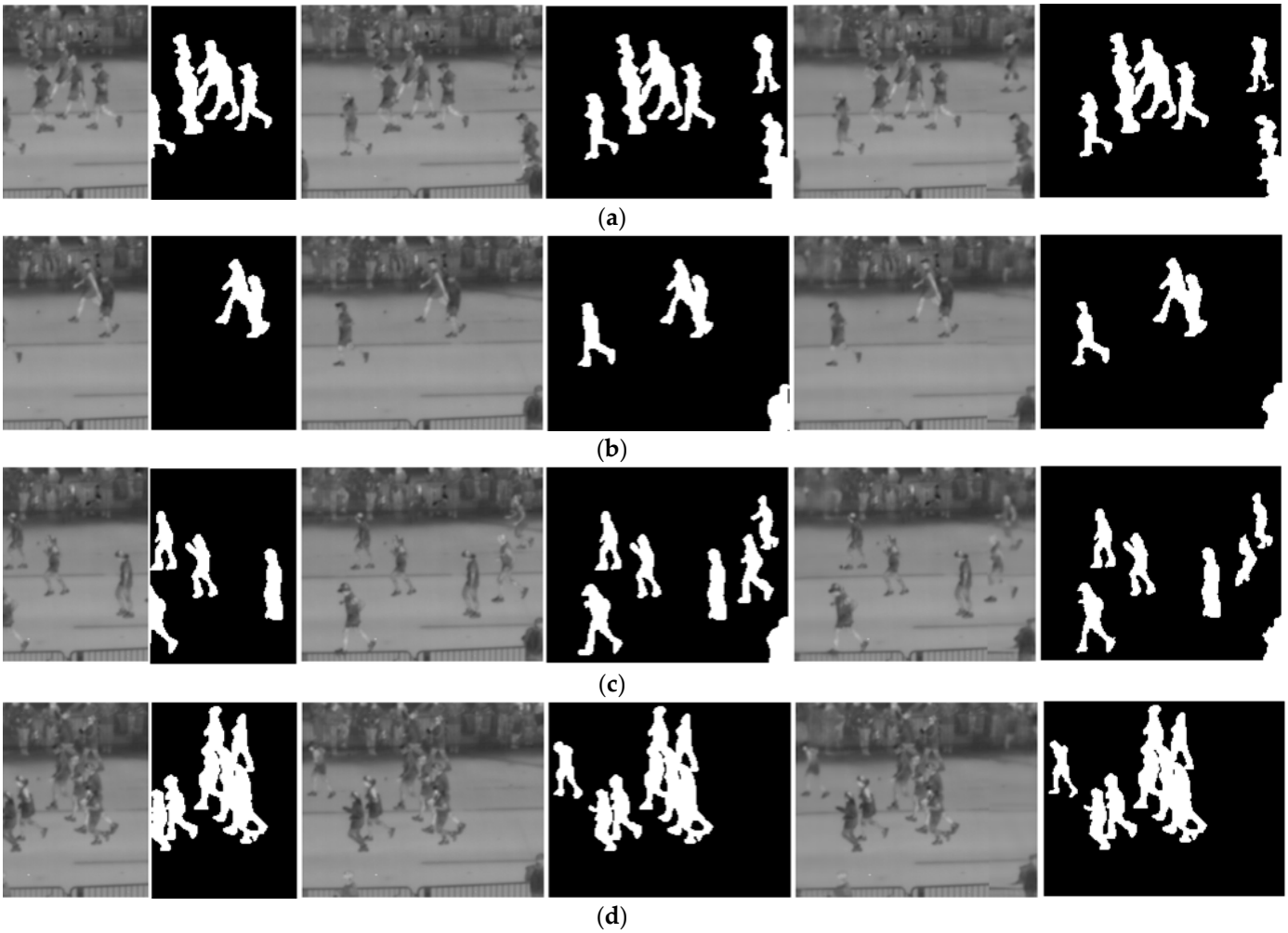


Figure 10. Examples of segmentation results before and after image prediction. (**a**–**d**) From left to right, the original input images, results with original input images, *GT* images, results with *GT* images, images predicted by the proposed method, and results with predicted images.

Table 13. Comparisons of segmentation accuracies with original images (or-detect) and with images predicted by the proposed method (pred-detect).

Methods	TPR	PPV	F1	ACC	IoU
or-detect	0.601	0.613	0.606	0.71	0.483
pred-detect (proposed)	0.887	0.847	0.866	0.914	0.730

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

In this subsection, the proposed method is compared with state-of-the-art methods. When measuring accuracy, the output image obtained by the proposed method is compared based on the similarity to the *GT* image. In Table 14, the conventional image prediction [26], image region prediction [14], and inpainting [17,19,21] methods were compared with the proposed IPGAN-2-based image prediction method. Figures 11–13 show the comparisons of the images obtained by all the methods. For a fair performance evaluation, previous methods [17,19,21], which typically use one image, were applied with sequential images (thermal images (120 × 160 × 20) and binary images (120 × 160 × 20)), as in the proposed method; accordingly, the input layers of these methods [17,19,21] were changed to the layers 0, 1, and 2 of the proposed method shown in Table 3. To evaluate the performance of

pred3-chan [14] against other methods fairly, the result before removing the parts that were not predicted well was used for the comparison, as explained in Section 4.2—see Figure 7f. Flipping, cropping, and combining were performed, as in Figure 2b, to predict the images on both sides, and an image of 200 × 160 × 1 was generated for comparison. As shown in Figures 11–13 and Table 14, the proposed method produced superior results to those of the state-of-the-art methods. The proposed predLreg method in Table 12 generated only the image to the left of the current image, while it generated left and right region images in Table 14; thus, the PSNR and SSIM values of the proposed method in Table 14 differ from the PSNR and SSIM values of the proposed predLreg in Table 12.

Table 14. Comparisons of accuracies of image prediction and human segmentation by the proposed method with those of the state-of-the-art methods.

Methods	Image Prediction		Mask R-CNN				
	PSNR	SSIM	TPR	PPV	F1	ACC	IoU
Haziq et al. [26]	22.843	0.8917	0.801	0.654	0.720	0.904	0.521
Liu et al. [17]	20.557	0.8454	0.638	0.626	0.631	0.864	0.432
Shin et al. [19]	22.181	0.8781	0.687	0.631	0.657	0.866	0.502
Nazeri et al. [21]	22.112	0.8724	0.651	0.672	0.661	0.890	0.514
pred3-chan [14]	25.146	0.8711	0.792	0.714	0.753	0.901	0.536
Proposed method	26.018	0.9437	0.887	0.847	0.866	0.914	0.730

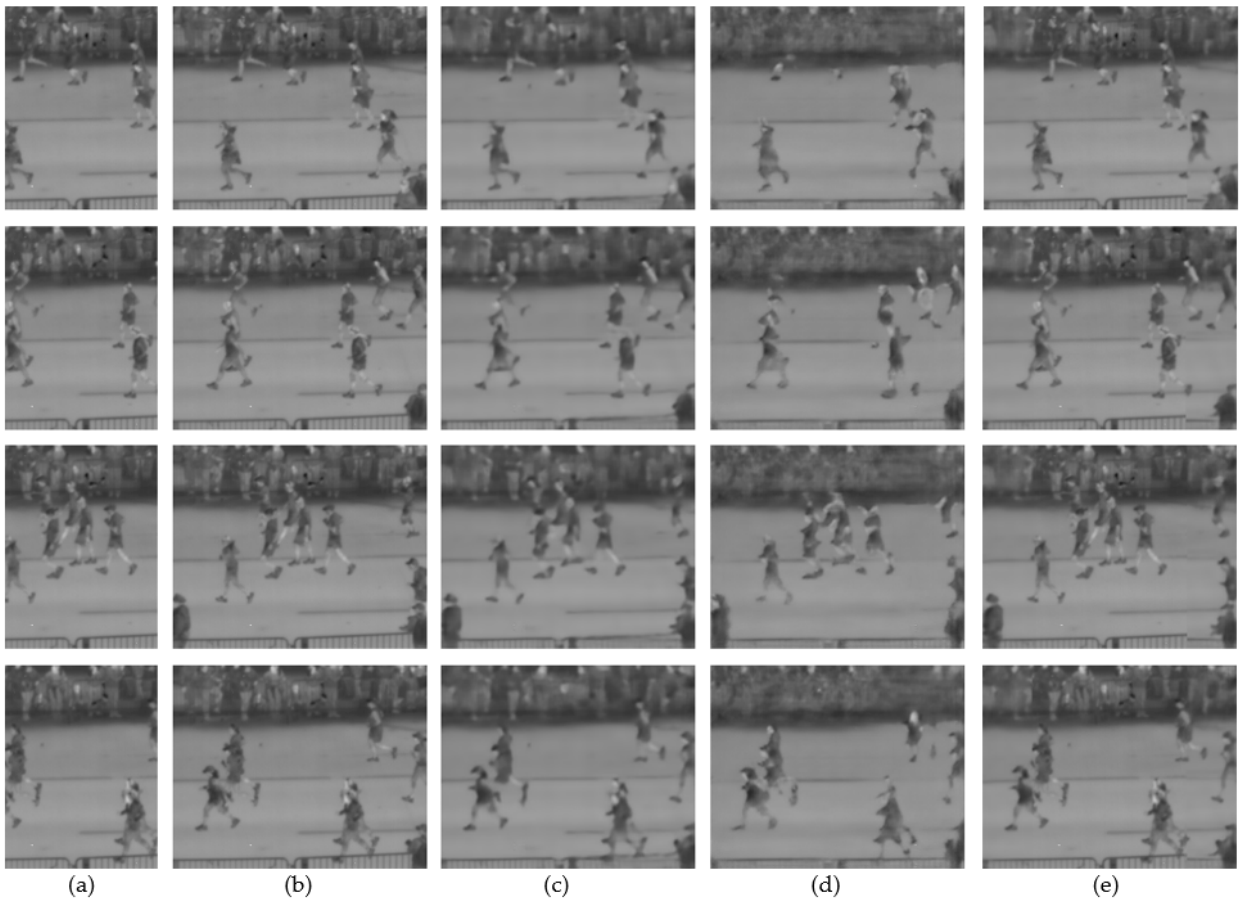


Figure 11. Comparisons of original images, *GT* images, the prediction results obtained by the state-of-the-art methods, and the proposed method: (**a**) original images, (**b**) *GT* images, and images predicted by (**c**) Haziq et al. [26], (**d**) Liu et al. [17], and (**e**) the proposed method.

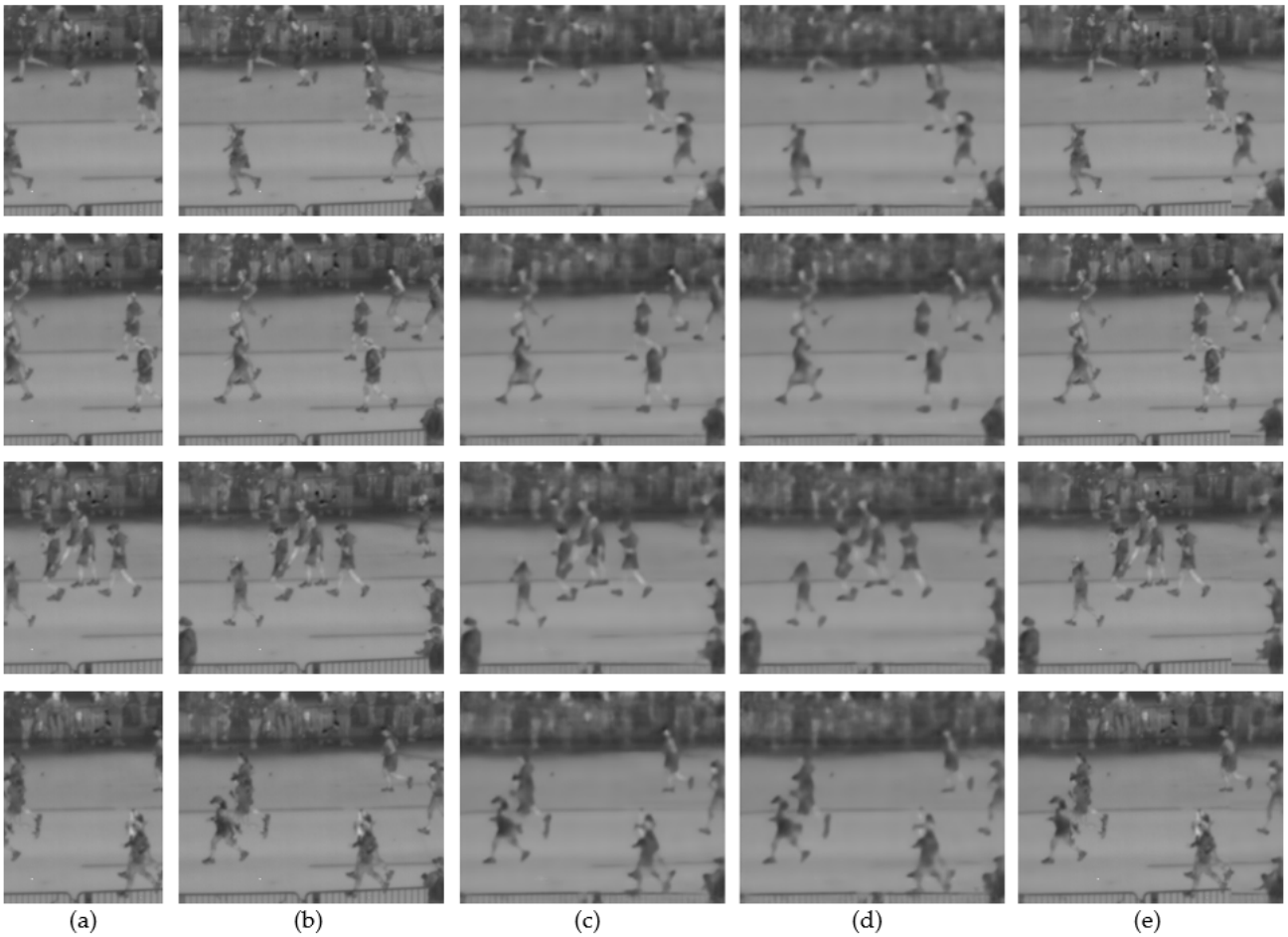


Figure 12. Comparisons of original images, *GT* images, and the prediction results obtained by the state-of-the-art methods and the proposed method: (**a**) original images, (**b**) *GT* images, and images predicted by (**c**) Shin et al. [19], (**d**) Nazeri et al. [21], and (**e**) the proposed method.

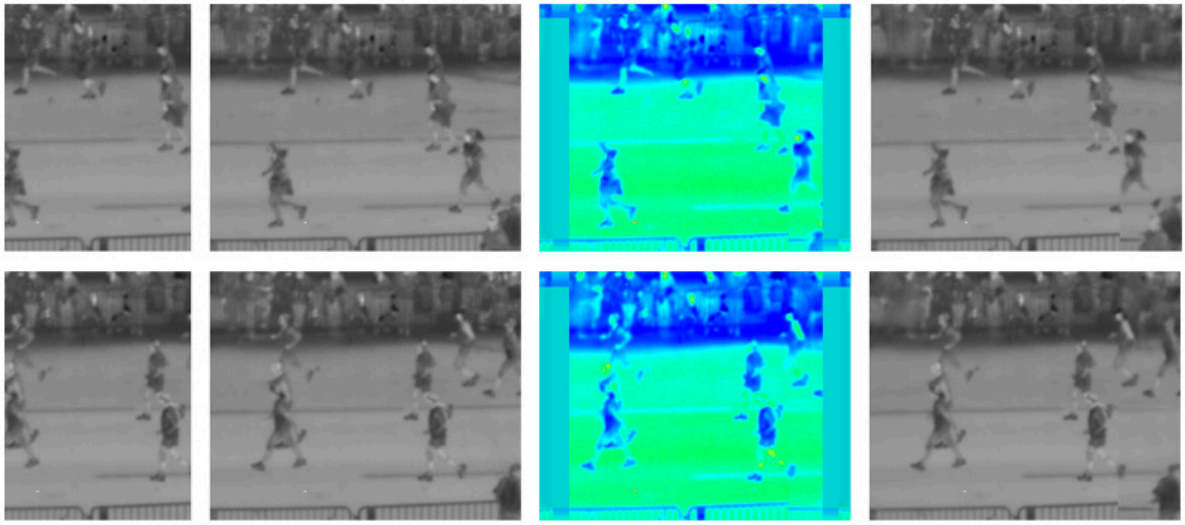


Figure 13. *Cont.*

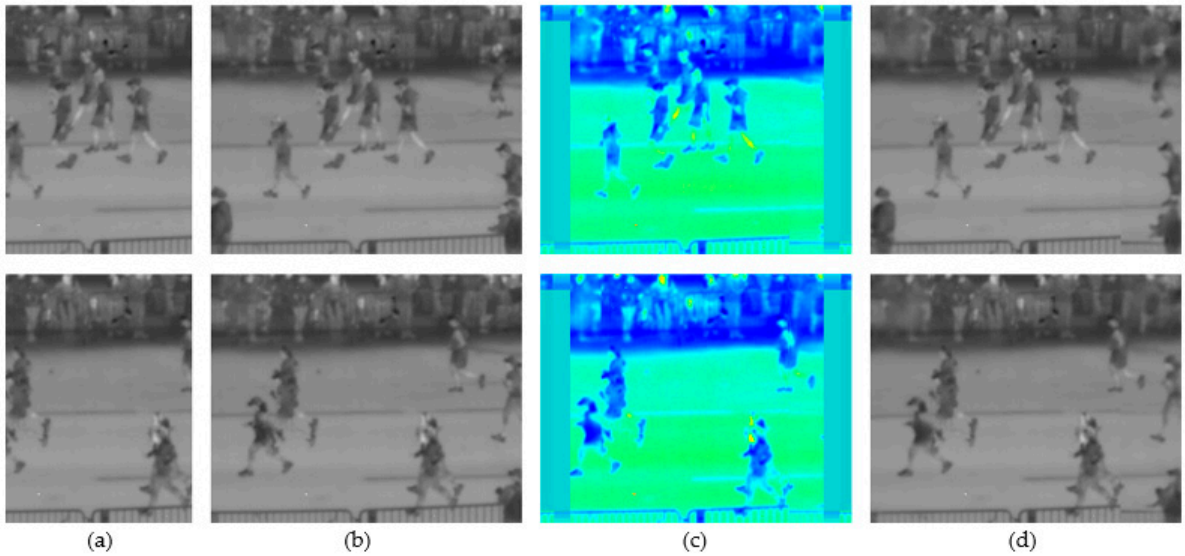


Figure 13. Comparisons of original images, *GT* images, and the prediction results obtained by the state-of-the-art methods and the proposed method: (**a**) original images, (**b**) *GT* images, and images predicted by (**c**) pred3-chan [14] and (**d**) the proposed method.

For the subsequent experiment, the performance was compared with Mask R-CNN human segmentation. The segmentation accuracy and output images are compared in Table 14 and Figures 14–16. Identical Mask R-CNN [42] based human segmentation was applied for all methods for a fair evaluation. As shown in Table 14 and Figures 14–16, the human segmentation performance was superior when the images obtained by the proposed method were used than when the images obtained by the state-of-the-art methods were used.

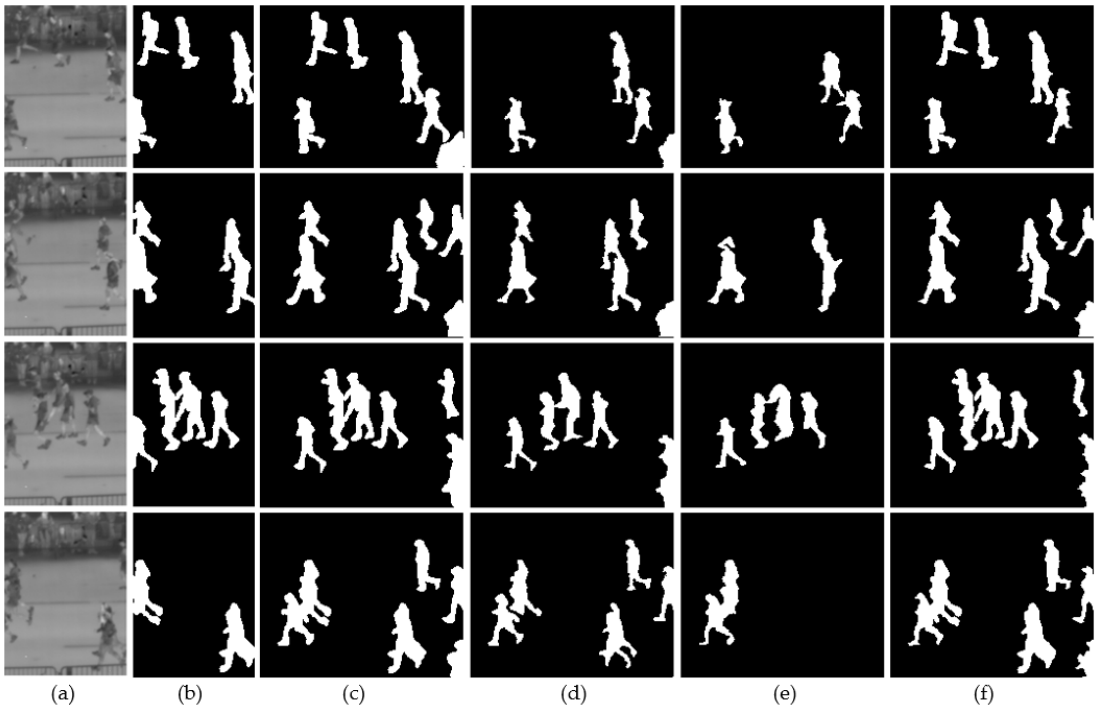


Figure 14. Comparisons of detection results using original images, *GT* images, and the predicted images obtained by the state-of-the-art methods and the proposed method: (**a**) original images; detection results using (**b**) original images and (**c**) *GT* images of Figures 11b, 12b and 13b; and the images predicted by (**d**) Haziq et al. [26], (**e**) Liu et al. [17], and (**f**) the proposed method.

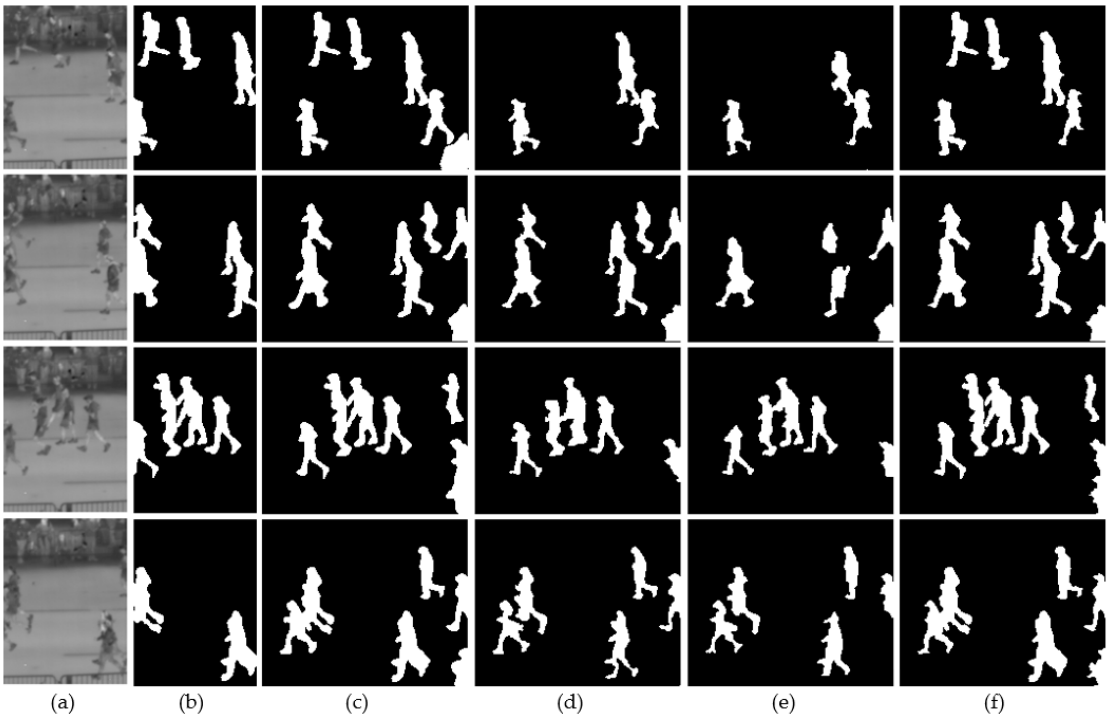


Figure 15. Comparisons of detection results using original images, *GT* images, and the predicted images obtained by the state-of-the-art methods and the proposed method: (**a**) original images; detection results using (**b**) original images and (**c**) *GT* images of Figures 11b, 12b and 13b; and the images predicted by (**d**) Shin et al. [19], (**e**) Nazeri et al. [21], and (**f**) the proposed method.

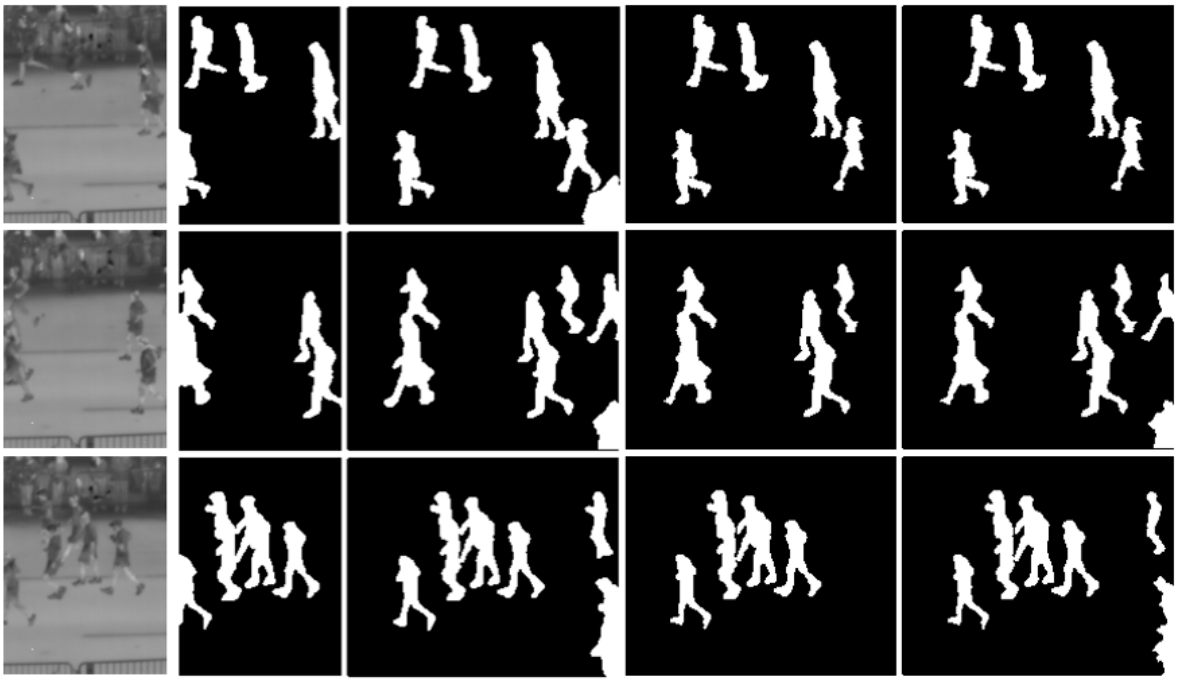


Figure 16. *Cont.*

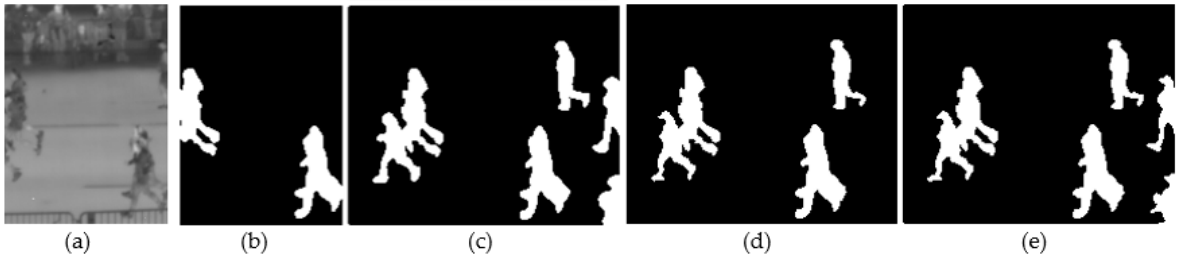


Figure 16. Comparisons of detection results using original images, *GT* images, and the predicted images obtained by the state-of-the-art methods and the proposed method: (**a**) original images; detection results using (**b**) original images and (**c**) *GT* images of Figures 11b, 12b and 13b; and the images predicted by (**d**) pred3-chan [14] and (**e**) the proposed method.

4.4. Experiments Using Different Datasets (Casia Dataset C and BU-TIV Marathon Dataset) for Training and Testing

In this section, additional experiments were conducted using Casia thermal image dataset C [43] and BU-TIV marathon dataset [15]. These two databases were acquired from different cameras, and they include different angle images with totally different backgrounds and foregrounds. The Casia dataset C includes thermal videos captured in outdoor environment using a thermal camera during nighttime. In addition, the Casia dataset C was captured under four walking conditions, namely slow, normal, fast walking, and normal walking with a bag. In the dataset, data of various humans including men and women are included. The total number of subjects and image sequences in this dataset are 153 and 1530, respectively. The pixel value of a thermal image ranges between 0 and 255. In this experiment, 2000 images (size of 320 $\times$ 240 $\times$ 1, and pixel depth of 8 bits) were used. For experiments, 2000 images of Casia dataset C and 2000 images of BU-TIV marathon dataset were used for training and testing. In addition, because the size of humans (height = 115 and width = 45 pixels) in images of Casia dataset C is much greater than that (height = 50 and width = 15 pixels) in images of BU-TIV marathon dataset, the images of Casia dataset C and images of BU-TIV marathon dataset were resized to make the size of humans in both datasets similar as shown in Figure 17. The experiments were conducted by two-fold cross validation. More specifically, the Casia dataset C was used for training, while the BU-TIV marathon dataset was used for testing in the fold-1 as shown in Figure 17. Then, the two datasets were switched for performing training and testing once again to perform two-fold cross validation. In Table 15, the results of fold-1 (train data = Casia dataset C, test data = BU-TIV dataset), fold-2 (train data = BU-TIV dataset, test data = Casia dataset C), and the average of fold-1 and fold-2 are presented. In addition, image prediction and human segmentation results are presented in Figures 18 and 19, and Figures 20 and 21, respectively. As shown in Table 15 and Figures 18–21, we confirm that our method can be adopted to the case of using two different databases for training and testing.

Figure 17. Example images of datasets in fold-1. (**a**) Example images used in training phase; (**b**) example images used in testing phase.

Table 15. Accuracies of image prediction and human segmentation by the proposed method using two different datasets.

Results	Image Prediction		Mask R-CNN				
	PSNR	**SSIM**	**TPR**	**PPV**	**F1**	**ACC**	**IoU**
Fold-1	24.984	0.9211	0.851	0.821	0.835	0.895	0.725
Fold-2	24.028	0.9064	0.835	0.802	0.818	0.889	0.705
Average	24.506	0.9137	0.843	0.811	0.826	0.892	0.715

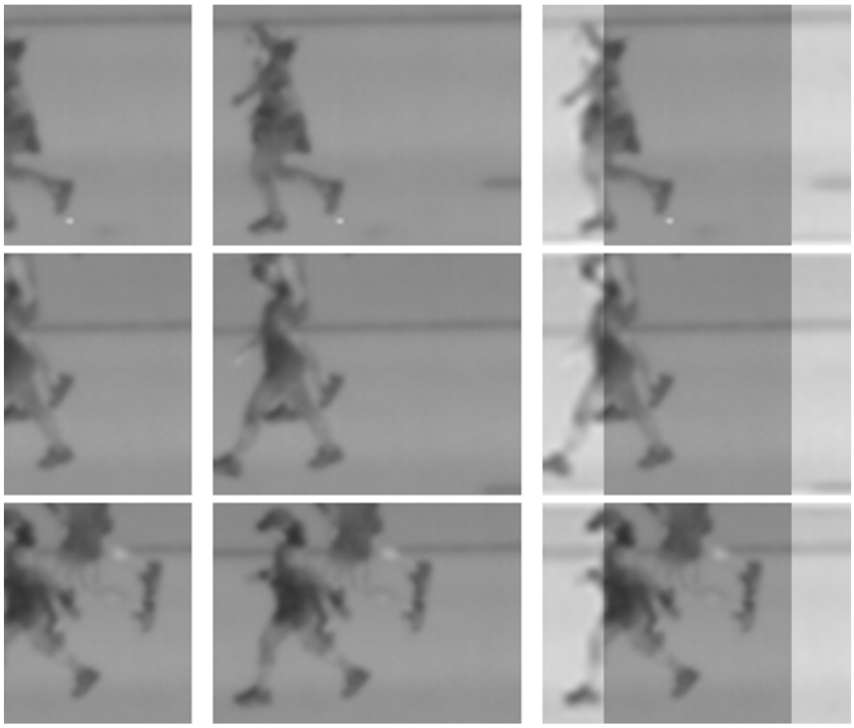


Figure 18. Example images of image prediction from fold-1: (**a**) original images; (**b**) *GT* images; (**c**) images predicted by the proposed method.

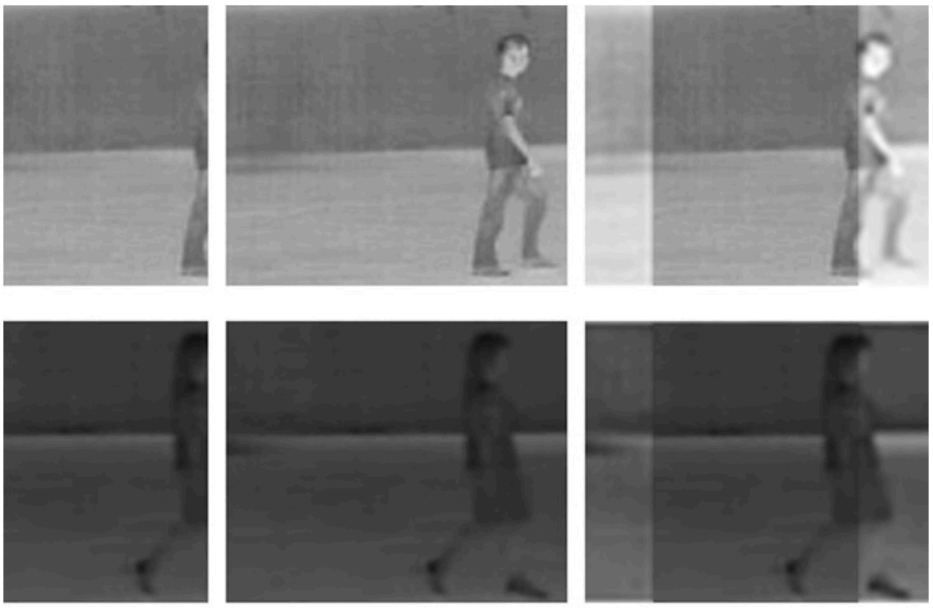


Figure 19. *Cont.*

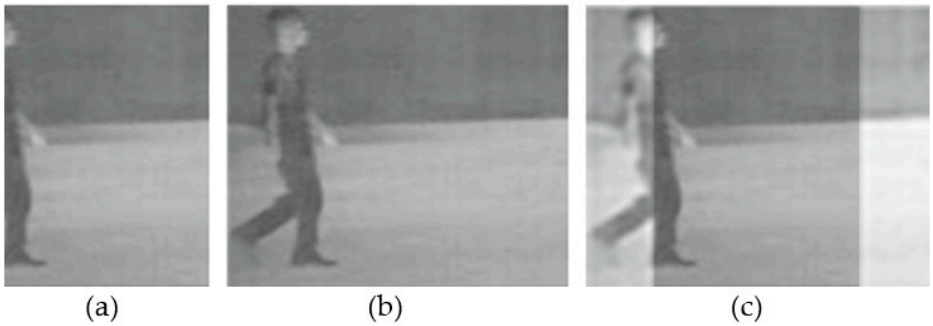


Figure 19. Example images of image prediction from fold-2: (**a**) original images; (**b**) *GT* images; (**c**) images predicted by the proposed method.

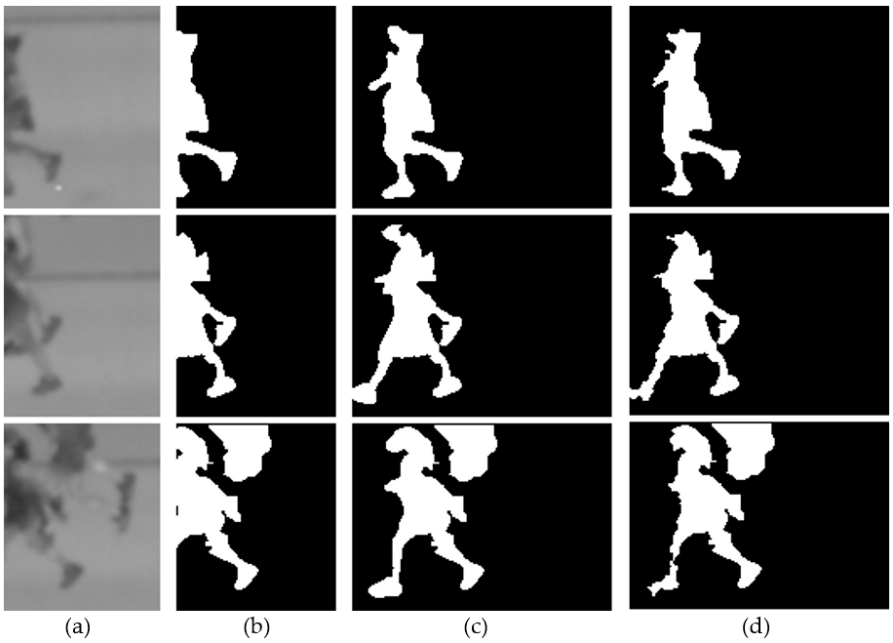


Figure 20. Example images of human segmentation from fold-1: (**a**) original images; segmentation results using (**b**) original images and (**c**) *GT* images; and the images predicted by (**d**) the proposed method.

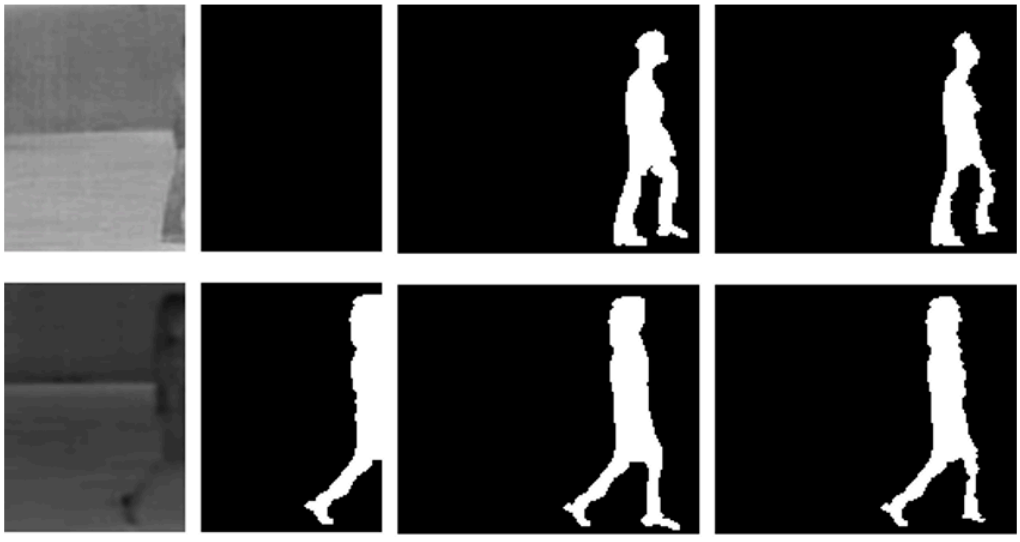


Figure 21. *Cont*.

Figure 21. Example images by human segmentation from fold-2: (**a**) original images; segmentation results using (**b**) original images and (**c**) *GT* images; and the images predicted by (**d**) the proposed method.

4.5. Processing Time

The processing times of the proposed image prediction method and the human segmentation method are shown in Table 16. Each component of the proposed method (as shown in Figure 2b) is shown in Table 16 as well. The processing time was measured in the environments described in Section 3.5.

Table 16. Processing time of the proposed method per image (unit: ms).

Methods	Component	Processing Time
Image prediction	IPGAN-2 (before flipping)	48.4
	IPGAN-2 (after flipping)	48.4
	Postprocessing	0.01
Human segmentation	Mask R-CNN	54.1
Total		150.91

As shown in Table 16, the processing time of the Mask R-CNN is higher than other components. The frame rate of the proposed image prediction method is approximately 10.33 frames per second (fps) (1000/(48.4 + 48.4 + 0.01)). The total frame rate including both image prediction and the human segmentation method is approximately 6.63 fps (1000/150.91). The time and space complexities of the proposed method are $O(2^n)$ and O(n) in training phase, respectively. They are O(n) and O(1) in testing phase, respectively.

5. Discussion

As shown in Figures 11–13, the persons in the predicted region of an image may be poorly segmented compared with the persons in the *GT* image. For example, it is difficult to detect a human body part with the proposed method when the pixel values corresponding to a human body part in the input image are similar to the pixel values corresponding to the background. In addition, the low-resolution thermal images used in this study have less spatial pattern information than the general visible light images, which may have contributed to the error.

Proposed IPGAN-2 predicts images on the left side not because the movement of humans is towards the left side. As shown in Figure 22a, IPGAN-2 predicts the left side of the current image at t-0 when the movement is towards the left side. In addition, as shown in Figure 22b, IPGAN-2 predicts the left side of the current image at t-0 after flipping images when the movement is towards the right side. Finally, we combine the two predicted regions with the current image at t-0. Thus, the prediction does not rely on the movement direction. By predicting both left and right sides of current image, we can increase the FOV of current image. In addition, the reason why we do not predict the left and right-side regions outside FOV at the same time using a single IPGAN is owing to experimental results as shown in Figure 7 in Section 4.2 (Ablation study). As shown in Figure 7, the performance of predicting the left and right sides of the current image at the same time (Figure 7a–d) is lower than predicting only one side of the image (Figure 7g).

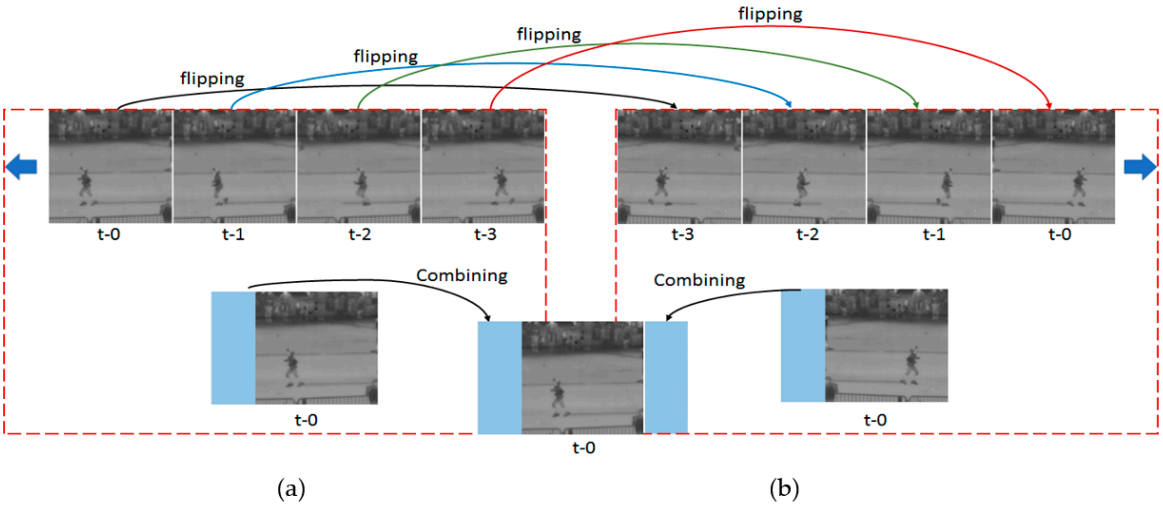


Figure 22. Example of our image prediction method.

Figure 23 shows examples of gradient-weighted class activation mapping (Grad-CAM) [44] images extracted from Conv2, Conv3, and Conv8, which are the layers in Mask R-CNN, which uses the images generated by IPGAN-2 as input.

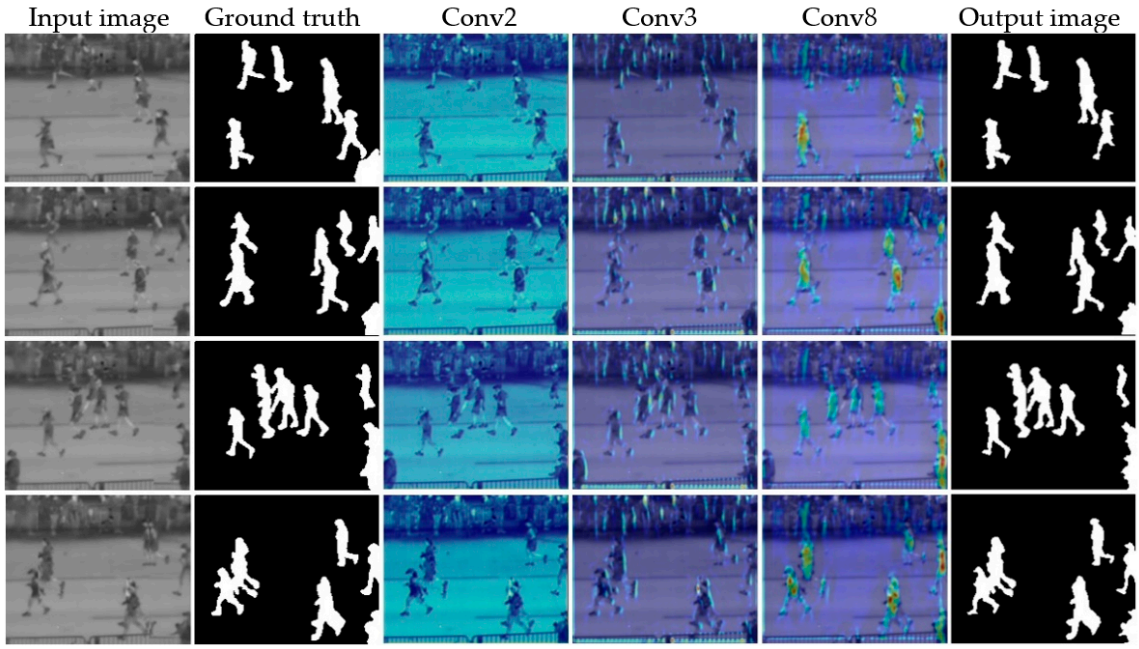


Figure 23. Example images extracted from Mask R-CNN layers by using Grad-CAM.

As shown in Figure 23, the Grad-CAM images extracted from the convolution layer (Conv2) of Mask R-CNN had almost no high activation regions. As convolution proceeded, activation regions were observed in the legs and head of a person and in the edge region of the front and back torso in the Grad-CAM images extracted from the convolution layer (Conv3). Figure 23 confirms that activation regions were observed in a more global region, including the torso, starting from the Conv8 layer, which signifies that more-accurate human segmentation is possible in the output image.

In this study, we proposed a method to predict image region outside FOV to restore a part of human body which has disappeared when a pedestrian leaves the camera FOV. There are several reasons that we started this study. For example, in case of tracking a

suspect in the CCTV camera system, our method helps to generate a body of the suspect after he or she has left the FOV of camera. In addition, the proposed method helps to track suspects continuously without losing them when a camera changes the view direction to a suspect who left the FOV.

Moreover, we conducted various experiments (Table 12 and Figure 7) in our ablation study to achieve good results. To validate the predicted images by the proposed method, we conducted human segmentation using the predicted images (Tables 13 and 14, and Figures 10 and 14–16). We measured the predicted images using SSIM and PSNR, and measured segmentation results using TPR, PPV, ACC, F1 score, and IoU, which confirms that the performance of our method is better than those of the state-of-the-art methods.

6. Conclusions

The IPGAN-2 method was proposed for image prediction for thermal images where the occurrence of noise is minimized while the wide regions to the left and right sides of the FOV in the current image are accurately generated. For improving the accuracy of image prediction, binary images corresponding to sequential input thermal images were used as input for IPGAN-2. For evaluating the performance of the proposed method, various ablation studies using original one-channel thermal images and comparative experiments using state-of-the-art methods were performed. The experimental results using an open database showed that the proposed IPGAN-2-based method had higher image prediction accuracy than other methods, including the state-of-the-art methods. The TPR, PPV, F1, ACC, and IoU of human segmentation using the proposed method were 0.887, 0.847, 0.866, 0.914, and 0.730, respectively, which are better results than those of the state-of-the-art methods. In the experimental results, the persons in the predicted region of an image may be more poorly segmented than the persons in the *GT* image. This could be because the pixel values corresponding to the human are similar to the pixel values corresponding to the background, which hindered the distinction between the human body part and background. Moreover, thermal images with less spatial pattern information and the factors in low-resolution images obtained from long distance may be affected the error.

In future work, image prediction using thermal and visible light images combined is planned to resolve such issues. For the proposed method, experiments were only conducted in a fixed-camera setting and not in a moving-camera setting; therefore, further experiments should be conducted to determine whether the proposed method is applicable in a moving-camera setting. In addition, further research is planned on image prediction in which the FOV of a visible light camera in a vehicle is expanded in four directions.

Author Contributions: Methodology, G.B.; validation, N.R.B.; 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 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-2021R1F1A1045587), 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).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Gong, J.; Zhao, J.; Li, F.; Zhang, H. Vehicle detection in thermal images with an improved yolov3-tiny. In Proceedings of the IEEE International Conference on Power, Intelligent Computing and Systems, Shenyang, China, 28–30 July 2020.
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. Zhang, X.; Chen, R.; Liu, G.; Li, X.; Luo, S.; Fan, X. Thermal infrared tracking using multi-stages deep features fusion. In Proceedings of the Chinese Control and Decision Conference, Hefei, China, 22–24 August 2020.
6. Svanström, F.; Englund, C.; Alonso-Fernandez, F. Real-time drone detection and tracking with visible, thermal and acoustic sensors. In Proceedings of the 25th International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021.
7. Liu, Q.; Li, X.; He, Z.; Fan, N.; Yuan, D.; Wang, H. Learning deep multi-level similarity for thermal infrared object tracking. *IEEE Trans. Multimed.* **2021**, *23*, 2114–2126. [CrossRef]
8. Liu, Q.; He, Z.; Li, X.; Zheng, Y. PTB-TIR: A thermal infrared pedestrian tracking benchmark. *IEEE Trans. Multimed.* **2020**, *22*, 666–675. [CrossRef]
9. Kang, B.; Liang, D.; Ding, W.; Zhou, H.; Zhu, W.-P. Grayscale-thermal tracking via inverse sparse representation-based collaborative encoding. *IEEE Trans. Image Process.* **2020**, *29*, 3401–3415. [CrossRef] [PubMed]
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] [PubMed]
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. 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]
14. 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. [CrossRef]
15. 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.
16. Image Prediction Generative Adversarial Network v2 (IPGAN-2). Available online: http://dm.dgu.edu/link.html (accessed on 25 March 2021).
17. 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.
18. 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.
19. 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] [PubMed]
20. 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.
21. 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.
22. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Akbari, Y. Image inpainting: A review. *Neural Process. Lett.* **2020**, *51*, 2007–2028. [CrossRef]
23. Liang, X.; Lee, L.; Dai, W.; Xing, E.P. Dual motion GAN for future-flow embedded video prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
24. Sedaghat, N.; Zolfaghari, M.; Brox, T. Hybrid learning of optical flow and next frame prediction to boost optical flow in the wild. *arXiv* **2017**, arXiv:1612.03777v2.
25. Mahjourian, R.; Wicke, M.; Angelova, A. Geometry-based next frame prediction from monocular video. *arXiv* **2017**, arXiv:1609.06377v2.
26. 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, Virtual, Waikola, HI, USA, 5–9 January 2021.
27. 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, Virtual, Seattle, WA, USA, 14–19 June 2020.
28. 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.

29. 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.
30. 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.
31. 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]
32. Rasouli, A. Deep learning for vision-based prediction: A survey. *arXiv* **2020**, arXiv:2007.00095v2.
33. Nvidia GeForce GTX TITAN X. 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. Python. Available online: https://www.python.org/download/releases/ (accessed on 25 March 2021).
36. Keras. Available online: https://keras.io/ (accessed on 25 March 2021).
37. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Peak Signal-to-Noise Ratio. Available online: https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio (accessed on 29 April 2021).
39. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
40. Powers, D.M.W. Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Mach. Learn. Technol.* **2011**, *2*, 37–63.
41. 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.
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. Tan, D.; Huang, K.; Yu, S.; Tan, T. Efficient night gait recognition based on template matching. In Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, 20–24 August 2006.
44. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *arXiv* **2016**, arXiv:1610.02391v4.