Illumination-Based Data Augmentation for Robust Background Subtraction

Dimitrios Sakkos* , Hubert P. H. Shum† and Edmond S. L. Ho‡
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, United Kingdom

***Abstract*—A core challenge in background subtraction (BGS) is handling videos with sudden illumination changes in consecutive frames. In this paper, we tackle the problem from a data point-of-view using data augmentation. Our method performs data augmentation that not only creates endless data on the fly, but also features semantic transformations of illumination which enhance the generalisation of the model. It successfully simulates flashes and shadows by applying the Euclidean distance transform over a binary mask generated randomly. Such data allows us to effectively train an illumination-invariant deep learning model for BGS. Experimental results demonstrate the contribution of the synthetics in the ability of the models to perform BGS even when significant illumination changes take place.**

***Index Terms*—Background subtraction, convolutional neural networks, synthetics, data augmentation, illumination-invariant**

I. Introduction

Background subtraction (BGS) has been an active research area in the past decades. The main task is to differentiate the foreground (i.e. moving objects) from the background (i.e. the static parts of a given scene) [1]. A large number of real-world applications, such as person re-identification [2], object tracking [3], gesture recognition [4], vehicle tracking [5], video recognition [6], crowd analysis [7] and even use cases of the medical domain [8], depend on accurate and robust background subtraction as a first step of their pipelines.

Sudden illumination changes signify a particularly difficult challenge, since they cannot be captured by a background model. Such changes in lighting conditions can be caused either by weather conditions or electric lights and result in color changes of a significant amount of pixels. Due to the difference of visual appearance in consecutive frames, BGS becomes inaccurate. The timing of these changes could be short, such as switching a light on/off, or a piece of cloud blocking the sun, making it tough for the system to adjust to the new condition in a timely manner.

State-of-the-art deep learning algorithms allow adapting to sudden illumination changes if a huge amount of training data is provided. However, obtaining labelled data is very costly and there is only limited datasets available in the community. As a solution, data augmentation methods are proposed to perform image-based operations on the data, such as mirroring or cropping, to synthesize a larger dataset. However, simple image tricks cannot effectively generate images with realistic illumination changes. Another solution is adding a small amount of noise to create a new, synthetic image that is similar to the original in context but different in color distribution. However, since the added noise does not have any semantic meaning, the synthetic images only slightly increase the generalisation power of the model, as they do not offer any additional knowledge of different lighting conditions of the same scene.

To overcome this challenge, we propose a new data augmentation technique by synthesising the light-based effects of different degrees of brightness. Such effects include shadows and halos of different size, placed in random locations of the input image. In addition, global illumination changes are also included, in order to increase the generalisation abilities of the model to scenes filmed at various times of the day and night. Such augmented data allows us to provide extra semantic information to the BGS model in terms of illumination for better generalisation performance. The results show that the proposed technique is superior to regular augmentation methods and can significantly boost the segmentation results even in scenes that feature illumination conditions unseen to the model. Our experiments indicate that the proposed method improves the BGS results in our quantitative and qualitative evaluations on the benchmark dataset SABS [9].

The main contributions of this work can be summarized as follows:

- A novel synthetic image generation method for robust background subtraction under challenging illumination conditions.
- An illumination-invariant deep neural network for background subtraction.

This paper is organised as follows. A literature review on the task of background subtraction is given in Section II. Section III outlines the proposed method for synthetic generation covering local, global and combined changes. In Section IV we introduce the dataset and we describe the training settings of our models. Section V follows with the presentation of our results and discussion. Finally, Section VI provides the

* email: dimitrios.sakkos@northumbria.ac.uk
† email: hubert.shum@northumbria.ac.uk
‡ email: e.ho@northumbria.ac.uk, *corresponding author*

978-1-7281-2741-5/19/$31.00 ©2019 IEEE

conclusion and future work.

II. Related Work

In this section, we will first review the related research in background subtraction using traditional approaches such as Gaussian Mixture Models and Principal Component Analysis. Then, we discuss on recent deep-learning based research.

Performing background subtraction on video with illumination change has been explored in the literature. In particular, Siva et al. [10] demonstrated that the pixel intensity values affected by sudden local illumination change can be modelled by combining a GMM with a conditional probabilistic function based on an extension of Zivkovic et al. [11]. Akilan et al. [12] proposed a feature fusing approach to fuse the color distortion, color similarity, and illumination measures to improve the performance of the GMM-based model. Vosters et al. [13] presented a statistical illumination model based on a PCA-based approach [14]. Such an approach is similar to Pillet et al. [15] in which a spatial likelihood model for modelling the relationship of neighboring pixels is proposed. When new frames are analysed, the model is updated. While the aforementioned approaches tried to improve the performance of background subtraction on videos with illumination change, the GMM-based approaches fail when there is a significant illumination change in consecutive frames [16]. PCA-based approaches are more robust in handling illumination changes in general. However, the lack of semantic knowledge in the scene limits the performance of PCA-based approaches.

In the past few years, the area of computer vision has grown rapidly thanks to deep learning. Deep neural networks are now the best performing models in background subtraction considering accuracy and robustness. This task's primary objective is to perform binary pixel-wise foreground/background classification in a given image or video. Clearly, since pixel-wise precision is the target, attention to detail is required. Similar to our proposed methods, there is existing work taking advantage of spatio-temporal information to improve the performance. A 3D convolution-based approached is proposed by Sakkos et al. [17] to exploit the relationship between a block of 10-frame for background subtraction tasks. In [18], the background model of the Kernel Density Estimation-based system is updated using information from previous frames. Group property information is exploited in both spatial and temporal domains in the sparse signal recovery based approached proposed by Liu et al [19]. A recent work [20] further demonstrated incorporating spatio-temporal constraints to improve [21] results in better performance. A successful deep-learning system requires a huge amount of training data, and it is costly to obtain labelled data with significant illumination variation. In this paper, we focus on a data augmentation approach to create synthetic data for training an illumination-invariant BGS network.

III. Methodology

In this section, we explain how we synthesise images of different illumination with both local and global changes, and then combine them as a unified augmentation method that covers all scenarios simultaneously.

A. Local changes

To synthesize local changes of illumination, we generate the synthetic images by locally altering the illumination of the input image, therefore creating either a "lamp-post" light source or a shadow effect. First, we randomly select a pixel of the image that serves as the centre of the illumination circle to be drawn: $p = I(w, h), w \in W, h \in H, I = W \times H$, where W, H the width and height of the input image I respectively. Once the coordinates of the centre pixel are determined, we randomly select the diameter d of the illumination circle. Since we want our model to be robust to both small and large shadows and flashes of light, we choose the diameter to be between one fifth and half of the smallest dimension of the input image: $d = k \times \min(W, H), k \in (\frac{1}{5}, \frac{1}{2})$.

Since modifying all pixels within the circle uniformly generate unrealistic results, we proposed a more sophisticated approach to model the effect of the light. First, we calculate the binary mask M_1 of the pixels to be altered using the following formula:

$$M_1(x, y) = 1 \Leftrightarrow (x - w)^2 + (y - h)^2 = d^2 \tag{1}$$

This means that the pixels of our mask have the value of 1 if they reside within the drawn circle and zero everywhere else. We then use the Euclidean Distance Transform (EDT) to model the light attenuation. Given a binary mask B, EDT is defined as:

$$EDT_x(B) = min_b(||x - b||_{L_2}), \quad \forall b \in B, \tag{2}$$

where L_2 is the Euclidean norm. Now, we can calculate the mask for local changes M_2 by applying the EDT on M_1:

$$M_2 = EDT(M_1) \tag{3}$$

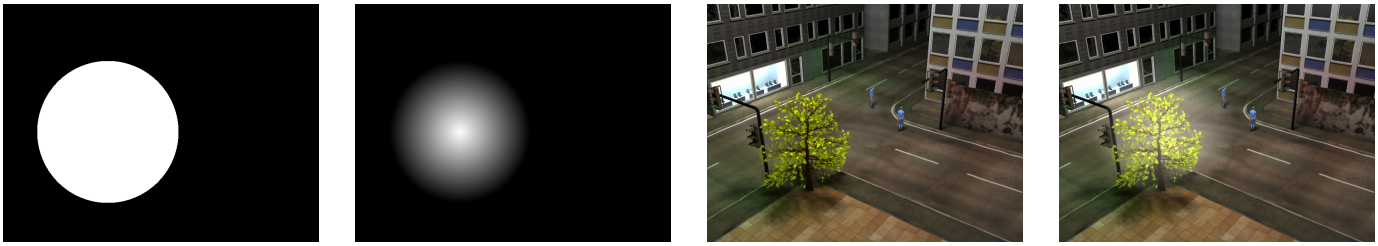
Once the new mask has been created, we proceed to alter the pixels of the original image that lie within the circle. The new synthetic image I_s is calculated as:

$$I_s = I \pm (M_2 \times z), \quad z \in [120, 160], \tag{4}$$

where I the original image, M_2 the mask calculated with the distance transform, z a random integer, and $\pm$ is either pixel-wise addition or subtraction, chosen with probability $p = 0.5$. When the addition operation is chosen, a lamp-post effect will be created in a random part of the image. Conversely, the subtraction operation creates shadows. The application of the aforementioned local masks are depicted in Figure 1. It can be seen that the final light source effect looks realistic.

B. Global changes

In some cases, global illumination changes can occur. For example, a lightning during a storm may instantly increase the brightness, and once the rain is over the global illumination will change again. In order to model such illumination changes, we need to alter the pixels of the whole image, rather than a small patch.



(a) The mask M_1 (b) The mask M_2 (c) Original image (d) After effect

Figure 1: The application of the mask for local changes. Subfigure (a): the initial binary mask M_1 is created by a circle of diameter $d = 179$ and centre coordinates $(322, 265)$. Subfigure (b): The mask M_2 after the application of the Euclidean distance transform on M_1. Subfigures (c) and (d) depict the original image and the lamp-post light source effect after the application of the mask M_2 on the input image respectively.

We synthesize global illumination changes as:

$$I_s = I \pm z, \quad z \in [40, 80], \tag{5}$$

where I, z and $\pm$ are as previously defined. In this case the illumination noise z needs to be slightly diminished, since the whole image is affected.

C. Combined changes

To capture both local and global illumination changes in the scene, we combine equation 4 and equation 5 into the following:

$$I_s = z_1 \pm (I \pm (M_2 \times z_2), z_1 \in [40, 80], z_2 \in [120, 160] \tag{6}$$

Sample images synthesised from our system can be found in figure 2. Since both the positioning and the intensity of the masks is random, this method can effectively cover all kinds of illumination changes. Additionally, hundreds of different synthetic images can be generated from a single frame. Therefore, given a small video, we can generate enough unique synthetic images to train a very deep network.

D. Illumination-invariant Deep Networks

We utilise the synthetic images to train multiple deep learning networks for BGS and evaluate their performances. Due to the use of images with synthetic illumination changes, these networks become invariant to lighting conditions.

To ensure the fairness of our experiments, all models used in this study have the same architecture. We follow the paradigm of previous background subtraction approaches [22], [23] and use a Unet architecture, which comprises an encoder and a decoder. We employ transfer learning and use the VGG16 model [24] as the encoder. Therefore, the weights of the encoder are initialised from those of VGG16, which has been pre-trained on Imagenet. VGG16 encompasses 13 convolutional layers, 5 pooling layers and 3 fully connected layers. Following the work of Long et al. [25], we remove the fully connected layers and make the network fully convolutional. Because of the pooling layers, the output of the encoder is 5 times smaller than the input. We use the decoder to recover the information that is lost from the downsampling operation via the use of upsampling blocks. Each block consists of a 2x2 bilinear interpolation operation which upsamples the feature maps, followed by two 3x3 convolutional layers with batch normalisation applied in-between (Figure 3b). To maximise the information recovered by the encoder, we add skip connections that connect the encoder to the decoder. In addition, the ReLu non-linearity is applied after each convolutional layer. Finally, once the spatial size has been restored, we add a final 3x3 convolutional layer, followed by a sigmoid layer to convert the output of the model to a foreground probability map. The architecture of the network is illustrated in Figure 3.

IV. Experiment settings

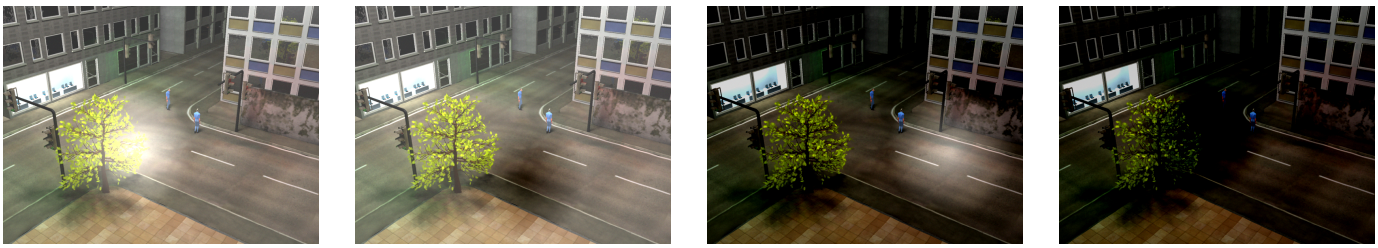
A. Dataset

In order to highlight the robustness of our augmentation process, we select the Stuttgart Artificial Background Subtraction dataset (SABS) [9].

The SABS dataset [9] contains 9 synthetic video sequences. The main challenge of the dataset stems from the sudden change of illumination over time. Although the foreground movements are the same in some sequences, the illumination is changing over time. In addition, different videos have very different lighting conditions, such as day-time and night scenes. In our experiments the sequence *Darkening* is used for training our models, which consists of 800 frames. The illumination of this scene is gradually changing from evening to night. For testing, the *Light Switch* video is used, which comprises 600 frames. This sequence only has night scenes and features light-switch effects in the middle of the video, where a store light is suddenly switched off. Since this effect is not present in the training video, *Light Switch* is an excellent candidate for measuring the generalisation abilities of our trained models. The rest of the video sequence in the SABS dataset [9] are not used because they are either day-time scenes and/or do not have significant illumination changes over time. Some example frames of the training and testing set are depicted in figure 5.

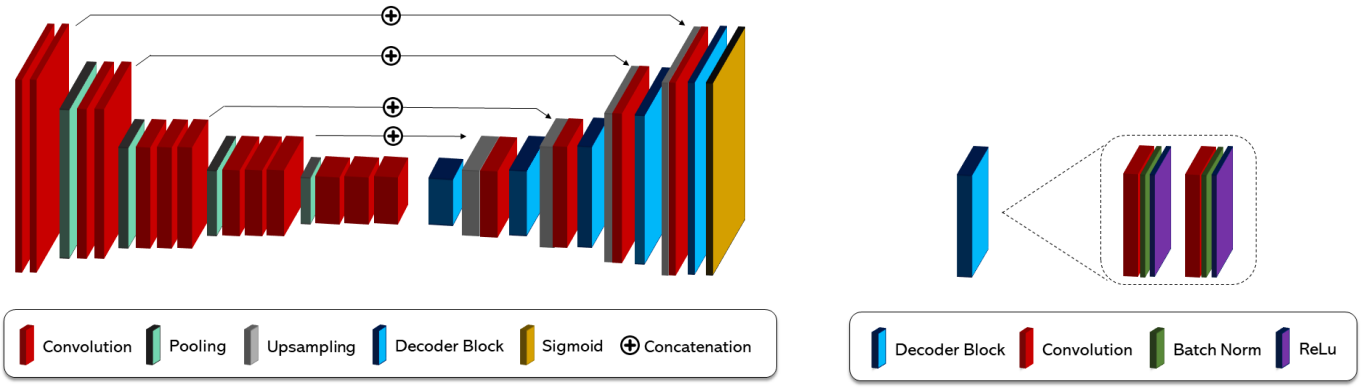
B. Training parameters

As mentioned above, we use the *Darkening* video sequence for training the model and the *Light Switch* video for testing. All models are trained with the same parameters. The initial



(a) G_bL_b (b) G_bL_d (c) G_dL_b (d) G_dL_d

Figure 2: Combination of global and local illumination changes. The subfigures (a) and (b) depict a combination of a brightening global filter with a bright and dark local filter respectively. On the other hand, subfigures (c) and (d) implement the darkening global filter.



(a) Overview (b) Decoder block

Figure 3: The CNN that was used for the experiments. The encoder is initialised from VGG16 [24] and is keep fixed during training.

learning rate is $lr = 0.001$ and is reduced by a factor of 0.1 if the model does not improve for 2 epochs. The training process ends after 5 epochs of no improvements. For optimisation, the Adam optimiser [26] is selected with betas $b_1 = 0.9$ and $b_2 = 0.999$. Finally, the batch size is set to 1.

To avoid overfitting, we freeze the encoder of our network. Specifically, the first 5 convolutional blocks of VGG16 are fixed and we only train the decoder. This training procedure yields better results according to our experiments.

Furthermore, for all augmenters, the probability of each training sample being augmented is set to 66.7%.

Most frames contain many more pixels of the background than the foreground - some frames might not depict any moving objects at all. Given this observation, the loss function needs to balance the classes as to not allow the model to be biased towards the background class. Therefore, we use the weighted cross-entropy loss, which is formally defined as follows:

$$G_s = wt[-\log\sigma(x)] + (1-t)[-\log(1-\sigma(x))], \quad (7)$$

where w is the weight coefficient, x is the predicted label, t is the target label and $\sigma(x) = \frac{1}{1+e^{-x}}$ is a sigmoid function. The weight w is calculated according to the ground truth frames with the following formula:

$$w = \frac{N}{2 \times [N_b, N_f]}, \quad (8)$$

where N denotes the number of pixels of all input frames and N_b,N_f are those pixels that belong to the background and foreground respectively.

C. Implementation details

We use the Keras library [27] for training our models. In addition, for the quick deployment of the proposed model, the *Segmentation models* [28] library is used. The Graphics Processing Unit (GPU) that was used in all our experiments is a GeForce GTX TITAN X.

D. Evaluation Metric

For evaluating our experiments, we use the following metrics: *F-Measure (FM), Intersection over Union (IoU), Matthews correlation (MC)*. We provide the formal definitions below:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

Settings	Recall ↑	Sp ↑	FPR ↓	FNR ↓	PWC ↓	FM ↑	Precision ↑	IoU ↑	Matthews ↑
L_a	0.5467	0.9962	0.0038	0.4533	1.4290	0.6412	0.7752	0.4719	0.6442
L_b	0.5958	0.9951	0.0049	0.4042	1.4219	0.6619	0.7444	0.4946	0.6589
L_c	0.6294	0.9954	0.0046	0.3706	1.3189	0.6903	0.7643	0.5271	0.6870

(a) Ablation studies for local changes

Settings	Recall ↑	Sp ↑	FPR ↓	FNR ↓	PWC ↓	FM ↑	Precision ↑	IoU ↑	Matthews ↑
G_{low}	0.7103	0.9927	0.0073	0.2897	1.3877	0.7051	0.6999	0.5445	0.6980
G_{med}	0.7082	0.9942	0.0058	0.2918	1.2464	0.7263	0.7454	0.5703	0.7202
G_{high}	0.6679	0.9952	0.0048	0.3321	1.2405	0.7155	0.7704	0.5570	0.7111

(b) Ablation studies for global changes

Table I: Ablation studies for local, global and combined changes

$$FM = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{12}$$

$$MC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{13}$$

where TP, TN, FP, FN denote the true positive, true negative, false positive and false negative pixels respectively.

V. Results

We perform extensive evaluations on the proposed method. In particular, a wide range of different augmentation settings (Table II) were evaluated. We also compare against the regular augmentation techniques. We implement a "default" augmenter which performs the following image transformations: *horizontal flipping, random cropping* and *noise addition*, as depicted in Figure 4. The *cropping* operation performs center cropping with random image sizes, whereas the *noise* option adds salt and pepper noise drawn from a Gaussian distribution. The amount of noise is fixed to 0.05. All operations have a 50% probability of taking place.

In the following section, we will evaluate the proposed method quantitatively to determine the optimal settings.

A. Quantitative Evaluations

In this experiment, we evaluate the performance of the proposed method using the commonly used metrics stated in section IV-D on the SABS dataset. The results of our experiments are presented in Table III. Even though the default augmenter improved the F-Measure by more than 7%, the proposed model, named GL, outperformed it by a very large margin of 16%. As a matter of fact, the proposed method obtains better results in every single metric. This highlights the effectiveness of our proposed method.

In addition, to determine the optimal settings, an ablation study is conducted and the details are explained in Section V-C.

Figure 4: Default augmentation techniques (from left to right): image mirroring, center cropping and adding noise.

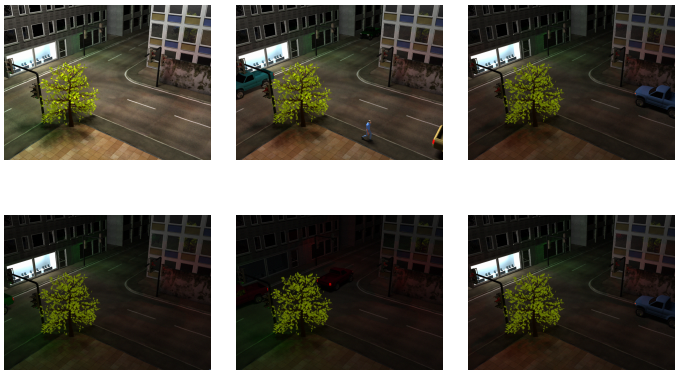


Figure 5: The SABS dataset that was used for evaluating the models. The first row depicts the training sequence *Darkening*, while the second row shows the testing video *LightSwitch*. The columns show frames from the start, middle and ending parts of the video. Note that in the middle of the *LightSwitch* sequence the store light switches off, causing major changes to the background.

B. Qualitative Evaluations

To provide visual comparison between the various models tested in this paper, we report the segmentation results of three frames taken from the start, middle and end of the video. The results are illustrated in Figure 8. The comparison indicates that the proposed data augmentation approach improves the quality of the segmentation masks, as it obtained solutions which are clearly closer to the ground truths compared to other models.

Name	Description	Threshold
baseline	No augmentation	0.8
default	Common augmentation: Mirror, crop and noise	0.7
L_a	Local changes with $z \in (80, 120), k \in (1/2, 2/3) \times G$	0.7
L_b	Local changes with $z \in (80, 120), k \in (1/5, 1/2) \times G$	0.7
L_c	Local changes with $z \in (120, 160), k \in (1/5, 1/2) \times G$	0.6
G_{low}	Global, low intensity changes with $z \in (20, 60)$	0.9
G_{med}	Global, medium intensity changes with $z \in (40, 80)$	0.6
G_{high}	Global, high intensity changes with $z \in (60, 100)$	0.8
GL	Global and local changes with $z_{global} \in (40, 80)$ and $z_{local} \in (120, 160)$	0.7

Table II: The different augmentation settings that were tested in our experiments. Parameters k, z and G denote the kernel size of the mask M_1, the illumination intensity in terms of pixel values and the resolution of the smallest dimension of the input image respectively. The last column shows the threshold that maximised the F-Measure of the segmentation mask.

Setting	No augm	Common augm	GL
Recall ↑	0.4606	0.5440	0.7687
Sp ↑	0.9933	0.9937	0.9941
FPR ↓	0.0067	0.0063	0.0059
FNR ↓	0.5394	0.4560	0.2313
PWC ↓	1.9172	1.6767	1.1189
FM ↑	0.5288	0.6025	0.7624
Precision ↑	0.6207	0.6750	0.7562
IoU ↑	0.3594	0.4311	0.6161
Matthews ↑	0.5253	0.5976	0.7567

Table III: Comparison between no augmentation, common augmentation and the proposed method which covers global and local illumination changes.

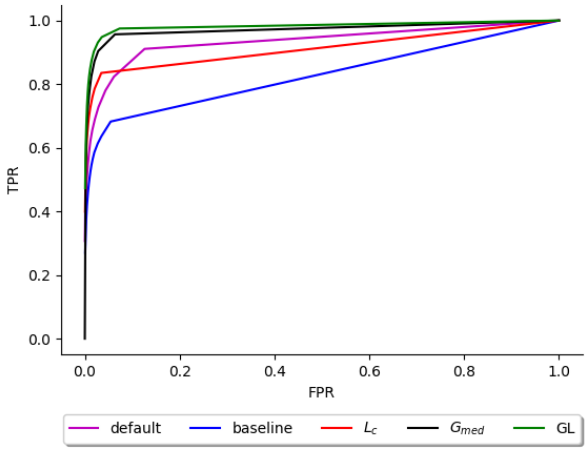


Figure 6: ROC curve

C. Ablation studies

Here we discuss the ablation studies that verify the optimal hyper-parameters of our method. The full list of our experiments can be found in Table II, whereas the result of each method is shown in Table I. By cross-checking these tables, it can be seen that using a smaller kernel is better for creating local effects. Also, greater changes in illumination yield better results. This is because of the fading effect caused by the distance transform, which is only strong in the centre of the circle. For global changes a much smaller noise value is needed, firstly because there is no fading and secondly due to the effect being applied to the whole image.

It is evident from Table I that both local and global change augmentations yield significant results. However, the best performing model according to our experiments encompasses both.

It is noteworthy that sub-optimal settings need a very high threshold to produce a good segmentation result. This is because in those frames of the *Light Switch* sequence where the light suddenly switches off, the model fails to identify the moving object due to low lighting. However, with the optimal settings of the proposed method, the model can generalise in all illumination conditions. The performance of each model under different threshold is depicted in Figure 7.

VI. Conclusion and Future Work

In this paper, we have presented a fast and easy method to synthesise training samples for the implementation of illumination-invariant models. The synthetic images are generated by artificially altering the pixel intensity values of the input image not only globally but also in small regions. A typical "lamp-post" light source effect can be approximated by applying the distance transform on a binary mask. We have tested the proposed method in the task of background subtraction. The experimental results indicate that the models trained using the dataset augmented with the new synthetics are more robust to illumination changes and are able to handle even intense lighting variations. As future work, more shapes can be explored, not only geometrical but also of arbitrary shapes, for representing shadows more realistically.

References

[1] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[2] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[3] Y. Shen, W. Hu, M. Yang, J. Liu, B. Wei, S. Lucey, and C. Chou, "Real-time and robust compressive background subtraction for embedded camera networks," *IEEE TRANSACTIONS ON MOBILE COMPUTING*, vol. 15, no. 2, pp. 406 – 418, 2016.

[4] H.-S. Yeo, B.-G. Lee, and H. Lim, "Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware," *Multimedia Tools and Applications*, 2013.

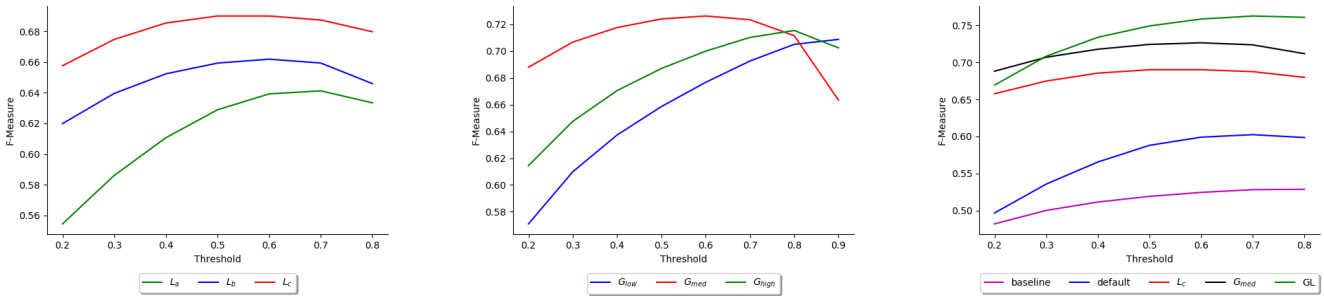


Figure 7: F-Measure values on different thresholds for each model.

[5] N. Sirikuntamat, S. Satoh, and T. H. Chalidabhongse, "Vehicle tracking in low hue contrast based on camshift and background subtraction," in *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, July 2015, pp. 58–62.

[6] C. Li and Y. Ming, "Three-stream convolution networks after background subtraction for action recognition," *Artificial Intelligence and Soft Computing*, pp. 12–24, 2019.

[7] X. Wang, C. Lu, J. Jia, and H. Li, "l_0 regularized stationary-time estimation for crowd analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 981–994, 2017.

[8] K. Them, M. G. Kaul, C. Jung, M. Hofmann, T. Mummert, F. Werner, , and T. Knopp, "Sensitivity enhancement in magnetic particle imaging by background subtraction," *IEEE TRANSACTIONS ON MEDICAL IMAGING*, vol. 35, no. 3, 2016.

[9] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," *CVPR*, 2011.

[10] P. Siva, M. J. Shafiee, F. Li, and A. Wong, "Pirm: Fast background subtraction under sudden, local illumination changes via probabilistic illumination range modelling," *2015 IEEE International Conference on Image Processing (ICIP)*, 2015.

[11] Z. Zivkovic and F. Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, pp. 773–780, 2006.

[12] T. Akilan, Q. J. Wu, and Y. Yang, "Fusion-based foreground enhancement for background subtraction using multivariate multi-model gaussian distribution," *Information Sciences*, vol. 430-431, pp. 414–431, 2018.

[13] L. Vosters, C. Shan, and T. Gritti, "Background subtraction under sudden illumination changes," *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010.

[14] B. R. N. Oliver and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[15] J. Pilet, C. Strecha, and P. Fua, "Making background subtraction robust to sudden illumination changes," *European Conference on Computer Vision*, 2008.

[16] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, 2014.

[17] D. Sakkos, H. Liu, J. Han, and L. Shao, "End-to-end video background subtraction with 3d convolutional neural networks," *Multimedia Tools and Applications*, vol. 77, no. 17, pp. 23 023–23 041, Sep 2018. [Online]. Available: https://doi.org/10.1007/s11042-017-5460-9

[18] D. Berjón, C. Cuevas, F. Morán, and N. García, "Real-time nonparametric background subtraction with tracking-based foreground update," *Pattern Recognition*, vol. 74, pp. 156–170, 2018.

[19] X. Liu, J. Yao, X. Hong, X. Huang, Z. Zhou, C. Qi, and G. Zhao, "Background subtraction using spatio-temporal group sparsity recovery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1737–1751, Aug 2018.

[20] S. Javed, A. Mahmood, S. Al-Maadeed, T. Bouwmans, and S. K. Jung, "Moving object detection in complex scene using spatiotemporal structured-sparse rpca," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 1007–1022, 2018.

[21] S. E. Ebadi and E. Izquierdo, "Foreground segmentation with tree-structured sparse rpca," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 2273–2280, 2018.

[22] L. A. Lim and H. Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognition Letters*, vol. 112, pp. 256 – 262, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518303702

[23] D. Zeng and M. Zhu, "Multiscale fully convolutional network for foreground object detection in infrared videos," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 617–621, 2018.

[24] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations (ICRL)*, pp. 1–14, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv e-prints*, p. arXiv:1412.6980, Dec 2014.

[27] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[28] P. Yakubovskiy, "Segmentation models," https://github.com/qubvel/segmentation_models, 2019.

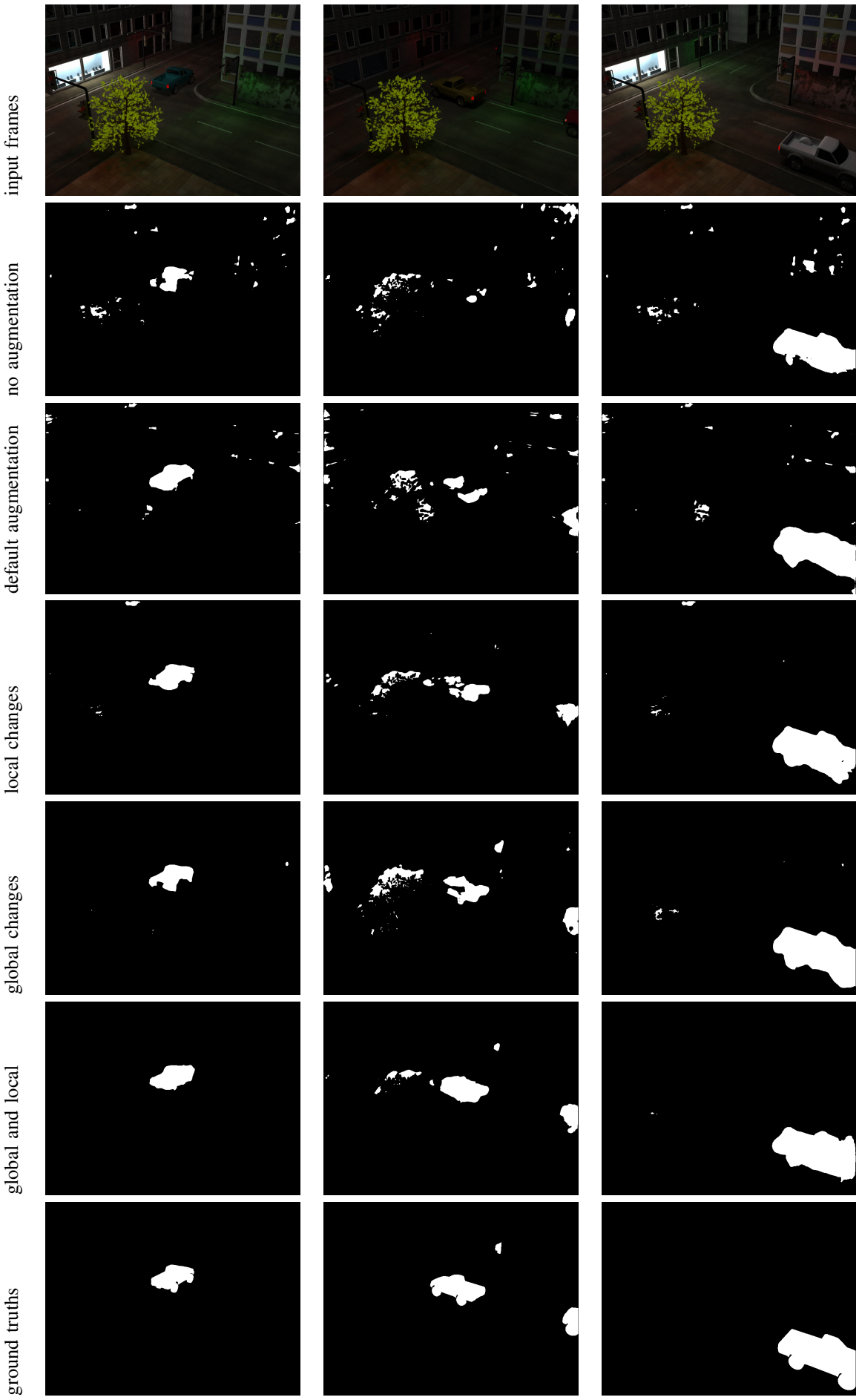


Figure 8: Comparison between different augmentation techniques.