Background/Foreground Separation: Guided Attention based Adversarial Modeling (GAAM) versus Robust Subspace Learning Methods

Maryam Sultana[1], Arif Mahmood[2], Thierry Bouwmans[3], Muhammad Haris Khan[4], Soon Ki Jung[1]
[1]School of Computer Science and Engineering, Kyungpook National University, South Korea.
[2] Department of Computer Science, Information Technology University (ITU), Lahore, Pakistan.
[3] Laboratoire MIA, LaRochelle, Universite deLaRochelle, France.
[4] Mohamed Bin Zayed University of Artificial Intelligence, United Arab Emirates.
[1]{maryam,skjung}@knu.ac.kr, [2]arif.mahmood@itu.edu.pk, [3]thierry.bouwmans@univ-lr.fr
[4] muhammad.haris@mbzuai.ac.ae

Abstract

Background-Foreground separation and appearance generation is a fundamental step in many computer vision applications. Existing methods like Robust Subspace Learning (RSL) suffer performance degradation in the presence of challenges like bad weather, illumination variations, occlusion, dynamic backgrounds and intermittent object motion. In the current work we propose a more accurate deep neural network based model for background-foreground separation and complete appearance generation of the foreground objects. Our proposed model, Guided Attention based Adversarial Model (GAAM), can efficiently extract pixel-level boundaries of the foreground objects for improved appearance generation. Unlike RSL methods our model extracts the binary information of foreground objects labeled as attention map which guides our generator network to segment the foreground objects from the complex background information. Wide range of experiments performed on the benchmark CDnet2014 dataset demonstrate the excellent performance of our proposed model.

1. Introduction

Background-Foreground separation and appearance generation is a critical step in many applications of computer vision such as smart cities traffic monitoring [3], human activity analysis [22], surveillance, and security [23]. Despite wide applications, this problem is challenging in the presence of complex scenes manifesting various conditions like illumination variations, bad weather, occlusion, intermittent object motion, and dynamic background.

To address these challenges in background-foreground separation, prior works have relied on Robust Subspace Learning (RSL) and Robust Principal Components Analysis (RPCA) techniques. The underlying idea of RSL based methods is to convert the input data matrix (video frames processed in batch format) into two components, low-rank and sparse information. Where the former component represents the background while the latter represents the foreground information. The RSL based techniques have shown promising results for background-foreground separation [12, 9, 4]. However, without using spatio-temporal constraints it is extremely challenging for RPCA based methods to handle dynamic background variations because they assume that the foreground is dynamic while the background is static. However, in the real complex environments this is a rather restrictive assumption. Despite the good performance, RSL based techniques cannot faithfully generate the appearance of foreground objects in complex scenes. Therefore, exploiting deep neural networks, specifically generative adversarial networks (GANs), to address this problem is a suitable alternative. For instance, robust deep auto-encoders can not only segment background-foreground information but also these methods can generate the appearance of the foreground objects [14].

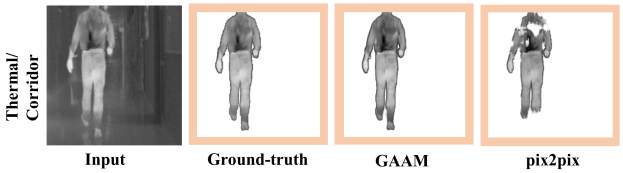

Figure 1. Foreground appearance generation by our proposed GAAM model exploiting guided attention based adversarial network and by an existing GAN based methodology pix2pix [8]. The video sequence 'Corridor' shown in this figure are from benchmark CDnet2014 dataset category 'Thermal'.

In the current work we aim to separate background-

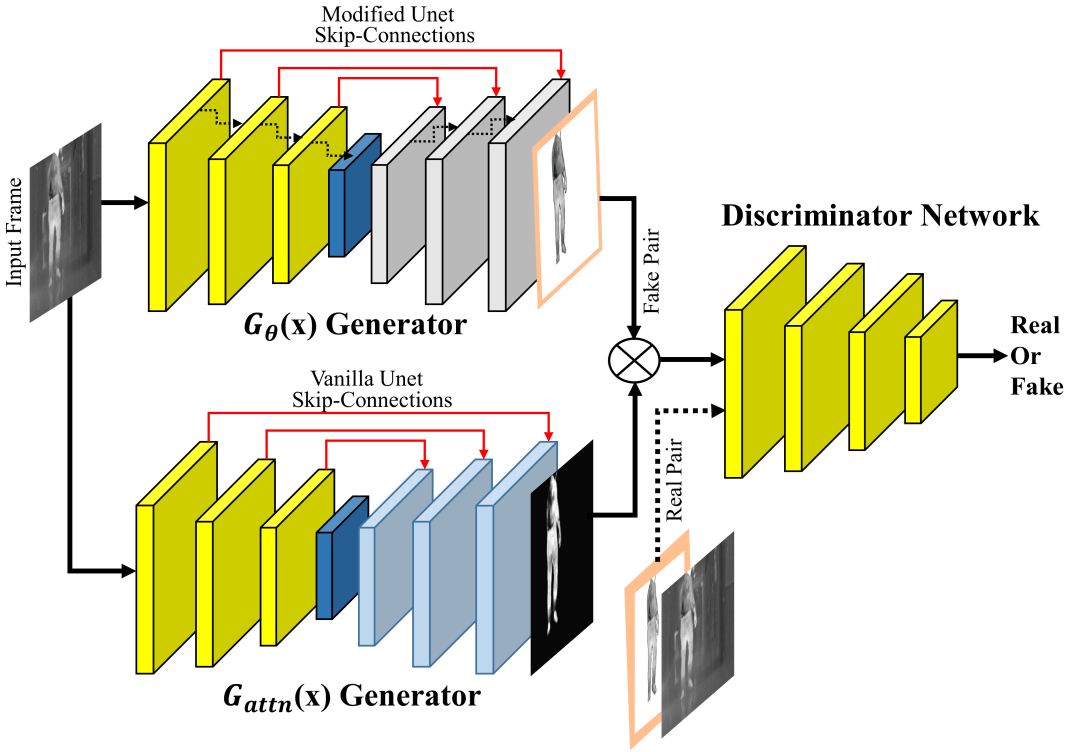

Figure 2. Overall architecture of our proposed GAAM model using guided attention based adversarial network. The input frame is given to the modified Unet based primary generator G_θ and vanilla U-net based based secondary generator G_{attn}. The primary generator outputs the foreground appearance information while the secondary generator outputs the attention. The G_{attn} probability map guides our proposed model to generate accurate foreground appearance by removing artifacts. While the discriminator role is same to perform classification between paired wise real vs fake samples. The video sequence 'Corridor' shown in this figure are from benchmark CDnet2014 dataset category 'Thermal'.

foreground information and generate more accurate appearance (see figure 1). To actualize this, we propose a Guided Attention based Adversarial Model (GAAM) for accurate background-foreground separation as well as foreground generation. Our attention module mimics the idea of separating background-foreground objects in complex scenes same as RSL based techniques. However, RSL based methods cannot generate the appearance of the required foreground objects. In contrast to that, our model has the ability to not only separate the background-foreground objects in complex scenes but it can also generate the appearance of foreground objects with high accuracy. This idea is visually appealing because our model cannot only separate the background-foreground objects in binary format but it can also provide the visual appearance of the separated foreground objects in real complex environments. Our main contributions are summarized as follows:

- We propose a novel approach that can more accurately separate background-foreground information in complex scenes and it can also generate the complete appearance of the foreground objects with high accuracy.
- We propose an attention based approach to suppress background image regions. This key aspect in our GAAM model separates background-foreground information and generates appearance in complex scenes simultaneously (see figure 2).
- We propose an efficient formulation of various loss functions to claim high accuracy in this task.

The rest of the paper is organized as follows: Section 2 explains related work in the domain of background-foreground separation in complex scenes. Our proposed GAAM method, including the details of model architecture, is explained in Section 3. The implementation details and the experimental results are briefly discussed in Sections 4 and 5, respectively. Finally, the conclusion of this study is presented in Section 6.

2. Related Work

Background-Foreground separation in complex scenes has remained an important research topic over the past two decades and so many studies have been conducted to address this challenging problem [5]. Whereas foreground

appearance generation is a relatively new domain in this research problem and is attracting attention [17]. The classical algorithms for background-foreground separation are based on robust subspace learning [15, 16, 10]. Javed *et al.* [11] proposed a RPCA based efficient technique to handle spatio-temporal information in complex scenes for background-foreground separation. The method performed well in many complex scenes however, it is a hybrid RPCA based methodology that addresses the offline data processing limitation. Wipf *et al.* [7] proposed a study based on the fusion of classical robust subspace learning approach with deep learning methodology. The technique presents the idea that Variational autoencoders can be considered as the natural evolution of RPCA with the potential to learn the nonlinear manifolds of unknown dimension cloak by entire data corruptions. Despite the excellent performance of RSL methods the limitations of high computational cost and offline data processing makes them unsuitable for real-time applications. Moreover, RSL based approaches can only segment foreground objects in binary format and lacks the ability to generate foreground appearance information.

Recently, robust deep learning based methods including generative adversarial networks have shown significantly high-performance in background-foreground separation and appearance generation [8]. For instance, Isola *et al.*[8] proposed a technique using conditional generative adversarial network that has the ability to generate appearance of the foreground objects. The training is done using cross-entropy loss term and ℓ_1 distance between generated output and the original image sample ground-truth. Sultana *et al.* [19] proposed a method called 'CcGAN' using conditional generative adversarial network as well. The model has the ability to not only segment the occluded foreground objects but also it can generate the missing information caused by the occlusion. Sultana *et al.*[18] also proposed a technique 'BslsGAN' that works using conditional least squares generative adversarial networks. The model has the potential to efficiently segment the foreground objects from complex background information. In comparison to above mentioned existing works our proposed method has an attention module with efficient loss terms and an effective generator network that has the potential to segment as well as generate the appearance information simultaneously.

3. Proposed Guided Attention Adversarial Model (GAAM)

Our proposed GAAM as shown in Figure 2 is a supervised learning technique for background-foreground separation and appearance generation. In order to generate the full appearance our goal is to train a generator network $G : X \rightarrow Y$. We are given paired input samples $x \in X$ and $y \in Y$, and the supervised setting assumes that x and y are drawn from distributions $P_{x \backsim X}(x)$ and $P_{y \backsim Y}(y)$. The paired image samples are the given input video frames with its corresponding foreground objects. The model learns a transformation of input data to foreground appearance generation. Therefore, based on the requirements described above, we propose to learn θ by minimizing the following objective functions described in the following section.

3.1. Loss Functions

For the generator G combined loss function is given below:

$$\begin{aligned}\min_{G}\ (&\mathcal{L}_{adv}(x,G(x),D)+\alpha_1\mathcal{L}_{hinge}(x,G(x))\\&+\alpha_2\mathcal{L}_{app}(x,G(x))+\alpha_3\mathcal{L}_{style}(x,G(x))),\end{aligned} \tag{1}$$

where $G(x) = G_{attn}(x) \otimes G_\theta(x)$, and G_θ is the modified Unet based generator and G_{attn} is the attention based secondary generator (architectural details are provided in section 3.2). The output of G_θ is a transformed image in terms of foreground appearance generation. While G_{attn} predicts a probability map that guides our proposed model to generate accurate foreground appearance by removing artifacts. In the final training objective function mentioned in Eq. (1) the loss term $\mathcal{L}_{adv}$ is the least squares adversarial loss given by Eq. (2), $\mathcal{L}_{hinge}$ is the hinge loss given by Eq. (5), $\mathcal{L}_{app}$ given by Eq. (6) is the appearance loss and $\mathcal{L}_{style}$ given by Eq. (7) is the feature matching neural style transfer loss in the image domain to make sure that $G(x)$ learns to not only separate the background-foreground objects but also generates its appearance with better accuracy. In the proposed system, all the three networks are trained jointly in end-to-end manner. The details of the adversarial loss are as follows:

$$\mathcal{L}_{adv}(x,G(x),D) = \frac{1}{2}\mathbb{E}_x[(D(x,G(x))-1)^2]. \tag{2}$$

The corresponding adversarial loss term for the discriminator is given by:

$$\begin{aligned}\min_{D}\ \mathcal{L}_{adv}(x,y,G(x),D) = &\frac{1}{2}\mathbb{E}_{x,y}[(D(x,y)-1)^2]+\\&\frac{1}{2}\mathbb{E}_x[(D(x,G(x)))^2].\end{aligned} \tag{3}$$

The over all discriminator loss consist of two terms including adversarial loss and hinge loss:

$$\min_{D}\ (\mathcal{L}_{adv}(x,y,G(x),D)+\alpha_1\mathcal{L}_{hinge}(x,y,G(x),D)) \tag{4}$$

The hinge loss in our proposed model is formulated as follows [13]:

$$\begin{aligned}\mathcal{L}_{hinge}(x,y,G,D) = &-\mathbb{E}_{(x,y)\sim p_{data}}[min(0,-1+D(x,y))]-\\&\mathbb{E}_{(x)\sim p_{data}}[min(0,-1-D(x,G))],\\\mathcal{L}_{hinge}(x,G,D) = &-\mathbb{E}_{(x)\sim p_{data}}[(D(x,G)-1)].\end{aligned} \tag{5}$$

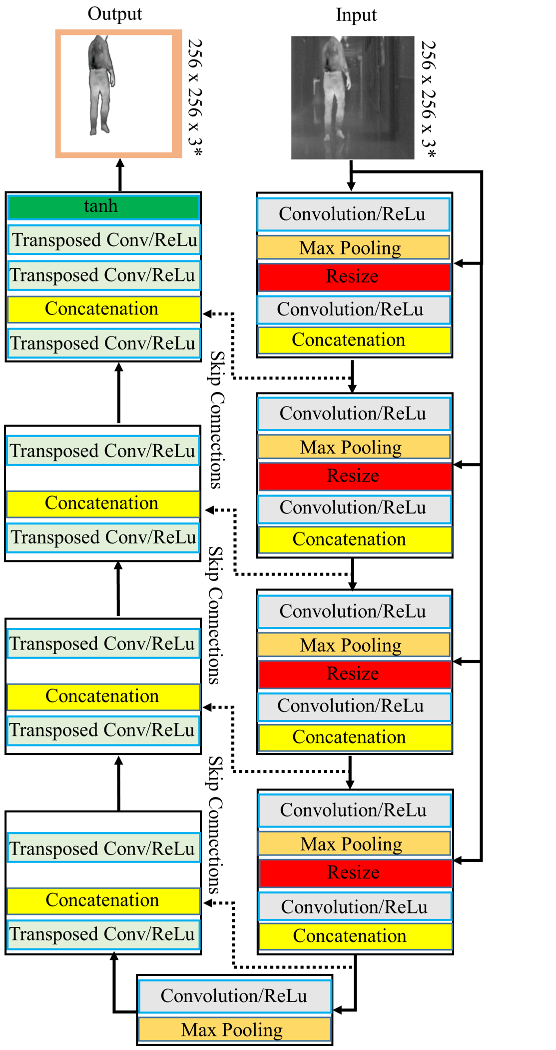

Figure 3. Generator network architecture in our proposed model is a Modified Unet. The video sequence ‘Corridor’ shown in this figure are from benchmark CDnet2014 dataset category ‘Thermal’.

The appearance loss is visual difference between the ground-truth information and output of the model:

$$\mathcal{L}_{app}(G) = \mathbb{E}_{x,y}[||G(x) - y||_1]. \quad (6)$$

The style loss is calculated between all convolutional layers of the discriminator network fed with generated output and the ground-truth:

$$\mathcal{L}_{style}(G) = \mathbb{E}_{x,y} \sum_{i=0}^{n_d} [||\Psi_i(G(x)) - \Psi_i(y)||_1], \quad (7)$$

where n_d are the discriminator layers. All loss terms constrain the generator model to learn foreground details and ignore the background clutter while maintaining the realistic touch of the foreground objects.

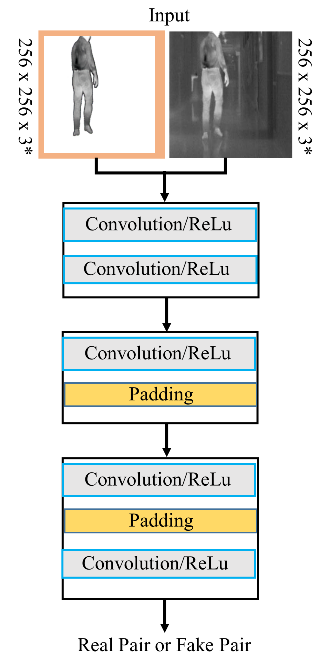

Figure 4. Discriminator network architecture in our proposed model is a PatchGAN. The video sequence ‘Corridor’ shown in this figure are from benchmark CDnet2014 dataset category ‘Thermal’.

3.2. Model Architecture

Our proposed GAAM model consists of a primary and a secondary generator and a single discriminator. The primary generator is a modified version of Unet model (G_θ) as shown in the Figure 3, while a vanilla Unet generator is exploited to generate attention (G_{attn}). The aim of the G_θ is to generate the appearance of the foreground objects whereas G_{attn} predicts the probability map that is considered as the attention mask. The output resolution of G_{attn} is the same as that of the input x where each pixel has a probability value between $0.00 - 1.00$, where 1.00 means foreground and 0.00 means background. The architecture of the G_θ is similar to Unet formulation but with added max-pooling layers in the down-sampling path and within block short-circuit connections. More specifically, in the encoding part, convolution with stride 1 is performed on original image samples as well as its resized version. The feature maps from both convolutional layers are concatenated within each block and input to the next block till the bottle-neck layer. Afterwards, in the decoder part we perform transposed convolutions with stride 1 and 2 to match the exact dimensions of feature maps for creating skip-connections between encoder-decoder as shown in the Figure 3. This modified version of Unet model helps in the blending of local as well as global features necessary for accurate foreground appearance generation. All the weights are randomly initialized and LeakyReLU activation is used. The discriminator net-

Algorithm	Algorithm Type	Challenge Categories							
		Baseline	Dynamic Bg	Thermal	Bad Weather	IOM*	Camera Jitter	Shadows	Average
MSCL [11]	RSL	0.87	0.85	0.82	0.83	0.80	0.83	0.82	0.83
DECOLOR [24]	RSL	0.92	0.70	0.70	0.76	0.59	0.77	0.83	0.75
TVRPCA [6]	RSL	0.84	0.55	0.69	0.78	0.57	0.63	0.71	0.68
SPRCA [10]	RSL	0.82	0.84	0.79	0.75	0.80	0.78	0.77	0.79
DeepBS [2]	DL	0.95	0.87	0.75	0.83	0.63	0.89	0.93	0.83
BSUV-net [20]	DL	0.96	0.79	0.85	0.87	0.74	0.77	0.92	0.84
DeepDC [1]	DL	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.91
GAAM	DL	0.96	0.90	0.93	0.95	0.82	0.91	0.94	0.91

Table 1. Background-Foreground separation: Quantitative performance comparison of our proposed GAAM model with RSL as well as Deep learning (DL) based existing methods on benchmark CDnet2014 dataset using F measure. The highest and the second highest results are shown in red and blue colors, where IOM* is category 'Intermittent Object Motion'.

work, as shown in Figure 4, in our proposed GAAM system works on the formulation of PatchGAN [8]. The goal of this formulation is to classify real vs fake over-lapping image patches. PatchGAN is a Fully Convolutional Neural Network (FCN) that has fewer parameters as compared to full image discriminator and it has the ability to process arbitrarily sized input images.

4. Implementation

In this study, we implement our proposed GAAM system using Tensorflow and execute the system on a single TitanX GPU. Different blocks in our proposed networks maintain the arrangement of convolution-BatchNorm-LReLu modules [8]. The training and testing samples are fixed to the resolution $256 \times 256 \times 3$. We use the Adam optimizer with a learning rate of 0.0002 and $\beta = 0.5$. We also exploit data augmentation in the training of our proposed model by random flipping of image samples. During testing the model is given input video frames from complex scenes to generate foreground appearance with high accuracy.

5. Experiments

Experiments are performed on 'Change Detection 2014 (CDnet2014)' [21] benchmark dataset. Seven challenging categories are selected including 'Baseline', 'Bad Weather', 'Camera Jitter', 'Dynamic Background', 'Intermittent Object Motion', 'Shadows' and 'Thermal'. For the training of our GAAM model we exploited 70% of video frames from each category while remaining 30% are used for the testing of scene-specific models. The qualitative and quantitative comparisons of our proposed model are done with various existing RSL based techniques such as MSCL [11], DECOLOR [24], TVRPCA [6], SRPCA [10], and deep learning based techniques including DeepBS [2], BSUV-net [20], DeepDC [1], pix2pix [8], CcGAN [19], BsLsGAN [18]. In our experiments we did two kinds of evaluation one for the binary background-foreground separation (attention maps) and the other is the foreground appearance generation. For the former we exploited F measure calculated as:

$$R = \frac{T_P}{T_P + F_N}, \qquad P = \frac{T_P}{T_P + F_P}, \tag{8}$$

$$F = \frac{2(P \cdot R)}{P + R}. \tag{9}$$

In the above equations F_N is False Negatives, F_P is False Positives, T_P is True Positives, P is precision and R is Recall. Therefore, to achieve high accuracy the model should have high value of F measure and low value of L_1 distance as shown in tables 1 and 2. For the appearance generation we used L_1 distance for comparison with existing state-of-the-art methods. More details about the results are discussed in the following sections.

5.1. Background-Foreground Separation Evaluation

We converted attention maps into binary format for the ease of quantitative comparison of our proposed GAAM model with RSL as well as deep learning based techniques. It can be seen in Table 1 that our proposed model has achieved best results in all challenging categories of CDnet2014 dataset. The quantitative results also show that RSL based methods are affected by the challenging conditions in the complex scenes. For instance, in the conditions like camera jitter, the background is dynamic and it leads RSL based techniques towards performance degradation. However, our proposed model is not affected by static or dynamic background scenes, hence achieved good performance. The visual results presented in Figure 5 shows that our GAAM model has excellent performance as compared to RSL and deep learning based methods.

5.2. Foreground Appearance Generation Evaluation

For the evaluation of foreground appearance generation, we exploited L_1 distance to calculate error between the generated output and ground-truth information. It can be seen in Table 2 that our proposed GAAM model has achieved best results in all challenge categories of CDnet2014 dataset

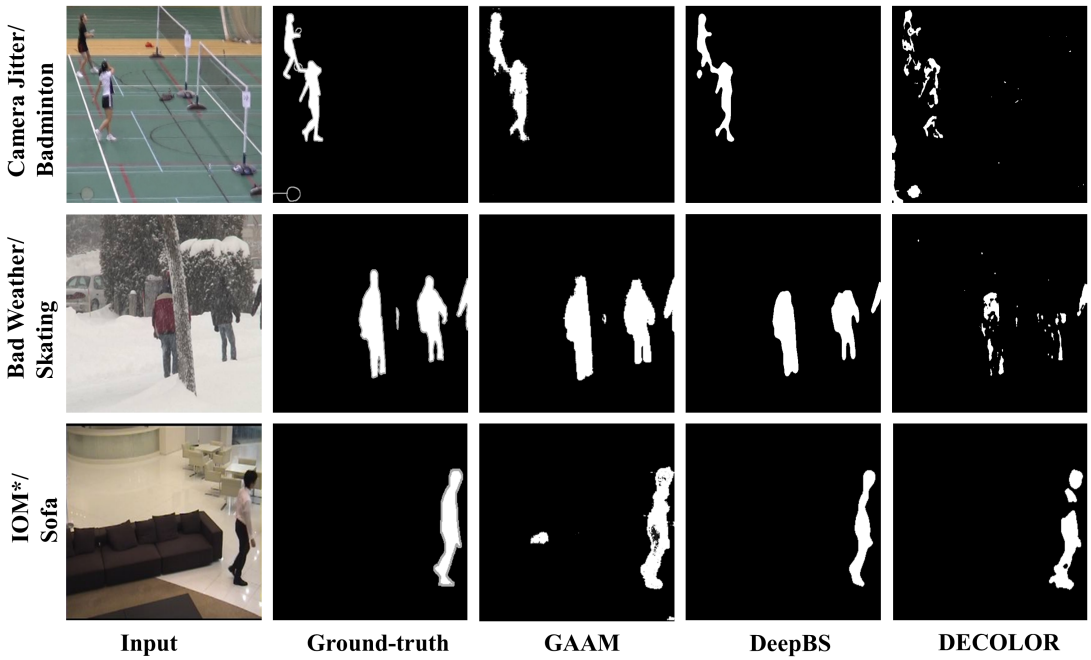

Figure 5. Background-Foreground separation performance visual comparison with existing methods on benchmark CDnet2014 dataset. The qualitative results show the performance degradation of RSL based technique DECOLOR [24] while deep learning based methods including our proposed GAAM has better results.

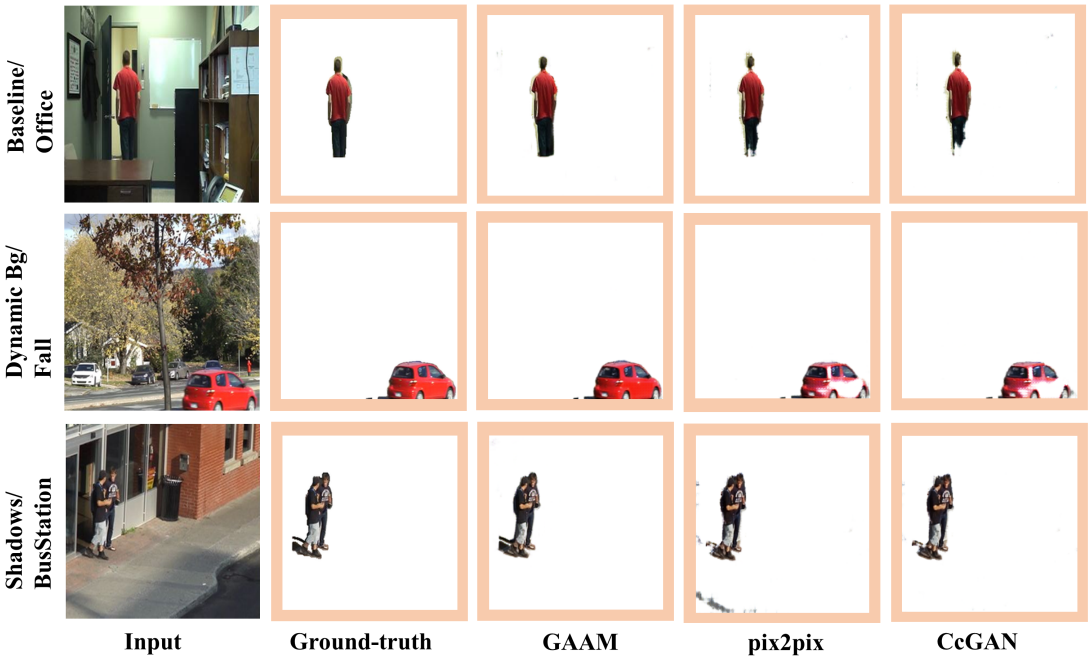

Figure 6. Visual performance comparison of foreground generation on benchmark CDnet2014 dataset. It can be seen in the figure that both pix2pix [8] and CcGAN [19] have shown performance degradation due to the fact that they lack attention module to estimate better foreground appearance. In contrast to that, GAAM has the potential to generate good quality foreground appearance information.

with lowest values of L_1 distance. The visual results presented in Figure 6 show that our proposed model is efficient in terms of foreground appearance generation. For instance, in the category 'Baseline' the video sequence 'Office' (first row Figure 5) the existing method like pix2pix [8] has artifacts in its generated foreground appearance. The reason

Challenge Categories	Algorithm			
	pix2pix [8]	CcGAN [19]	BslsGAN [18]	GAAM
Baseline	0.0167	0.1114	0.0890	0.0089
Dy Bg*	0.2987	0.1753	0.0765	0.0035
Thermal	0.1334	0.2000	0.0999	0.0125
Bad Weather	0.1770	0.1041	0.1480	0.0097
IOM**	0.2901	0.2228	0.1300	0.0131
Camera Jitter	0.1040	0.2165	0.1689	0.0052
Shadows	0.0212	0.1040	0.1190	0.0043

Table 2. Foreground Appearance Generation: Quantitative performance comparison of our proposed model with Deep learning (DL) based existing methods on benchmark CDnet2014 dataset using L_1 distance. The highest and the second highest results are shown in red and blue colors, where Dy Bg* is category 'Dynamic Background' and IOM** is category 'Intermittent Object Motion'.

Training loss Terms	F measure
$\mathcal{L}_{adv}$	0.83
$\mathcal{L}_{adv} + \alpha_1\ \mathcal{L}_{hinge}$	0.84
$\mathcal{L}_{adv} + \alpha_1\ \mathcal{L}_{hinge} + \alpha_2\ \mathcal{L}_{app}$	0.84
$\mathcal{L}_{adv} + \alpha_1\ \mathcal{L}_{hinge} + \alpha_2\ \mathcal{L}_{app} + \alpha_3\ \mathcal{L}_{style}$	0.91
Generator Network Formulation	L_1 Distance
Without Attention	0.2943
With Attention	0.0052

Table 3. Ablation study: performance comparison of different loss terms in our proposed objective function on CDnet2014 dataset, category 'Camera Jitter'.

behind this fact is, although pix2pix is an efficient adversarial learning based technique, however, only cross-entropy loss term and vanilla Unet lacks the ability to generate the accurate pixel-level appearance information of foreground objects. On the other hand, our proposed model, working with guided attention based modified Unet formulation and efficient loss terms enhances the quality of the generated image.

5.3. Ablation Study

We performed two kinds of ablation studies to evaluate the significance of different components in the proposed objective function. First part of the ablation study highlights the importance of various loss terms, while the second term is about the effect of attention module in our proposed GAAM model. The analysis is performed as follows:

- It can be seen in Table 3 that adding hinge loss in the least square adversarial objective function improves the results. Moreover, the additional regularization in terms of image and feature domain that are actually appearance and style losses further improve the quality of the generated image. Note that in this category of ablation study, the network architecture is kept fixed.

- The second part of the ablation experiments are to highlight the significance of the attention module. It can be seen in Table 3 that adding the attention module improves the results of foreground appearance generation. In practice without attention module the foreground appearance generation suffer from perturbations of background texture/color and artifacts around the region of interests. While with the predicted attention maps our GAAM model learns to segment the foreground objects accurately, thus maintaining the high visual quality of the results as shown in Figure 7. Note that in this category of ablation study, the final objective function is kept fixed as Eq. (1).

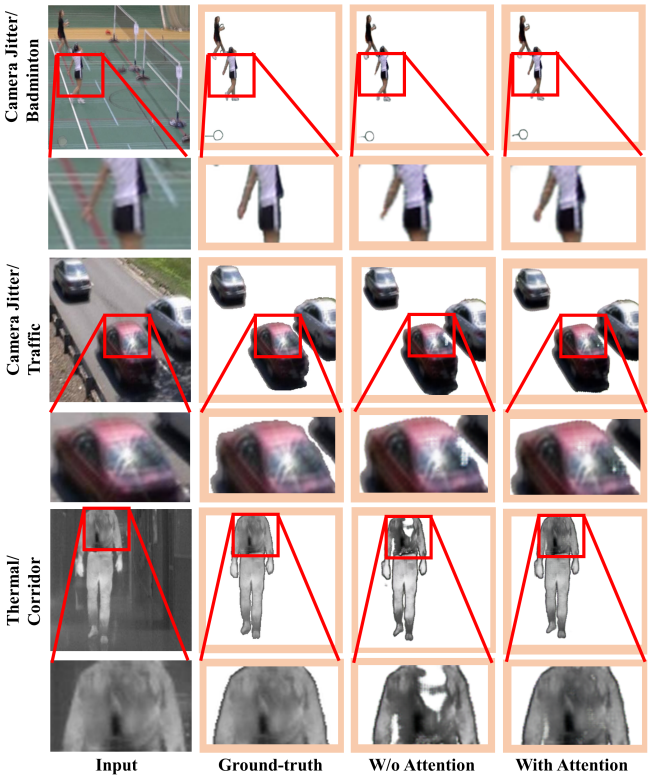

Figure 7. Effects of attention module in our proposed GAAM method output for foreground appearance generation on benchmark CDnet2014 dataset. The visual results show that without the attention module there could be missing information in the appearance generation of the foreground objects. Nonetheless, the attention guided module generates better quality foreground information.

6. Conclusion

In this work a deep learning based algorithm is proposed for background-foreground separation as well as foreground generation. The proposed model works with guided attention based adversarial module that has the efficiency to extract pixel level boundaries of the background-foreground region separation. Unlike RSL methods our model extracts binary information of foreground objects which are labeled as attention maps. The attention maps guide the generator network to segment the foreground objects from the complex background scenes. Experiments

performed on benchmark CDnet2014 dataset demonstrated excellent performance of the proposed model compared with various existing state-of-the-art RSL methods as well as deep learning based techniques.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2019R1A2C1010786).

References

[1] Sirine Ammar, Thierry Bouwmans, Nizar Zaghden, and Mahmoud Neji. Deep detector classifier (deepdc) for moving objects segmentation and classification in video surveillance. *IET Image Processing*, 2020.

[2] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76:635–649, 2018.

[3] Chris Baber, Natan Sorin Morar, and Faye McCabe. Ecological interface design, the proximity compatibility principle, and automation reliability in road traffic management. *IEEE Transactions on Human-Machine Systems*, 2019.

[4] Thierry Bouwmans, Sajid Javed, Maryam Sultana, and Soon Ki Jung. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Networks*, 2019.

[5] Thierry Bouwmans and El Hadi Zahzah. Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014.

[6] Xiaochun Cao, Liang Yang, and Xiaojie Guo. Total variation regularized rpca for irregularly moving object detection under dynamic background. *IEEE transactions on cybernetics*, 46(4):1014–1027, 2016.

[7] Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE CVPR*, pages 1125–1134, 2017.

[9] Sajid Javed, Arif Mahmood, Somaya Al-Maadeed, Thierry Bouwmans, and Soon Ki Jung. Moving object detection in complex scene using spatiotemporal structured-sparse rpca. *IEEE Transactions on Image Processing*, 28(2):1007–1022, 2018.

[10] Sajid Javed, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Spatiotemporal low-rank modeling for complex scene background initialization. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1315–1329, 2016.

[11] Sajid Javed, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Background–foreground modeling based on spatiotemporal sparse subspace clustering. *IEEE Transactions on Image Processing*, 26(12):5840–5854, 2017.

[12] Sajid Javed, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Spatiotemporal low-rank modeling for complex scene background initialization. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1315–1329, 2018.

[13] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

[14] Jiancheng Ni, Susu Zhang, Zili Zhou, Lijun Hou, Jie Hou, and Feng Gao. Background and foreground disentangled generative adversarial network for scene image synthesis. *Computers & Graphics*, 97:54–66, 2021.

[15] Peng Pan, Yongli Wang, Mingyuan Zhou, Zhipeng Sun, and Guoping He. Background recovery via motion-based robust principal component analysis with matrix factorization. *Journal of Electronic Imaging*, 27(2):023034, 2018.

[16] Behnaz Rezaei and Sarah Ostadabbas. Moving object detection through robust matrix completion augmented with objectness. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1313–1323, 2018.

[17] Sijie Song, Wei Zhang, Jiaying Liu, Zongming Guo, and Tao Mei. Unpaired person image generation with semantic parsing transformation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[18] Maryam Sultana, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Dynamic background subtraction using least square adversarial learning. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3204–3208. IEEE, 2020.

[19] Maryam Sultana, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Complete moving object detection in the context of robust subspace learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[20] Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. Bsuv-net: a fully-convolutional neural network for background subtraction of unseen videos. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2774–2783, 2020.

[21] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 393–400. IEEE, 2014.

[22] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018.

[23] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 2019.

[24] Xiaowei Zhou, Can Yang, and Weichuan Yu. Moving object detection by detecting contiguous outliers in the low-rank representation. *IEEE T-PAMI*, 35(3):597–610, 2013.