



Background-Foreground Segmentation Using Multi-Scale Attention Net (MA-Net): A Deep Learning Approach

Vishruth Boraiah Gowda^{1,2*}, Gopalakrishna Madigondanahalli Thimmaiah^{2,3}, Megha Jaishankar^{2,4}, Chaitra Yuvaraj Lakkondra^{2,5}

¹ Department of Information Science and Engineering, SJB. Institute of Technology, Bengaluru 560060, India

² Affiliated to Visvesvaraya Technological University, Belgavi 590018, Karnataka, India

³ Department of Artificial Intelligence and Machine Learning, SJB. Institute of Technology, Bengaluru 560060, India

⁴ Department of Artificial Intelligence and Machine Learning, Ramaiah. Institute of Technology, Bengaluru 560060, India

⁵ Department of Computer Science and Engineering, SJB. Institute of Technology, Bengaluru 560060, India

Corresponding Author Email: vishruth1711@gmail.com

<https://doi.org/10.18280/ria.370304>

ABSTRACT

Received: 21 February 2023

Accepted: 18 April 2023

Keywords:

background subtraction, self-attention mechanism, multi-scale feature fusion, deep neural network architecture, background-foreground segmentation

Background subtraction serves as a critical foundation for numerous computer vision tasks, and a variety of traditional techniques have been proposed. In recent years, deep neural network architectures have emerged as a promising approach, with the UNET architecture being a notable example. However, UNET is considered an older architecture with limitations. To address these limitations, a novel neural network technique called Multi-scale Attention Net (MA-Net) is proposed, which incorporates a self-attention mechanism for adaptively integrating local features with their global dependencies. The attention mechanism within MA-Net enables the capture of complex contextual dependencies. Two distinct blocks are developed for the MA-Net: The Position-wise Attention Block (PAB) and the Multi-scale Fusion Attention Block (MFAB). While PAB models the interdependencies between features from spatial dimensions, representing pixel dependencies in a global view, MFAB capitalizes on fused multi-scale semantic feature fusion to capture channel dependencies between feature maps, effectively segmenting the foreground from the background. The proposed method is evaluated using the CDNET 2014 dataset and demonstrates improved performance under Shadow, dynamic background, and illumination challenges. This study highlights the potential of the MA-Net for advancing the field of background-foreground segmentation in computer vision tasks.

1. INTRODUCTION

Video surveillance has gained prominence in recent times, with a wide range of applications including intelligent traffic monitoring, optical motion capture, and human-computer interfaces. To achieve effective foreground detection in video surveillance, it is crucial to have a robust background model. Our work aims to design an efficient background model for this purpose. This process of designing a background model for identifying the foreground pixels is known as background subtraction, as illustrated in the study [1].

Background subtraction forms the basis of computer vision, and it is used as a preprocessing step in many applications namely object tracking, surveillance, and human-computer interaction. Traditional techniques for background subtraction have been proposed, including GMM (Gaussian mixture models) and KDE (kernel density estimation) [2, 3]. However, these methods are sensitive to variations in lighting and camera motion, and they have difficulty dealing with complex backgrounds and dynamic scenes.

Off late, techniques based on deep learning techniques are applied for background subtraction, and they have shown promising results [4-6].

Feature extraction is an important step in comparing the

background image and the video frames. The right features must be chosen to reveal the necessary information. Most algorithms use grayscale or the intensities of the three colors RGB as features. In a few cases, the intensity of pixels is combined with other features that are manually designed, such as in the study [7, 8]. Additionally, it is pertinent to select the appropriate feature region. Features could be extracted from pixels, blocks, or patterns. However, features from pixel-wise categories often produce noisy segmentation because it does not take into account local correlation, whereas block-wise or pattern-wise features may be insensitive to small changes.

Segmentation is a key step in processing each video frame by using a background model. It is achieved by taking features either from pixels or regions of pixels that are the same in both frames and by adopting a distance measure, such as the Euclidean distance, for determining how similar the pixels are. Each pixel is then assigned a background or foreground label based on its similarity to a similarity threshold. The entire background subtraction system is composed of these building blocks, and its reliability is dependent on and limited by the performance of each individual block. This means that if one module does not perform well, the overall system will not function optimally. Background subtraction has been extensively studied, resulting in a wide range of algorithms

that can be used for this purpose. Many of the best methods currently in use are based on algorithms developed long ago, some of which are described earlier.

There are various segmentation models in deep learning architecture, such as Unet, Unet++, PAN, PSP Net, and Manet. However, only the Unet architecture has been used for designing a background subtraction model. With this in mind, we decided to adopt the Ma-net architecture because of its recent success in medical image segmentation [9].

Specifically, our proposed MA-Net is composed of two main blocks: PAB (Position-wise Attention Block) and MFAB (Multi-scale Fusion Attention Block). PAB models interdependencies between features from spatial dimensions, which represent dependencies of pixels spatially in a global view. MFAB relies on fused multi-scale semantic features are for capturing channel dependencies between feature maps for effectively segmenting foreground from background. Our experiments on the CDNET 2014 dataset show that MA-Net outperforms the state-of-the-art methods in challenging scenarios, particularly under the shadow, illumination changes and dynamic background. In summary, our contribution in this paper is a deep neural network architecture, MA-Net (Multi-scale Attention Net) that combines the self-attention mechanism with the multi-scale feature fusion to adaptively integrate local features with their global dependencies for background subtraction. We illustrate our method’s effectiveness on the CDNET 2014 dataset, and our results demonstrate that it performs better than a few of the state-of-the-art methods under challenging scenarios.

Further discussed, related works in section 2. Then, in section 3, we discussed the proposed work and architecture. Afterward, in section 4, we discussed the results and a comparative study. Finally, in section 5, we summarized the entire discussion with a conclusion and an overview of future possibilities.

2. RELATED WORK

Traditional background subtraction methods include GMM (Gaussian mixture models) and KDE (kernel density estimation) [2, 3]. These methods are sensitive to variations in lighting and camera motion, and they have difficulty dealing with complex backgrounds and dynamic scenes. Wang et al. [6] propose a survey on recent advances in achieved in background subtraction, including deep learning-based methods. They reviewed various deep learning architectures such as DCNN, RNN (Recurrent Neural Network), and GAN (Generative Adversarial Network) for background subtraction

Piccardi et al. [4] proposed a review of background

subtraction techniques, including both traditional and deep learning-based methods. He has discussed the pros and cons of each method, and he highlighted the potential of deep learning-based techniques in background subtraction.

Other related works include the use of UNet architecture for background subtraction [10]. UNet is a popular architecture for image segmentation tasks, but it is considered an older architecture. Our proposed MA-Net is based on UNet architecture, but it also incorporates a self-attention mechanism to adaptively integrates local features with their global dependencies, which is not present in the UNet architecture.

Another related work is the use of attention mechanisms in image segmentation tasks [11, 12]. These methods have shown that attention mechanisms can improve the performance of image segmentation tasks by selectively focusing on important regions of the input. In our proposed MA-Net, we also use an attention mechanism for background subtraction but in a different way than the previous works as we have also incorporated a fusion of multi-scale features for capturing channel dependencies between feature maps for effectively segmenting foreground from background.

In this paper, we propose a deep neural network architecture, MA-Net (Multi-scale Attention Net) that combines the self-attention mechanism with the multi-scale feature fusion which adaptively integrates local features with their global dependencies for background subtraction. Our method is different from the previous works as it is based on a self-attention mechanism that captures complex contextual dependencies and multi-scale feature fusion which captures channel dependencies between feature maps. Our experimental results on the CDNET 2014 dataset demonstrate the effectiveness of our method in challenging scenarios.

3. PROPOSED APPROACH

Proposed approach is a deep neural network architecture called MA-Net (Multi-scale Attention Net) for background subtraction. The basis for our proposed method is UNet architecture, but it also incorporates a self-attention mechanism for adaptively integrating local features with their global dependencies. The attention mechanism in MA-Net allows it to capture complex contextual dependencies and the multi-scale feature fusion allows it to capture the channel dependencies between feature maps for effectively segmenting foreground from background. The detailed architecture of MA-Net has discussed in section 3.1 and MA-Net architecture of proposed model is shown in Figure 1.

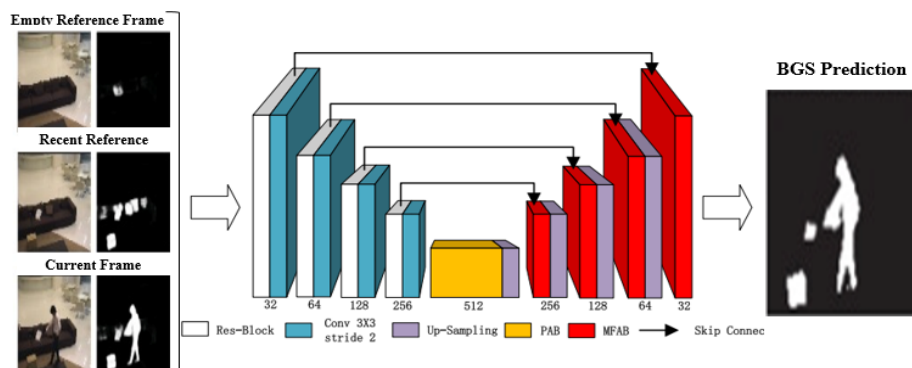


Figure 1. Proposed model of Ma-Net architecture

3.1 Ma-Net architecture

The Ma-Net (Multi-scale Attention Net) architecture proposed [13, 14] in this paper has two stages: the encoder stage and the decoder stage. The encoder stage is responsible for extracting features from the input image, while the decoder stage is responsible for reconstructing the foreground mask from the extracted features.

The encoder stage is comprised of many convolutional layers, which are designed to learn multi-scale features from the input image. In order to capture the dependencies between the features, we use a self-attention mechanism, which is applied after each convolutional layer. This allows the network to adaptively integrate local features with their global dependencies. The encoder stage of the MA-Net architecture is comprised of many convolutional layers, where batch normalization and activation by ReLU function is applied after each layer. The convolutional layers are designed to learn multi-scale features from the input image. To capture the dependencies between the features, we use a self-attention mechanism, which is applied after each convolutional layer. The self-attention mechanism is implemented using a PAB (Position-wise Attention Block) [15]. The PAB uses multi-head self-attention for modeling interdependencies between features in spatial dimensions, which represent the spatial dependencies between pixels in a global view. This allows the network to adaptively integrating local features with their global dependencies. The PAB (Position Attention Block) simulates a wide range of rich spatial contextual information across local feature maps as shown in Figure 2. Given a local

feature map, $I \in R^{H \times W \times 1024}$ as input, it is fed into a 3×3 convolution layer to obtain $I' \in R^{H \times W \times 2048}$. Then, 1×1 Convolution layers are utilized to generate $A \in R^{H \times W \times 256}$, $B \in R^{H \times W \times 256}$, and $C \in R^{H \times W \times 2048}$, respectively. A and B are reshaped to $R^{N \times 256}$ and $R^{256 \times N}$, respectively, and then a matrix multiplication is performed between $A \in R^{N \times 256}$ and $B \in 256 \times N$. Where N is the number of pixels. After that, the softmax function is used to obtain the spatial attention feature map, $P \in R^{N \times N}$. Where P_{ji} denotes the impact of the i^{th} position on the j^{th} position in the feature map as shown in Eq. (1).

$$P_{ji} = \frac{\exp(A_i, B_j)}{\sum_{i=1}^N \exp(A_i, B_j)} \quad (1)$$

The decoder comprises of many up-sampling layers, which are used to reconstruct the foreground mask from the extracted features. The decoder stage also uses the self-attention mechanism to focus specifically on significant features and to improve the accuracy of the foreground mask. The up-sampling layers in the decoder are followed by a convolutional layer and a batch normalization layer. To further improve the accuracy of the foreground, mask a self-attention mechanism is used in the decoder stage. We implement the self-attention mechanism using a MFAB (Multi-scale Fusion Attention Block) which uses a fused multi-scale semantic features for capturing channel dependencies between any feature maps and detail architecture shown in Figure 3.

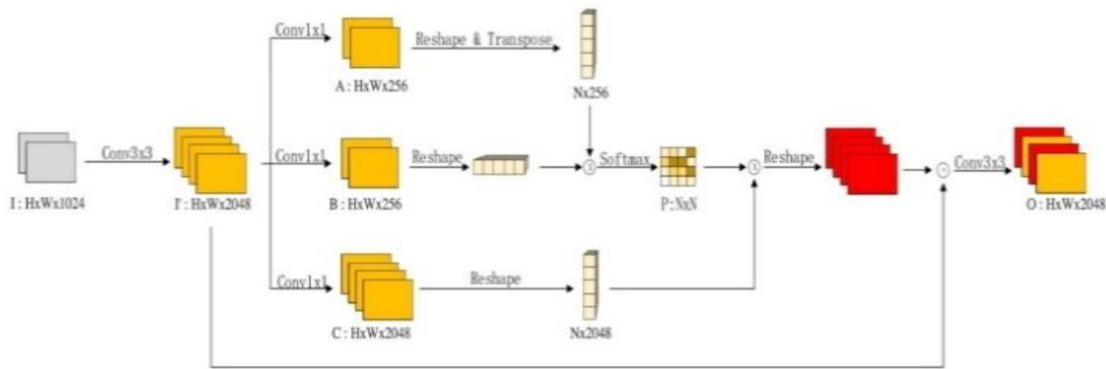


Figure 2. Position wise attention block: Used for foreground segmentation from background

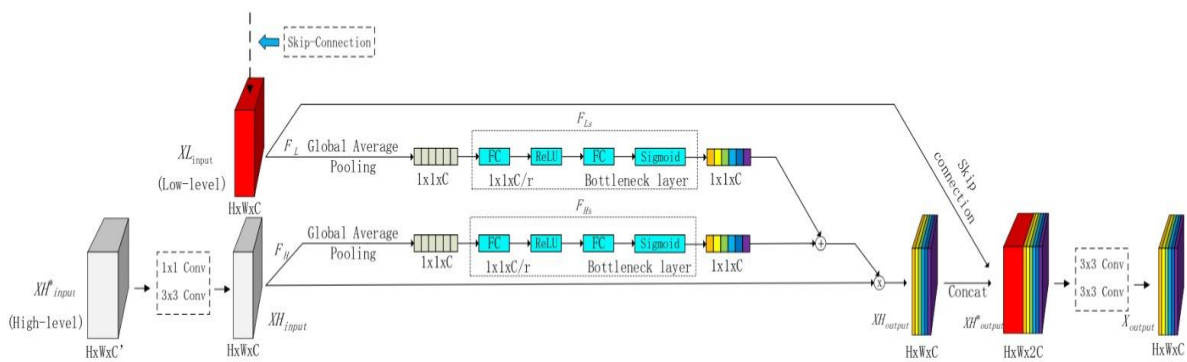


Figure 3. Multi scale fusion attention block: To capture the channel dependencies between any feature maps

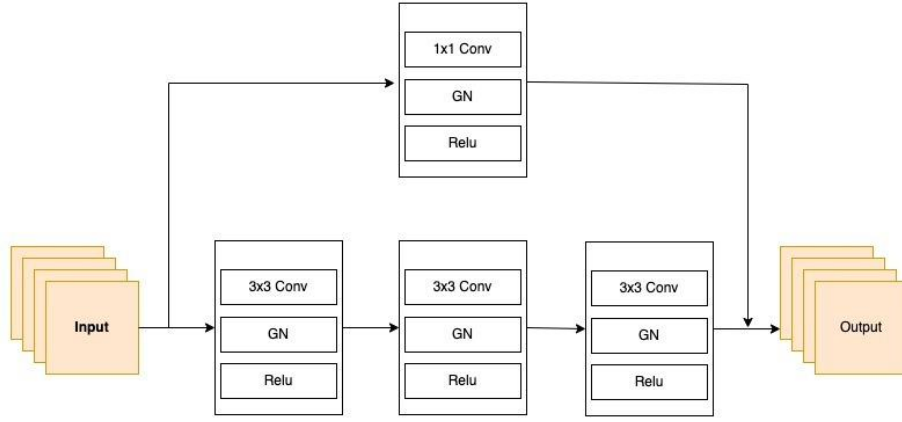


Figure 4. The architecture of the Res-block in our proposed method includes convolution layers, Group Normalization (GN), and a residual connection

Meanwhile, $C \in \mathbb{R}^{H \times W \times 2048}$ is reshaped to $C \in \mathbb{R}^{N \times 2048}$. A matrix multiplication is performed between the spatial attention map and $C \in \mathbb{R}^{N \times 2048}$, and the result is reshaped to $O' \in \mathbb{R}^{H \times W \times 2048}$. After that, an element-wise sum operation is performed between I' and O' . Finally, the final output, $O \in \mathbb{R}^{H \times W \times 512}$, is obtained through a 3×3 convolution layer. The final output, O , is as shown in Eq. (2).

$$O_i = \alpha \sum_{i=0}^N (P_{ij} C_i) + I'_j \quad (2)$$

The network produces foreground mask as the final output which indicates the pixels that belong to the foreground object.

In our approach, a deep neural network architecture is proposed for background subtraction that combines the ResNet50 encoder network with the self-attention mechanism in the decoder part. For capturing the spatial and channel dependencies of feature maps, we adopt two mechanisms based on self-attention concepts they are: the PAB (Position-wise Attention Block) and the MFAB (Multi-scale Fusion Attention Block). The PAB captures spatial dependencies between feature maps, while the MFAB combines channel dependencies between feature map by merging high and low-level feature information.

The ResNet50 architecture is based on the residual learning framework, which aims to alleviate the problem of vanishing gradients in very deep networks by using skip connections that bypass some of the layers. This allows the network to learn residual functions, which are the differences between the desired output and the input, rather than learning the underlying mapping function from the input to the output. In our semantic segmentation model, the ResNet50 architecture is used as the encoder, which is responsible for extracting features from the input image.

Inspired by the position attention module in the study [16], we use PAB to record the spatial interdependence between any two position feature maps. The PAB simulates a broader variety of rich spatial contextual information across local feature maps by using a 3×3 convolution layer, 1×1 convolution layers, matrix multiplication and softmax function. The resulting features are then passed to the decoder part of the network which uses upsampling layers to increase the spatial resolution and produce the final segmentation map.

Overall, our proposed approach combines the ResNet50 encoder network with a self-attention mechanism in the

decoder part which captures dependency at spatial and channel levels in feature maps that can improve the background subtraction's overall performance as shown in Figure 4.

4. EXPERIMENTAL RESULTS

For achieving better feature representation to segment background-foreground, we introduce a new network named MA-Net (Multi-scale Attention-Net) that incorporates a dual attention mechanism. In the experimental results section, the performance of the proposed MA-Net architecture is evaluated and compared with state-of-the-art methods on the CDNET 2014 dataset. The CDNET 2014 dataset is a challenging dataset for background subtraction, as it contains videos with various scenarios such as shadows, illumination changes, and dynamic backgrounds. In most background subtraction datasets, the count of pixels belonging to background is substantially more in comparison to the count of foreground pixels. This imbalance in the class creates significant issues for loss functions adapted commonly like as mean-squared error and cross-entropy. A viable option for imbalanced binary datasets is the Jaccard index. The Jaccard index is a better choice for as a loss function for probability maps. In this case, the Jaccard index measures the similarity between the predicted probability map and the ground truth probability map. The Jaccard index as a loss function for probability maps can be represented in Eq. (3).

$$L(P, Q) = 1 - J(P, Q) = 1 - \frac{|P \cap Q|}{|P \cup Q|} \quad (3)$$

where, P and Q are the predicted and ground truth probability maps, respectively, $J(P, Q)$ is the Jaccard index, and $L(P, Q)$ represents the loss function.

4.1 Dataset and evaluation details

For evaluating the performance of our proposed MA-Net (Multi-scale Attention-Net) network, we used the CDNET-2014 dataset. It has 53 videos under natural settings which are categorized into 11 categories by subsuming various challenging scenarios like night videos, dynamic backgrounds, shadows etc. The resolution (spatial resolution) of videos ranges from 320×240 to 720×526 pixels. Every video is annotated with the region of interest labeled as either 1)

foreground, 2) background, 3) hard shadow, or 4) unknown motion.

The performance of our network is evaluated by employing a 4-fold cross-validation strategy on as it is intuitive and simple. Where all videos from the data set are grouped into 4-folds in every category as evenly as possible. The proposed approach BGS algorithm on three of the folds and testing is performed on the fold which remains, this process is replicated for all four combinations, details of different categories along with videos chosen for training and testing samples as which are mentioned in Table 1. Where the entry “yes” in every fold corresponds to the video which has been chosen for testing and videos that fall under blank entries from that category are used for training purpose. Results obtained on full CDNet-2014 dataset could be uploaded in the evaluation server for comparing with the state-of-the-art methods.

When algorithm's performance, is measured, pixels having unknown motion labels was ignored and pixels having hard shadows is considered as background. This method of treating hard shadow as background is in line with CD-NET 2014 for change detection task. for evaluating performance, we use metrics listed by CD-Net 2014 namely F-score (F1), precision (Pr) and recall (Re).

4.2 Training details

To train the Ma-Net network for background-foreground segmentation, we have used the ADAM optimizer with a learning rate of 0.0005, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. The mini-batch size was set to 8 and the number of epochs was set to 50. The loss function used in the training process was a binary cross-entropy loss, which is commonly used for binary classification problems. The binary cross-entropy loss is calculated as the negative log-likelihood of the binary classification problem, and is given by the following Eq. (4):

$$L = -[y * \log(p) + (1-y) * \log(1-p)] \quad (4)$$

where, y is the ground truth label (0 for background and 1 for foreground), p is the predicted probability of the foreground, and \log is the natural logarithm. The binary cross-entropy loss

penalizes the model more when it predicts a higher probability for the wrong class. During the training process, data augmentation techniques were used to improve the robustness of the model, including random data augmentation at the beginning of each epoch, global illumination changes, and random Gaussian noise. The input frames were fixed to a size of 224×224 pixels, randomly cropped from the input frame, to leverage parallel GPU processing during training. In the evaluation step, thresholding with a threshold of $\theta = 0.5$ was applied to the output of the sigmoid layer of the network to obtain binary maps as shown in Figures 3 and 4 respectively. No scaling or cropping was applied to the inputs during evaluation. The visual results obtained for different categories of CDNet-2014 dataset has been shown in Table 2 and Table 3 respectively.

4.3 Comparison with state of the art

The proposed MA-Net architecture performance is measured by employing several metrics like recall, accuracy, F1-score and precision. The result obtained during evaluation of four-fold dataset is tabulated in Table 4, Table 5, Table 6 and Table 7, the overall result tabulated in Table 8 is obtained by taking the average of that. The proposed MA-Net architecture is compared with few state-of-the-art techniques. The results in Table 8 demonstrate that the proposed MA-Net architecture outshines other state-of-the-art techniques in terms of recall, F1 score, accuracy and precision. This signifies the effectiveness of the MA-Net architecture in dealing with challenging scenarios such as dynamic backgrounds, shadows, and illumination changes. Additionally, results also illustrate that the proposed MA-Net architecture fares well with various types of noises like salt-and-pepper noise, speckle noise, and Gaussian noise. This indicates that the proposed MA-Net architecture is able to handle noise effectively and maintain high performance. In summary, results obtained in the experiment signify the effectiveness achieved by the proposed MA-Net architecture in background subtraction and its ability to handle challenging scenarios such as dynamic background, shadows, and illumination changes. It also shows robustness to different types of noise as shown in Table 9.

Table 1. Categorization of CD-NET 2014 dataset in to four folds

Category	Video	Fold-1- Test	Fold-2 - Test	Fold-3- Test	Fold-4 – Test
Baseline	Highway	Yes			
	Pedestrians		Yes		
	Office			Yes	
	Pets2006				Yes
Camera Jitter	Badminton	Yes			
	Traffic		Yes		
	Boulevard			Yes	
	Sidewalk				Yes
Shadow	Copy Machine	Yes			
	Bus Station		Yes		
	Cubicle			Yes	
	People In Shade			Yes	
	Bungalows				Yes
Dynamic Background	Backdoor				Yes
	Overpass	Yes			
	Fountain02	Yes			
	Fountain01		Yes		
	Boats		Yes		
	Canoe			Yes	
	Fall				Yes

Table 2. Proposed method visual results of different standard datasets


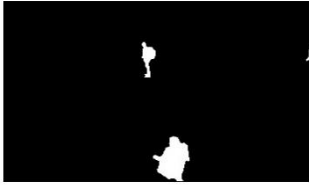
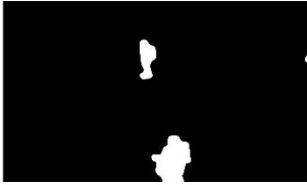

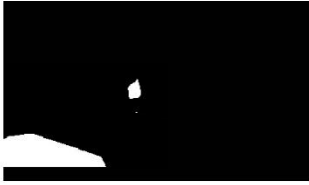



















Datasets	Current Frame	Ground Truth	Prediction of the Network
Baseline			
Camera Jitter			
Dynamic Background			
Low FrameRate			
Night			

Table 3. Proposed method visual results of different standard datasets

Datasets	Current Frame	Ground Truth	Prediction of the Network
PTZ			
Shadow			
Snowfall			

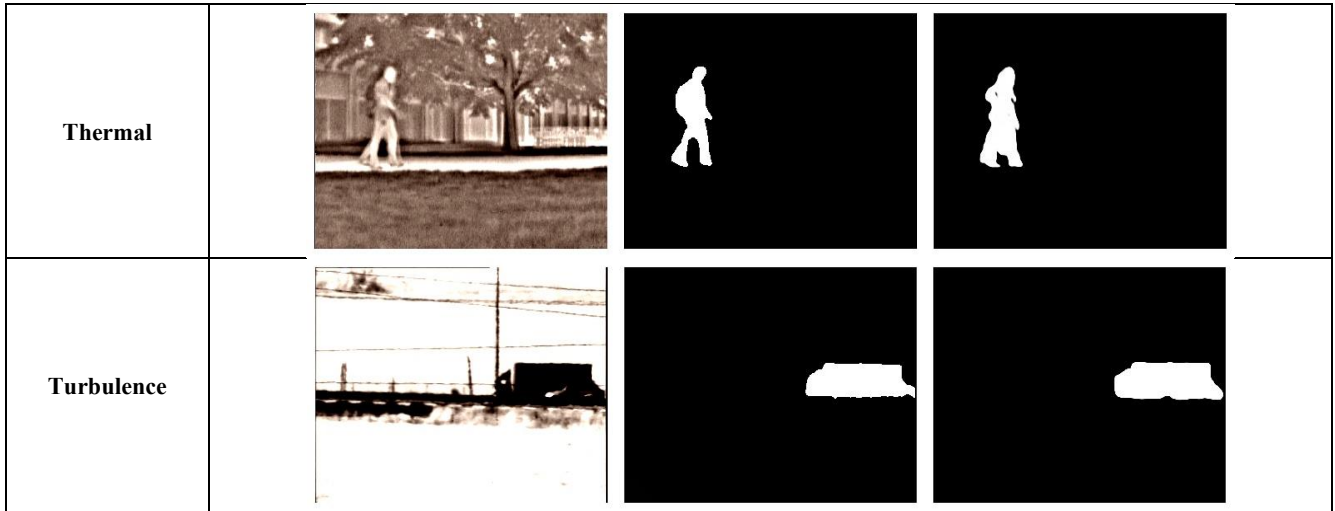


Table 4. Results obtained from evaluation metric on all categories of Fold-1

Category	Test Accuracy	Test_Precision	Test_Recall	Test_F-Score	Test_tp	Test_fp	Test_tn	Test_fn
Baseline	0.9986	0.999619	0.985434	0.992476	1580094	603.0005	15512442	23356
Camera Jitter	0.993275	0.859632	0.975335	0.913836	629889	102854	16914238	15929
BadWeather	0.986292	0.998001	0.579998	0.733636	337487	676.0005	17294592	244389
Dynamic Background	0.997761	0.985296	0.946439	0.965477	1089813	16264	33641910	61675
IntermittentObject Motion	0.985342	0.977089	0.757867	0.853628	743934	17444	16406484	237681
Low Framerate	0.99915	0.893557	0.146079	0.251107	1276	152.0005	8941883	7459
Night Videos	0.958127	0.488634	0.183775	0.267096	46667	48838	5813508	207268
PTZ	0.970965	0.376545	0.899484	0.53086	272710	451533	15846218	30475
Shadow	0.991531	0.973203	0.905536	0.938151	1141859	31441	16484728	119117
Thermal	0.982596	0.917748	0.761219	0.832187	1516803	135942	33020306	475794
Turbulence	0.998832	0.964105	0.680632	0.79794	41443	1543	17912985	19446
Full	0.989159	0.901564	0.837148	0.868162	7404206	808421	1.98E+08	1440358

Table 5. Results obtained from evaluation metric on all categories of Fold-2

Category	Test Accuracy	Test Precision	Test Recall	Test_F-Score	Test_tp	Test_fp	Test_tn	Test_fn
Baseline	0.999418	0.983866	0.967578	0.975654	205379	3368.001	17405157	6882
CameraJitter	0.984665	0.965294	0.887087	0.924539	1614669	58054	15309642	205524
BadWeather	0.986301	0.9956	0.886606	0.937947	1815975	8026	15484073	232258
DynamicBackground	0.997623	0.88194	0.951395	0.915352	339568	45456	26019379	17348
IntermittentObjectMotion	0.992784	0.837646	0.589735	0.692162	115075	22304	13967944	80055
LowFramerate	0.994702	0.98418	0.82561	0.897948	95121	1529.001	3964452	20092
NightVideos	0.976229	0.817799	0.490757	0.61341	373071	83118	18938496	387124
PTZ	0.965026	0.357007	0.28851	0.319124	146191	263299	17066428	360520
Shadow	0.995418	0.990207	0.898334	0.942035	642146	6351	16526547	72673
Thermal	0.965639	0.932061	0.922012	0.927009	3776288	275260	12935925	319416
Turbulence	0.997912	0.933114	0.537989	0.682488	40304	2889	17879802	34612
Full	0.986614	0.922832	0.840429	0.879705	9160918	766045	1.76E+08	1739373

Table 6. Results obtained from evaluation metric on all categories of Fold-3

Category	Test Accuracy	Test Precision	Test Recall	Test_F-Score	Test_tp	Test_fp	Test_tn	Test_fn
Baseline	0.991454	0.907625	0.980571	0.942689	1230276	125213	16124753	24377
CameraJitter	0.956779	0.812174	0.80199	0.80705	1290449	298433	12368896	318610
BadWeather	0.991395	0.975804	0.808483	0.884299	470973	11678	13728628	111566
DynamicBackground	0.995128	0.945234	0.980433	0.962512	1085086	62869	16178459	21656
IntermittentObjectMotion	0.961368	0.960358	0.619081	0.752849	84330	3481.001	1293531	51888
LowFramerate	0.987601	0.982232	0.750637	0.850958	373746	6761.001	10054271	124159
NightVideos	0.981729	0.664892	0.348118	0.456976	102184	51501	12946772	191349
PTZ	0.992236	0.174975	0.956022	0.29581	29217	137761	17749138	1344.001
Shadow	0.996261	0.988588	0.965322	0.976817	2729988	31513	31793931	98072
Thermal	0.976745	0.971838	0.776347	0.863162	1291933	37438	15912881	372184
Turbulence	0.999652	0.985066	0.692815	0.813489	13654	207.0005	17974836	6054
Full	0.988308	0.920852	0.868245	0.893775	8702507	747987	1.66E+08	1320588

Table 7. Results obtained from evaluation metric on all categories of Fold-4

Category	Test_Accuracy	Test_Precision	Test_Recall	Test_F-score	Test_tp	Test_fp	Test_tn	Test_fn
Baseline	0.994933	0.78705	0.944173	0.858482	274557	74286	17499351	16234
CameraJitter	0.959826	0.37334	0.899115	0.527603	119888	201235	5009391	13452
BadWeather	0.997485	0.991376	0.957936	0.974369	852761	7418	16939704	37446
DynamicBackground	0.984158	0.877624	0.947832	0.911378	1452159	202489	16091859	79926
IntermittentObjectMotion	0.974049	0.994486	0.797389	0.885098	164121	910.001	1435294	41702
LowFramerate	0.991096	0.972456	0.917482	0.94417	1209866	34268	14716633	108815
NightVideos	0.990564	0.924558	0.586284	0.717551	211353	17246	17256430	149143
PTZ	0.996714	0.939039	0.919141	0.928983	223342	14499	10134292	19648
Shadow	0.993938	0.989319	0.96282	0.975889	3201427	34565	22737303	123626
Thermal	0.990517	0.922443	0.704822	0.799081	335013	28167	17261480	140303
Turbulence	0.996071	0.984128	0.859583	0.917649	388651	6268	17293684	63488
Full	0.991832	0.937168	0.914133	0.925507	8434630	565498	1.56E+08	792291

Table 8. Comparison of methods for unseen videos from CDNet-2014

Method	Recall	Precision	F-Scores
Ours	0.8649	0.9206	0.8917
BSUV-net [17]	0.8203	0.8113	0.7868
WisenetMD [18]	0.8179	0.7535	0.7668
BSUV-net 2.0 [19]	0.8136	0.9011	0.8387
FgSegNet v2 [20]	0.5119	0.4859	0.3715
SWCD [21]	0.7839	0.7527	0.7583

Table 9. Comparison of methods according to the per-category F-Score for unseen videos from CDNet-2014

Method	Bad Weather	Low Frame Rate	Night	PTZ	Thermal	Shadow	IntObj Motion	Camera Jitter	Dynamic Background	Base Line	Turbulence	Overall
Ours	0.8825	0.7360	0.5137	0.5186	0.8553	0.9582	0.7959	0.7932	0.9386	0.9423	0.8028	0.8917
BSUV Net2.0 [11]	0.8844	0.7902	0.5857	0.7037	0.8932	0.9562	0.8263	0.9004	0.9057	0.9620	0.8174	0.8387
FgSeg NetV2 [12]	0.3277	0.2482	0.2800	0.3503	0.6038	0.5295	0.2002	0.4266	0.3634	0.6926	0.0643	0.3715
BSUV Net [16]	0.8713	0.6797	0.6987	0.6282	0.8581	0.9233	0.7499	0.7743	0.7967	0.9693	0.7051	0.7868
SWCD [17]	0.8233	0.7374	0.5807	0.4545	0.8581	0.8779	0.7092	0.7411	0.8645	0.9214	0.7735	0.7583
Wise net MD [18]	0.8616	0.6404	0.5701	0.3367	0.8152	0.8984	0.7264	0.8228	0.8376	0.9487	0.8304	0.7535

5. CONCLUSION

We have developed a new deep-learning technique to achieve background subtraction in videos that have not been seen before. We have also introduced a video-agnostic evaluation method for evaluation which treats every video from the dataset as unseen. The choice of cross-validation strategy is made as it is found to be highly beneficial in the evaluation of BGS algorithms in future. The input to the network includes the current frame, two reference frames from different time scales, and semantic information for all three frames. To enhance the generalization capacity of the network, we have formulated a simple yet effective illumination-change model. Our experimental results on the CDNet-2014 dataset show that the network outperforms current state-of-the-art video-agnostic background subtraction algorithms in terms of all performance metrics, indicating the potential of deep-learning algorithms for unseen or unlabeled videos. As a part of future work, we plan to further explore different

architectures and techniques for improving the performance under challenging categories such as "PTZ" and "Night videos," as well as investigate different background models for the reference frames. In the proposed work, we have focused on delivering a high-performance, supervised background subtraction algorithm especially for unseen category videos by not paying too much attention towards without computational complexity.

REFERENCES

- [1] Babae, M., Dinh, D.T., Rigoll, G. (2018). A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76: 635-649. <https://doi.org/10.1016/j.patcog.2017.09.040>
- [2] Elgammal, A., Harwood, D., Davis, L. (2000). Non-parametric model for background subtraction. In: Vernon, D. (eds) *Computer Vision — ECCV 2000*. ECCV 2000.

- Lecture Notes in Computer Science, vol 1843. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45053-X_48
- [3] Zivkovic, Z., Van Der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7): 773-780. <https://doi.org/10.1016/j.patrec.2005.11.005>
- [4] Piccardi, M. (2004). Background subtraction techniques: a review. In 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583), The Hague, Netherlands, pp. 3099-3104. <https://doi.org/10.1109/ICSMC.2004.1400815>
- [5] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. <https://doi.org/10.48550/arXiv.1804.03999>
- [6] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156-3164.
- [7] Heikkilä, M., Pietikainen, M. (2006). A texture-based method for modeling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4): 657-662. <https://doi.org/10.1109/TPAMI.2006.68>
- [8] Bilodeau, G.A., Jodoin, J.P., Saunier, N. (2013). Change detection in feature space using local binary similarity patterns. In 2013 International Conference on Computer and Robot Vision, Regina, SK, Canada, pp. 106-112. <https://doi.org/10.1109/CRV.2013.29>
- [9] Fan, T., Wang, G., Li, Y., Wang, H. (2020). Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8: 179656-179665. <https://doi.org/10.1109/ACCESS.2020.3025372>
- [10] Lee, S.H., Lee, G.C., Yoo, J., Kwon, S. (2019). WisenetMD: Motion detection using dynamic background region analysis. *Symmetry*, 11(5): 621. <https://doi.org/10.3390/sym11050621>
- [11] Tezcan, M.O., Ishwar, P., Konrad, J. (2021). BSUV-Net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction. *IEEE Access*, 9: 53849-53860. <https://doi.org/10.1109/ACCESS.2021.3071163>
- [12] Lim, L.A., Keles, H.Y. (2018). Foreground segmentation using convolutional neural networks for multiscale feature encoding. *Pattern Recognition Letters*, 112: 256-262. <https://doi.org/10.1016/j.patrec.2018.08.002>
- [13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [14] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146-3154.
- [15] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. MICCAI 2015. *Lecture Notes in Computer Science()*, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28
- [16] Tezcan, O., Ishwar, P., Konrad, J. (2020). BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Snowmass, CO, USA, pp. 2774-2783. <https://doi.org/10.1109/WACV45572.2020.9093464>
- [17] Işık, Ş., Özkan, K., Günal, S., Gerek, Ö.N. (2018). SWCD: A sliding window and self-regulated learning-based background updating method for change detection in videos. *Journal of Electronic Imaging*, 27(2): 023002-023002. <https://doi.org/10.1117/1.JEI.27.2.023002>
- [18] St-Charles, P.L., Bilodeau, G.A., Bergevin, R. (2016). Universal background subtraction using word consensus models. *IEEE Transactions on Image Processing*, 25(10): 4768-4781. <https://doi.org/10.1109/TIP.2016.2598691>