IMAGE EDITING BASED DATA AUGMENTATION FOR ILLUMINATION-INSENSITIVE BACKGROUND SUBTRACTION

A PREPRINT

Dimitrios Sakkos
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, United Kingdom
dimitrios.sakkos@northumbria.ac.uk

Edmond S. L. Ho*
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, United Kingdom
e.ho@northumbria.ac.uk

Hubert P. H. Shum
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, United Kingdom
hubert.shum@northumbria.ac.uk

Garry Elvin
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, United Kingdom
garry.elvin@northumbria.ac.uk

July 14, 2020

ABSTRACT

A core challenge in background subtraction (BGS) is handling videos with sudden illumination changes in consecutive frames. While the use of data augmentation has been shown to increase robustness, the modelling of realistic illumination changes remains less explored and is usually limited to global, static brightness adjustments. In this paper, we focus on tackling the problem of background subtraction using augmented training data, and propose an augmentation method which vastly improves the model's performance under challenging illumination conditions. In particular, our framework consists of a local component that considers direct light/shadow and lighting angles, and a global component that considers the overall contrast, sharpness and color saturation of the image. It generates realistic, structured training data with different illumination conditions, enabling our deep learning system to be trained effectively for background subtraction even when significant illumination changes take place. We further propose a post-processing method that removes noise from the output binary map of segmentation, resulting in a cleaner, more accurate segmentation map that can generalise to multiple scenes of different conditions. Experimental results demonstrate that the proposed system outperforms existing work, with the highest F-measure score of 81.27% obtained by the full system. To facilitate the research in the field, we open the source code of this project at: https://github.com/dksakkos/illumination_augmentation

***K*eywords** Background subtraction; Convolutional neural networks; Synthetics ; Data augmentation; Illumination-invariant.

1 Introduction

The challenge in background subtraction (BGS) is to identify the pixels belonging to the background, which comprises the static areas in image such as the sky and roads, from the foreground, the areas that move against the background such as cars and humans [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], action recognition [7], crowd analysis [8] and even use cases from the medical domain [9–11], depend on accurate and robust background subtraction as a first step in their pipelines.

*Corresponding author

Sudden illumination changes provide 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 colour changes involving a significant number of pixels. Due to the difference in 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. In addition, a scene and the objects that appear in it will drastically transform during the night. It is necessary for an algorithm to be able to adjust in these kind of conditions.

State-of-the-art deep learning algorithms allow adaptation to sudden illumination changes if a huge amount of training data is provided. However, obtaining labelled data is very costly and there are 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 than the original in context but different in colour distribution. A major advantage is that each synthetic image will be unique, due to the added noise being random. However, the downside is that the added noise does not have any semantic meaning. Therefore, although the synthetic images do increase the generalisation of the model simply by obscuring pixels in the original image, they do not offer any additional knowledge regarding differing lighting conditions in the same scene. So the synthetic images only slightly increase the generalisation power of the model.

To overcome this challenge, we propose a new data augmentation technique which synthesises 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 pilot study, published in [12], demonstrated the effectiveness and feasibility of such an approach. In this paper, we extend the work by introducing new data augmentation techniques to handle additional variations to the input image locally and globally. A wide range of new experiments are conducted to evaluate the effectiveness of the new framework. 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.

We further propose a post-processing method that can successfully remove noise from the output binary map of segmentation. The method is based on the fact that contiguous frames have minimal changes between them and thus, the potential areas of the output that include foreground objects can be limited. Our experiments indicate that the proposed method improves the BGS results in our quantitative and qualitative evaluations on the benchmark dataset SABS [13].

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

- A novel synthetic image generation method which focuses on local and global image effects for robust background subtraction under challenging illumination conditions.
- An illumination-invariant deep neural network for background subtraction.
- A post-processing technique based on temporal coherence for the refinement of the segmentation results.

The rest of the paper is organised as follows. First, we review related work on background subtraction and particularly focus on illumination-aware systems in Section 2. Second, we present how we synthesize images by including local, global and combined illumination changes and present our new post-processing technique, in Section 3. Third, we present the dataset we created and explain how we train the network in 4. Next we present the experimental results and discuss the performance of our proposed methods in Section 5. Finally, we conclude the paper and discuss future directions in Section 6.

2 Related Work

In this section, we review works that are related to this paper. We first review traditional approaches of background subtractions, which involves statistical models like Gaussian Mixture Models and Principal Component Analysis. We then review background subtraction method that utilize deep-learning, with a particular focus on supervised methods that yield promising performance. We finally look at the problem of illumination in background subtraction and existing solutions, which motivate our research.

2.1 Traditional Approaches

Background subtraction (BGS) in video is a popular research topic, and the manipulation of illumination to improve accuracy has many research. Siva et al. [14] 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. [15]. Boulmerka and Allili [16] combine a GMM with inter-frame correlation analysis and histogram matching. Chen et al. [17] use a number of GMMs to construct spanning trees for hierarchical superpixel segmentation. They report that extending their model with optical flow for modeling temporal information increases the segmentation accuracy. Shen et al. [3] propose an efficient approach to BGS by reducing the dimensionality of the input data with a random projection matrix. Finally, they apply a GMM on the projected data.

Principal Component Analysis-related techniques are used for modelling the background of a video with an eigenspace. Since PCA retains the most significant eigenvectors, the foreground of the input image cannot be represented by the background model, as long as it is not static. The foreground can then be recovered with a difference image between the output of the model and the input frame [18]. Candès et al. [19] developed an efficient algorithm (RPCA) for decomposing the data into a low-rank matrix and a sparse matrix, which are representing the background and foreground in the BGS scenario, respectively. Recently, Ibadi and Isquierdo [20] extended RPCA by using a tree-structured sparse matrix to represent the input images. Although their method performs well on standard datasets, it fails in videos with sudden illumination changes like the *Light Switch* sequence of the *SABS* dataset.

The major weakness of GMM-based illumination-aware methods is that it cannot model significant illumination changes across consecutive frames as mentioned in [21], which happens frequently in real-world environments such as switching on and off the light. At the same time, while PCA methods have better robustness in modelling illumination changes, it lacks the semantic knowledge of the scene, resulting in a sub-optimal performance.

2.2 Deep-learning Based Approaches

Deep learning has improved system performance significantly in many areas. Here, we review how deep learning techniques and how they have been used in background subtraction in the past.

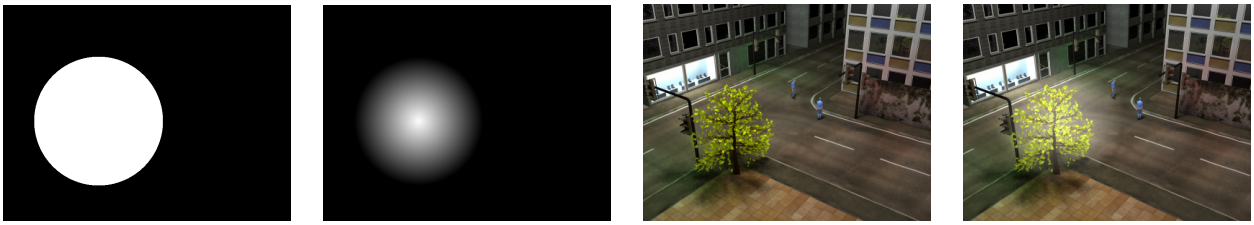
Deep learning approaches use variants of the fully convolutional network (FCN) proposed by Long et al. [22]. This is a special kind of convolutional neural networks with no fully connected layers, specifically designed for dense prediction tasks such as image segmentation. Most background subtraction methods follow the trend of recent generic image segmentation networks [23–27] and treat videos as a collection of images while disregarding the temporal information. Following the success of earlier approaches in object detection [28], image dehazing [29], segmentation [30] etc., Lim and Keles [31] and Zeng and Zhu [32] attempt to improve their binary maps by employing multi-scale feature aggregation. While [32] realise this idea simply by concatenating features from different layers, [31] employ multi-scale inputs, as previously done by Lu [33]. Wang et al. [34] also adopt the same preprocessing, but they refine the original CNN output by feeding it into another CNN.

A 3D convolution-based approached is proposed by Sakkos et al. [35] to exploit the relationship between a block of 10-frame for background subtraction tasks. In [36], 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 [37]. A recent work [38] further demonstrated incorporating spatio-temporal constraints to improve [20] results in better performance. We also employ deep-learning to perform background subtraction. However, we particularly focus on the problem of poor performance in different illumination conditions and propose methods to tackle it.

2.3 Dealing with Illumination

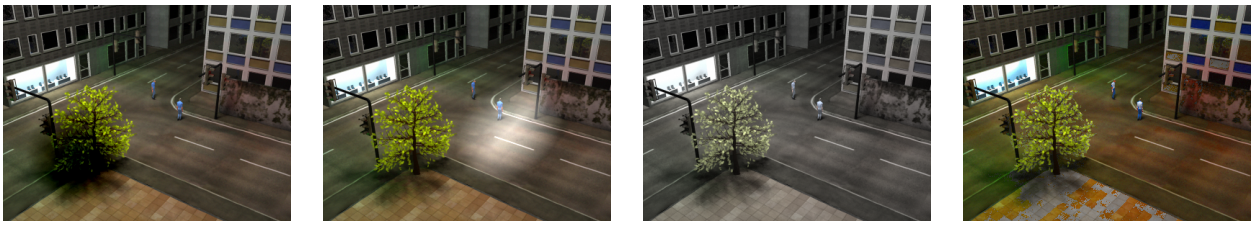
A major challenge for background subtraction is the illumination conditions, in which pixels belonging to the same object may look different under different illuminations. At the same time, deep-learning requires a huge amount of training data and annotating the lighting condition could be difficult and time-consuming. While existing deep-learning networks may automatically predict illumination information [39] and generate the illumination map [40,41], such approaches are typically used indoor with a limited number of light sources.

To solve this problem, Sakkos et al. [42] trained a multi-task generative adversarial network that modeled pairwise dark and bright images. The core idea is to apply gamma correction [43] to synthesize images under different illuminations, with a generative adversarial network [44] attempting to learn the data probability distribution for generating such images. A major advantage is that the method can synthesize pairwise images of different lighting conditions for effective training in both indoor or outdoor scene. However, such a method does not explicitly model the light sources that consist of different properties [45], and therefore generate images that may not necessarily realistic. In this work,



(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.



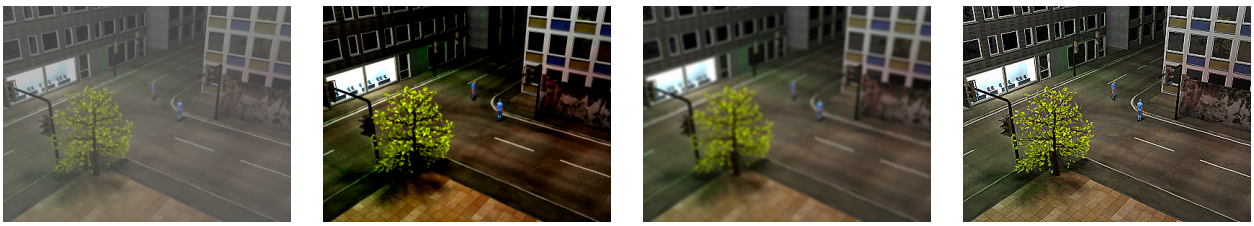
(a) Ellipsoid-shaped shadow (b) Ellipsoid-shaped flash of light (c) Low saturation (d) High saturation

Figure 2: (a) and (b): With the use of ellipsoid masks, we can produce a larger variety of effects. (c) and (d): The effect of colour saturation.

we solve the problem with a data augmentation approach, in which we perform image-based modelling of the light sources to synthesize different lighting conditions. Such a method create more realistic training images such that even a simple VGG16 [46] network can perform highly accurate background subtraction.

3 Methodology

In this section, we introduce the proposed data augmentation approaches and explain how a wide range of local (Section 3.1) and global (Section 3.2) illumination change effects can be synthesised using the proposed system. Our local illumination methods focus on editing a local region to simulate lamp-post and shadow effects caused by light sources from different angles (for example, Figure 1 and Figure 2 (a) and (b)). For global illumination change effects, we propose editing the entire image by altering the contrast, sharpness and image saturation (for example, Figure 2 (c) and (d), and Figure 3). In addition, we present a post-processing output refinement method (Section 3.4) that takes into account temporal information to further enhance segmentation results. Finally, we incorporate the data augmentation and output refinement approaches into a unified deep neural network architecture 3.5 to perform background subtraction.



(a) Low contrast (b) High contrast (c) Blurring (d) Sharpening

Figure 3: Global changes include contrast augmentation, blur and sharpness. Examples of after decreasing and increasing the effects are shown above.

Symbol	Description
I	input image
W	width of I
H	height of I
i	center of the local mask
d	diameter of the local mask
M_1	initial mask for local changes
k	kernel size of the mask M_1
z	illumination intensity in terms of pixel values
M_2	adjusted mask for local changes
c	contrast factor
F_b	blurring filter
F_s	sharpening filter
s	saturation parameter
c	contrast parameter
r	post-processing mask
p	model prediction
p'	refined model prediction
N	number of pixels of all input frames

Table 1: Table of symbols.

3.1 Local changes

Light sources can roughly be divided into *direct* and *indirect* lights. A direct light refers to light falling on a specific area or the surface of an object. In this research, we define this as a local illumination change since only a small region in the image is being affected. Our pilot study [12] demonstrated the effectiveness of an efficient approach to synthesize images with local illumination change. In this section, we first review the technique proposed in [12]. Next, we present a new local illumination augmentation approach to simulate a direct light source from different angles, including lamp-post and shadow effects.

3.1.1 Lamp-post and shadow effects

The shape of different light sources can vary significantly. However, circular shapes, such as street lights and light bulbs are common. In [12], we proposed to edit the intensity of a circular region on an image to simulate a "lamppost" light source (i.e. by increasing the pixel value) or a shadow effect (i.e. by decreasing the pixel value). Firstly, our method generates the center i of the circular region randomly:

$$i = I(w, h), w \in W, h \in H, I = W \times H \tag{1}$$

where W, H the width and height of the input image I respectively.

Next, the diameter d of the circular region is again determined randomly in order to create different illumination change effects. The range of the diameter size is defined as follow:

$$d = k \times \min(W, H), k \in (\frac{1}{5}, \frac{1}{2}). \tag{2}$$

where k is the kernel size of the binary mask M_1.

To replicate a realistic fading out effect near the edge of the circular region, we gradually decrease the brightening/darkening effect from the center to the edge (i.e. the attenuation of light). Specifically, we first calculate the binary mask M_1 of the pixels to be altered using the following formula:

$$M_1(x, y) = 1 \Leftrightarrow (x - w)^2 + (y - h)^2 \leq d^2 \tag{3}$$

Consequently, all pixels inside the circular region on the mask image will have the value of 1 and zero everywhere else. The light attenuation effect can then be simulated by using the Euclidean Distance Transform (EDT). Given a binary mask B, EDT is defined as:

$$EDT_x(B) = min_b(||x - b||_{L_2}), \quad \forall b \in B, \tag{4}$$

where L_2 is the Euclidean norm. By applying the EDT on M_1, the mask for local changes M_2 can be calculated as follows:

$$M_2 = EDT(M_1) \tag{5}$$

Having created the new mask for the local change, the intensity of the pixel inside the circular region (i.e. the masking region) on the image will be edited using the formula:

$$I_s = I \pm (M_2 \times z), \quad z \in [120, 160], \tag{6}$$

where I and I_s are the original and new synthetic image, respectively. z a random integer to further produce a wider variety of synthetic images, and $\pm$ is either pixel-wise addition or subtraction, chosen with equal probability. In summary, a lamp-post effect can be simulated by using the addition operation in Eq. 6. On the other hand, shadow effects can be created when the subtraction operation is used. An example of the circular masks M_1, M_2 and the image editing effect are illustrated in Figure 1.

3.1.2 Illumination angle effects

Although the method described above is very effective in creating a very realistic "lamp-post" lighting effect, it does not cover all variations of flashes and shadows. In reality, many lights often shine at an angle, resulting in a lit area that is not perfectly circular. As a result, the circular-shaped lighting effect has to be distorted and transformed in order to simulate the real-world effect. In computer graphics, such an effect can be created by estimating the 3D position and orientation of the light source in the virtual world, as well as the location and 3D shape of the objects in the scene. A realistically lit image can be generated by rendering the scene using techniques such as ray-tracing [47]. However, the aforementioned approach is computationally costly and requires detailed 3D information about the scene which is not available in 2D video and images. On the other hand, from our observation, the light effect closely resembles an ellipse rather than a circle in most cases. This motivates us to improve the realism of the lighting effect by using different mask shapes directly in the image space.

In particular, we propose using an ellipse-shaped mask in our simulation to cover these cases. Such a light-weight approach can effectively improve the realism of the lighting effect while minimizing the additional computation required for the framework in the training stage. More specifically, the following formula is used instead:

$$M_1(x, y) = 1 \Leftrightarrow \frac{(x-w)^2}{d^2} + \frac{(y-h)^2}{{d_0}^2} \leq 1 \tag{7}$$

where $d_0 = a_d * d, \quad a_d \in [0.3, 0.8]$. We also randomly rotate the ellipse along its axis.

3.2 Global changes

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

We synthesize global illumination changes as:

$$I_s = I \pm z, \quad z \in [40, 80], \tag{8}$$

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.

3.2.1 Contrast augmentation

In addition to image brightness augmentation, we further enhance the variance of our global augmentation settings with contrast changes. The contrast of an image plays an important role in highlighting different objects in the scene. Low contrast images usually look softer and flatter, as well as lacking shadows and highlights. In reality, various occurrences can result in low contrast images. One of the common situations is lens flare in the image, where a bright light source scatters the light directly into the lens. Inspired by this observation, we propose a new data augmentation approach that varies the contrast of the image to improve the robustness of the framework. Specifically, we alter the contrast of the original image by applying the following formula:

$$I_s = 128 + c * (I - 128) \tag{9}$$

where c is the contrast factor. For $c < 1$ the contrast is decreased; conversely we can increase it by setting $c > 1$. In our experiments, we let $c \in [0.2, 2]$. Example images are shown in figure 9.

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

Figure 4: 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.

3.2.2 Sharpness augmentation

Finding the sharp borders between different objects in the image will provide a clear separation between foreground and background and will certainly contribute positively toward the background subtraction task. In contrast, the blurring effects caused by low illumination as well as motion blur will have a negative impact. Inspired by this observation, we propose incorporating sharpness/blurring in the data augmentation algorithm.

Blurring can be easily achieved by convolving the original image with a $n \times n$ low-pass filter, which is averaging the neighboring pixels of the input image. Specifically, we use $F_b = \frac{1}{25} \times [1]^{5x5}$.

Sharpening an image can be done in the same manner, with the use of the filter

$$F_s = \frac{1}{9}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{10}$$

.

An example of the edited images is illustrated in Figures 3c-3d.

3.2.3 Color saturation

The color saturation of an image refers to the intensity of the color. The higher the saturation, the more colorful the image is. On the other hand, an image resembles a grey-scale one when the saturation is very low. Such a difference in image appearance is similar to the situation when illumination changes significantly. Here, we propose to further incorporate color saturation in the data augmentation process. This improves the framework by enhancing its robustness against significant changes in the color of the image caused by illumination changes.

We edit the saturation of the whole image by converting it from the RGB to the HSV colour space and directly changing the saturation attribute. Specifically, we scale the second dimension of the HSV space which corresponds to saturation using a parameter $s \in [0, 2]$. For $s < 1$, colour is diminished; conversely for $s > 1$ the colours become more saturated. An example of the edited image is illustrated in Figures 2c-2d.

3.3 Combined changes

To capture both local and global illumination changes in the scene, we combine Eq. 6 and Eq. 8 into the following:

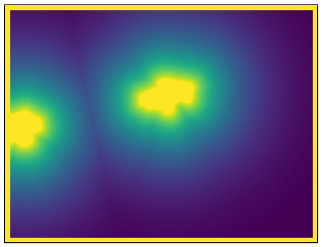
$$I_s = z_1 \pm (I \pm (M_2 \times z_2), z_1 \in [40, 80], z_2 \in [120, 160] \tag{11}$$

Sample images synthesised from our system can be found in figure 4. 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.

3.4 Output refinement via temporal coherence

While the proposed augmentation method works for still images and videos alike, in the latter case we can exploit motion information to refine the segmentation results as follows:

(a) Segmentation mask of the previous frame



(b) Probabilistic mask

Figure 5: The (a) probability mask that is used for refining the output, created from the (b) segmentation mask of the previous frame. Bright colours indicate high probability, whereas dark colours represent low probability values.

Lemma 3.1 *Let $o_t = (i, j)$ be a pixel of an object at time t. Then, the corresponding $o_{t+1} \in \{(i \pm \delta i, j \pm \delta j)\}$, where δ is a small integer.*

Based on this, we can create a refining map to highlight the areas of the input image that are likely to contain pixels of the foreground in the next frame. The map acts as a weight matrix that refines the probabilities of each pixel of the model output. Essentially, this refining map needs to be designed in a way that the predicted foreground probability of the next frame is not scaled down. This is a desired property since the change between two contiguous frames is minimal and most foreground pixels remain in the same class. Secondly, those pixels of the refining mask that are adjacent to foreground pixels need to be assigned with a probability value very close to 1, as it is highly possible for the foreground object to move into this area. As the distance becomes larger, the values need to be gradually scaled down. Eventually, the pixels that are furthest away from the foreground have the smallest probability.

Given each timestamp t and a video frame F_t, we can construct the refining mask r_t in the following way: First, we obtain the model output p_t, the pixels of which represent the probability of them belonging to the foreground class. Then, r_t can be generated by applying the Euclidean distance transform on p_t:

$$r_t = EDT(p_t), \tag{12}$$

where EDT is as defined in Eq. 4.

While this is a valid approach for existing moving objects, we need to account for new objects entering the scene at any moment. As a result, we set the values of r_t located around the border to 1. Therefore, the mask does not penalise new objects entering the frame. The end result of the refining mask r_t is depicted in figure 5.

Having defined the process of creating the probability map, the refinement is performed in a post-processing manner. During testing, we obtain the model output of the current frame and calculate the probability map, which is used to filter the model output on the next frame. Thus, p_{t+1} can be refined by scaling its probability values according to r_t as follows:

$$p'_{t+1} = p_{t+1} \times r_t, \tag{13}$$

where $\times$ is the pixel-wise multiplication operator and p'_{t+1} denotes the refined segmentation result.

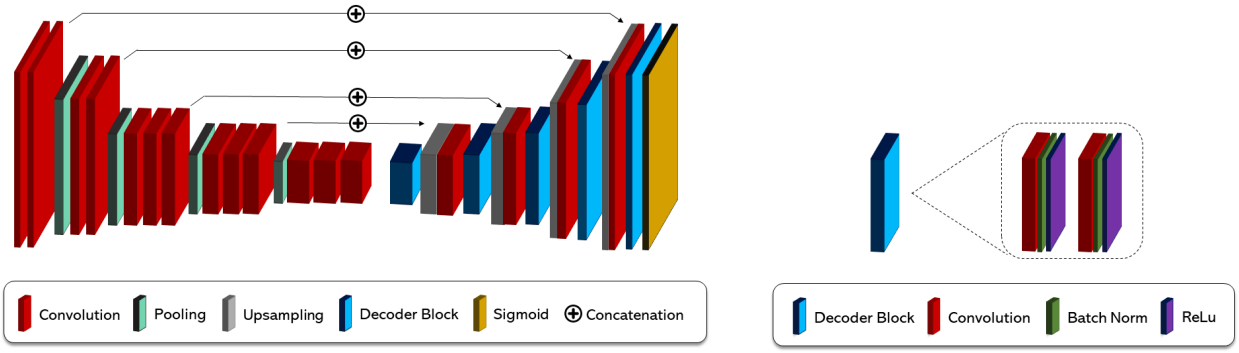
3.5 Illumination-invariant Deep Networks

In this work, we incorporate the proposed local and global data augmentation techniques into a deep learning framework for background subtraction. To keep all other variables fixed, the same network architecture is used for all experiments. We use a VGG16 backbone [46], which is transformed to a fully convolutional network by removing the fully connected layers and appending a decoder. As seen in Figure 6, the decoder mirrors the encoder and has the same number of channels. Additionally, the encoder features are concatenated with those of the decoder which are of the same size, to enable information flow. With the proposed illumination-focused data augmentation techniques, the proposed model is more robust on inputs with significant illumination changes. The experimental results are presented in Section 5.

4 Experiment settings

4.1 Dataset

In this work, a wide range of illumination-focus data augmentation techniques is proposed. In order to evaluate the performance of the new methods on BGS tasks, a dataset with significant illumination changes is needed. In particular,



(a) Overview

(b) Decoder block

Figure 6: The CNN that was used for the experiments. The encoder is initialised from VGG16 [46] and is keep fixed during training. ReLu layers are used after every convolution and are omitted from Figure (a) for clarity.

the Stuttgart Artificial Background Subtraction dataset (SABS) [13] fulfil all the requirements and it is used in other illumination-aware BGS approaches in the literature [12, 42].

The SABS dataset [13] contains videos with challenging illumination conditions and makes the BGS difficult. There is a wide range of environmental lighting conditions (such as day-time and night scenes), as well as other light sources (such as switching on/off the lights inside the shops in the street scenes). Following our pilot study [12], the sequence *Darkening* is used for training our models in all of our experiments. The *Light Switch* video is then used as the unseen testing sequence. In figure 8, a number of sample training and testing frames are illustrated. Note that we did not use the rest of the SABS dataset [13] since those videos do not contain any significant illumination changes and thus not suitable for evaluating our method.

Since the images are generated on-the-fly during training and are not saved on disk, the training dataset is slightly different for each experiment. However, when the same parameters are used, those differences become non-significant due to a large amount of generated data. The list of experiments and their hyper-parameters is given in Table 2.

4.2 Training parameters

In this section, the parameters used in training the deep learning models are explained. Firstly, a mini-batch approach with the batch size set to 1 is used. Next, we used the Adam optimiser [48] with betas $b_1 = 0.9$ and $b_2 = 0.999$. Thirdly, to avoid overfitting, the training process is terminated if there is no improvement after 5 epochs. The initial learning rate is $lr = 0.001$ and is reduced by a factor of 0.1 if the model does not improve for 2 epochs.

As another measure against overfitting, we freeze the encoder of our network. Specifically, the first 5 convolutional blocks of VGG16 are fixed and we only train the decoder. Therefore, although the total number of parameters is around 19M, we only train 4.3M.

The optimal ratio between the unaltered, original images and the new, augmented training samples is not the same for every problem. It can depend on the size of the training set and the variance of the images' appearance. In our case, we find that augmenting 2/3 of our training set yields the best results.

We observe that most frames contain more pixels of the background than the foreground - some frames might not even depict any moving objects at all. Considering this, we conclude that the loss function needs to balance the classes as to not allow the model to be biased towards the background class. As a solution, we use the weighted cross-entropy loss, which is formally defined as follows:

$$G_s = wt[-\log \sigma(x)] + (1-t)[-\log(1-\sigma(x))], \tag{14}$$

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:

$$w = \frac{N}{2 \times [N_b, N_f]}, \tag{15}$$

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.

4.3 Implementation details

We use the Keras library [49] for training our models. Furthermore, for the quick deployment of the proposed model, the *Segmentation models* [50] library is used. The full code is uploaded on GitHub [2]. The Graphics Processing Unit (GPU) that was used in all our experiments is a GeForce GTX TITAN X.

4.4 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:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$FM = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{19}$$

$$MC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{20}$$

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

5 Results

We perform extensive evaluations on the proposed method. In particular, a wide range of different augmentation settings (Table 2) 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 7. 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. To maximise the variance of the augmented data's appearance, 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.

5.1 Quantitative Evaluations

To quantitatively evaluate the performance of the proposed data augmentation techniques as well as the new post-processing method, we follow the commonly used metrics as in the previous work on the SABS dataset. The details are stated in section 4.4.

The results of our experiments are presented in Table 3 and 4. We first review the performance of the basic local and global data augmentation approaches presented in our pilot study [12]. In Table 3, the common augmentation approach achieved a 7% improvement in the main evaluation metric F-Measure (FM) over the baseline which does not employ any data augmentation, which indicates that general augmentation methods can improve results significantly. We further presented the results obtained by our combined (global and local) data augmentation, namely *GL*, proposed in our pilot study [12]. Our method achieved an outstanding 76.24%, which is significantly higher than the common augmentation approach by 16% and the baseline by 23%. Our method also outperformed other methods in every single metric. These results highlight the importance of targeted, task-specific data augmentation and demonstrates the superiority of the proposed method against illumination-agnostic augmentation.

Next, we focus on the new data augmentation approaches proposed in this work. The results, which are depicted in Table 4, show that all 4 of the newly proposed methods further improve the model performance in segmenting the

[2] `https://github.com/dksakkos/illumination_augmentation`

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
GL_{refine}	The GL model, after applying the post-processing method	0.6
GL_{sb}	GL plus sharpening and blurring augmentation	0.3
GL_s	GL plus colour saturation augmentation	0.3
GL_c	GL plus contrast augmentation	0.1
GL_e	GL with ellipsoid-shaped masks for local changes	0.3
GL_{all}	GL plus all the above	0.3
GL_{AD}	GL plus all the above plus default augmentation	0.2

Table 2: 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.

Settings	Recall ↑	Sp ↑	FPR ↓	FNR ↓	PWC ↓	FM ↑	Precision ↑	IoU ↑	Matthews ↑
No augm	0.4606	0.9933	0.0067	0.5394	1.9172	0.5288	0.6207	0.3594	0.5253
Common augm	0.5440	0.9937	0.0063	0.4560	1.6767	0.6025	0.6750	0.4311	0.5976
GL [12]	0.7687	0.9941	0.0059	0.2313	1.1189	0.7624	0.7562	0.6161	0.7567

Table 3: Comparison between no augmentation, common augmentation and method proposed in our pilot study [12] which covers global and local illumination changes.

foreground objects. The F-Measure values are ranging from 77.09% to 79.04%, a very good performance which is further increased to 79.96% when we combine all new methods. Finally, when the proposed augmentation methods are combined with the default ones, the model accuracy reaches an excellent 80.6%. That is an improvement of 4.36% FM score than *GL* proposed in our pilot study [12]. Therefore, we can deduce that each new method introduces modifications to the training data that offer improvements in different areas, and also all methods complement each other.

We further evaluate the performance of the proposed post-processing method and the results are presented in Table 5. Evidently, our method improves the result of all experiments. This improvement fluctuates between 0.1% and 0.84% with an average of 0.65%. Our best model reaches an F-Measure score of 81.27%. This shows that the post-processing can further boost the performance even the new augmentation techniques have already achieved a high accuracy level.

Settings	Recall ↑	Sp ↑	FPR ↓	FNR ↓	PWC ↓	FM ↑	Precision ↑	IoU ↑	Matthews ↑
Ellipsoid masks	0.7806	0.9942	0.0058	0.2194	1.0836	0.7709	0.7614	0.6272	0.7654
Sharp/Blur	0.7925	0.9941	0.0059	0.2075	1.0587	0.7776	0.7633	0.6361	0.7723
Saturation	0.7813	0.9945	0.0055	0.2187	1.0515	0.7763	0.7714	0.6344	0.7710
Contrast	0.7752	0.9955	0.0045	0.2248	0.9601	0.7904	0.8062	0.6535	0.7857
All	0.8097	0.9948	0.0052	0.1903	0.9478	0.7996	0.7898	0.6661	0.7948
All+Default	0.8191	0.9949	0.0051	0.1809	0.9211	0.8060	0.7932	0.6750	0.8014

Table 4: Individual contributions of new augmentation methods.

method	FM	Post-processed
Baseline (GL) [12]	0.7624	0.7697
Sharp/Blur	0.7776	0.7852
Saturation	0.7763	0.7846
Contrast	0.7904	0.7988
Ellipsoid masks	0.7709	0.7719
All	0.7996	0.8055
All+Default	0.8060	0.8127

Table 5: Accuracy improvements of our post-processing method. Numbers represent FM scores

Figure 7: Default augmentation techniques (from left to right): image mirroring, center cropping and adding noise.

5.2 Qualitative Evaluations

In this section, we visualize the segmentation results to evaluate the proposed augmentation methods qualitatively. We picked three representative frames from the start, middle and end in the testing video sequence for a fair comparison. The selected frames have different illumination conditions which allow us to evaluate the performance of the methods in all situations. The BGS results depicted in Figure 10 and 11 show that the proposed augmentation approaches generated higher quality segmentation masks. In particular, local illumination augmentation leads to masks with much fewer false positives, effectively suppressing noise. On the other hand, global augmentation offers significant improvement on the true positives, while discarding some noise as well. However, the results are further improved when combining global and local augmentation, with the model predictions showing minimal noise and being accurate even when the foreground is very dark. In this case, the contrast and colour/brightness difference between the foreground and background objects is very low, therefore our contrast and blurring augmentation helps the model segment the foreground more accurately.

5.3 Ablation studies

In this section, we justify the selection of hyperparameters used in this work. A comprehensive list of the experiments and the corresponding settings are presented in Table 2. In general, we investigate the kernel size used for local augmentation, in addition to the range of the pixel intensity change. Furthermore, we provide the optimal threshold which maximises the model performance. The quantitative results of ablation testing can be found in Table 6, while the performance of each model under different threshold is depicted in Figure 9.

6 Conclusion and Future Work

In this paper, we have proposed a data augmentation framework to generate structured training data for background subtraction on videos with significant lighting changes. Specifically, to improve the realism of image synthesis, we have proposed to separate the illumination changes into local and global components, and we proposed some novel designs to effectively model the respective illumination effects. We have further proposed a post-process technique to refine the background mask for generating more accurate results. Our framework has improved the training process and generalisation of a deep neural network for background subtraction, as it can provide an unlimited amount of training data that represents challenging illumination conditions. Experimental results have shown our method outperforming existing work, and achieving the highest score of 81.27% in all comparative setups.

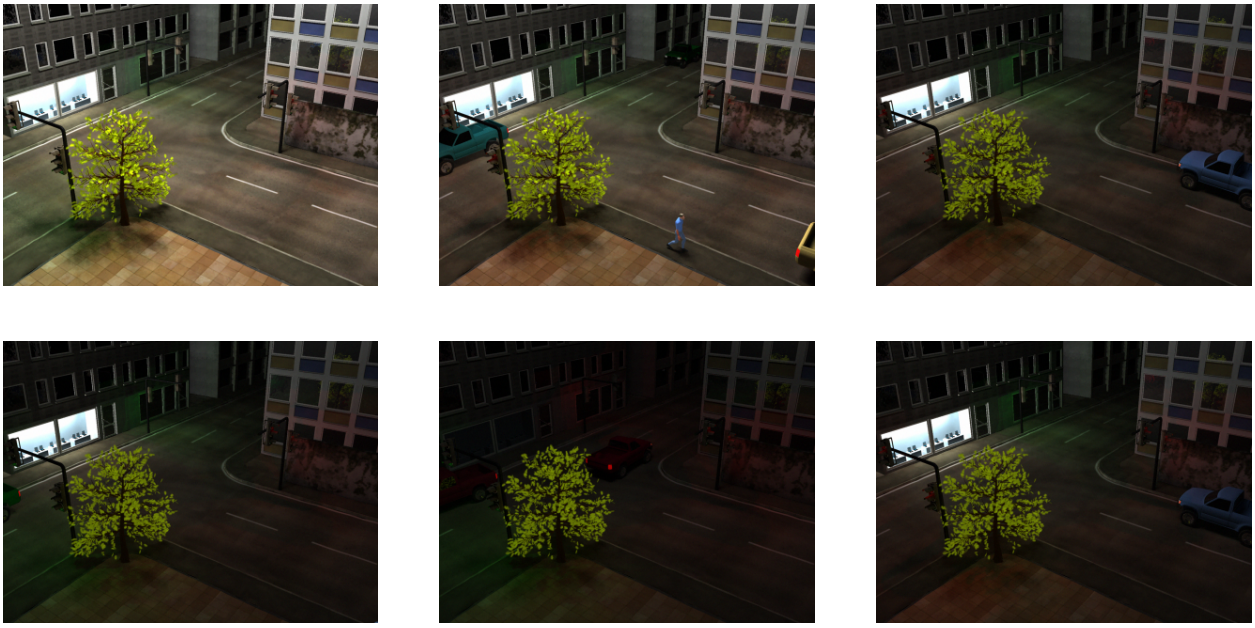


Figure 8: The SABS dataset used for evaluating the models. The first row depicts the training sequence *Darkening*, while the second row shows the testing video *LightSwitch*. The columns show frames from the start, middle and ending parts of the video. Note that in the middle of the *LightSwitch* sequence the store light switches off, causing major changes to the background.

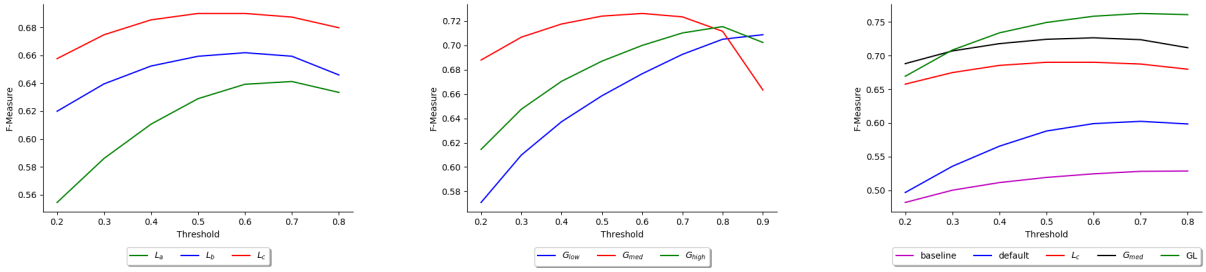


Figure 9: F-Measure values on different thresholds for each model.

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 6: Ablation studies for local and global changes

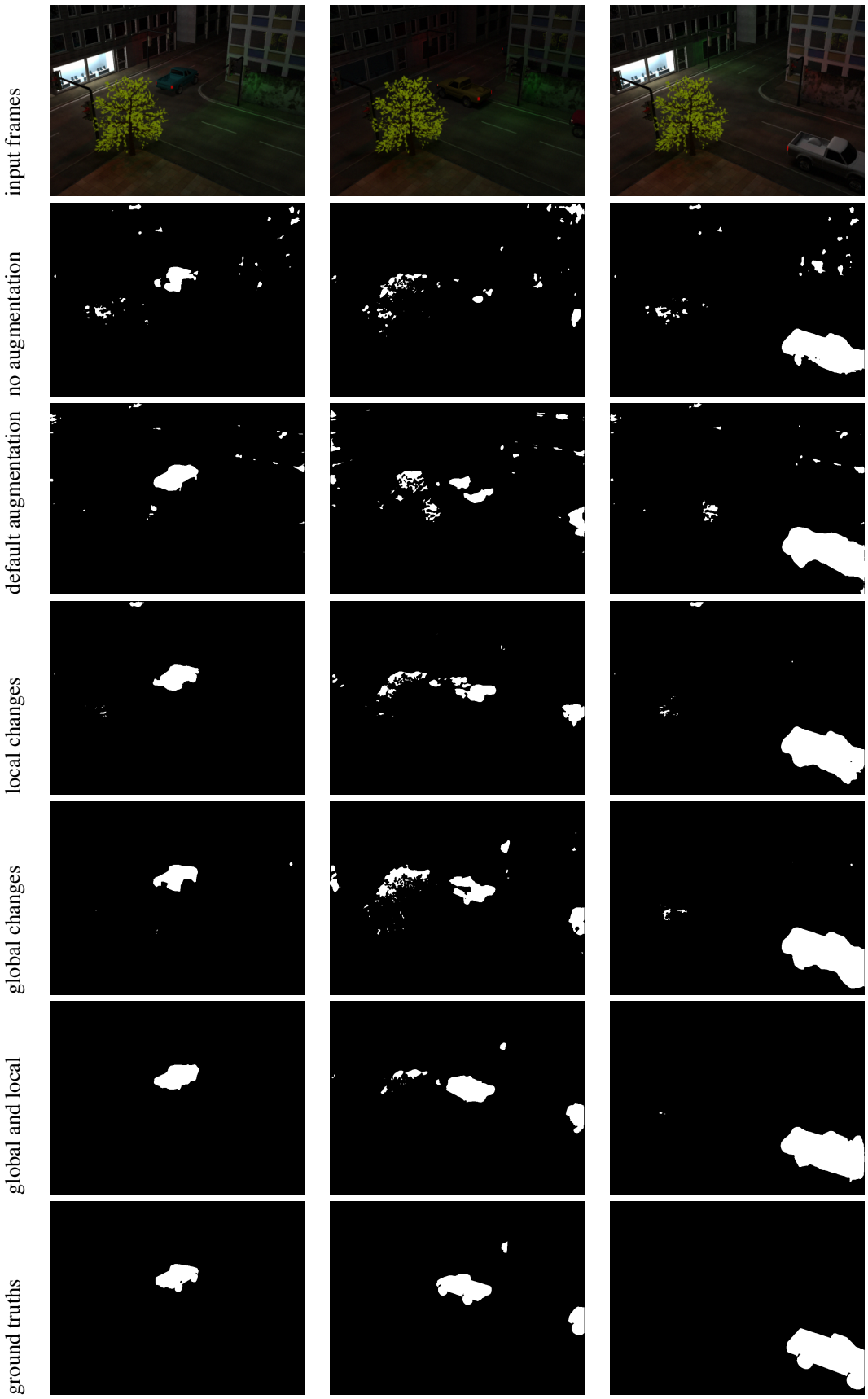


Figure 10: Comparison between different augmentation techniques.

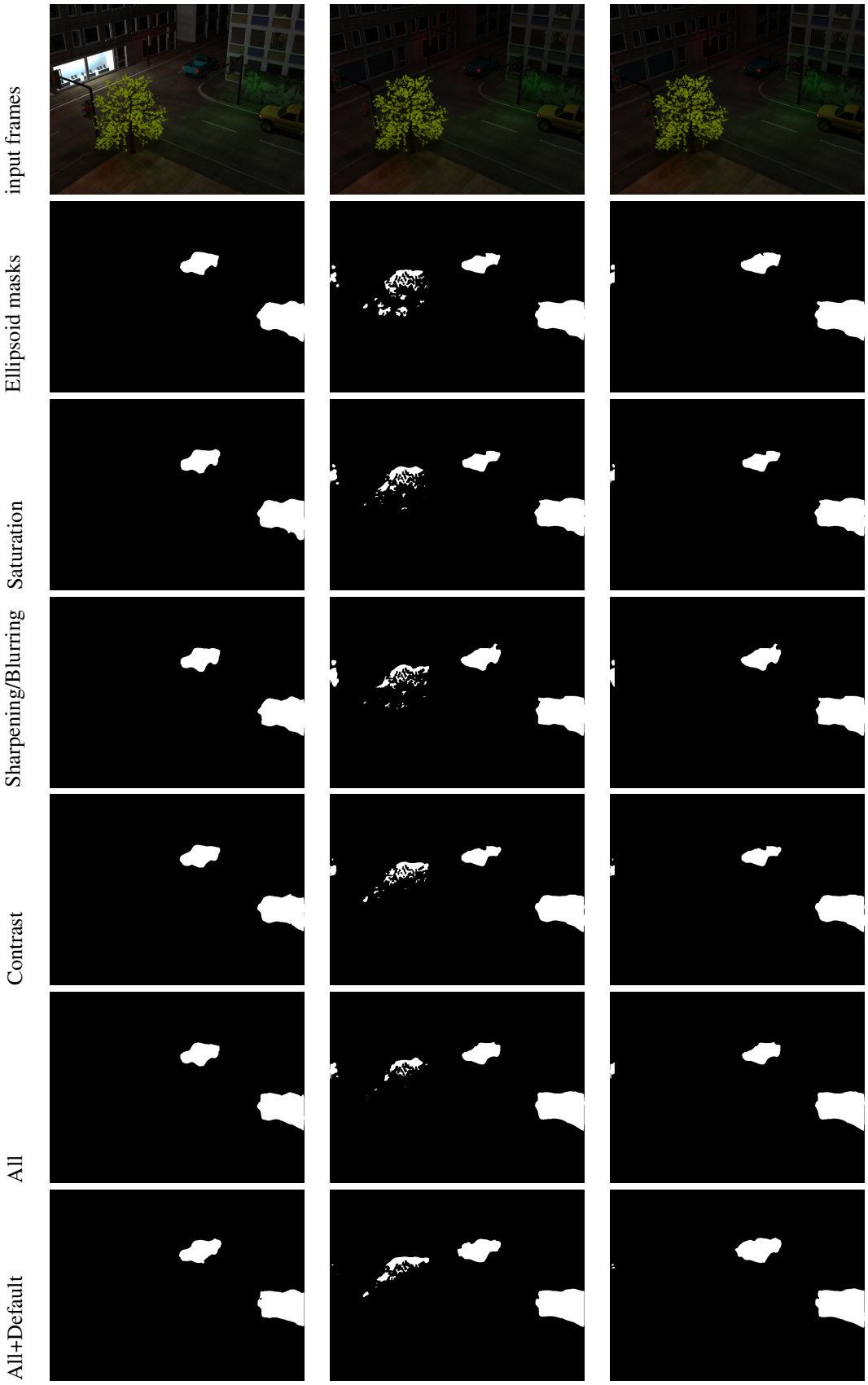


Figure 11: Comparison between the proposed augmentation techniques.

There are several interesting future directions for this research.

Firstly, the geometric information of the objects in the scenes directly affects the illuminations. While we currently model the effects by adjusting the intensity of the pixels, the results can potentially be improved if we explicitly model the geometry and deduce the occlusion information between objects. This will allow us to generate more realistic lighting and shadowing effects.

Secondly, the location of applying the local effects is randomly selected in our current design. This has the advantage of improving the robustness of the system. However, this also has the disadvantage of creating scenes that may be semantically incorrect. We would like to learn the correlations between illumination effects and the locations to apply them from real-world images.

Finally, under the current design, the data augmentation is used as a pre-processing method for creating more training data. On one hand, this is advantageous as the computation cost of training the deep network becomes independent of that of the data augmentation process. Therefore, the augmented data can be pre-computed and reused repeatedly. However, it is likely that the generated training data might not be the most effective samples for training the network and maximising its generalisation capabilities. From this point of view, coupling image generation with training and optimising both tasks makes sense. Hence, in the future, we would like to explore methods like adversarial training and measure their impact on data augmentation.

Acknowledgements

The project was supported in part by the Royal Society (Ref: IES\R1\191147 and IES\R2\181024) and Defence and Security Accelerator (Ref: DSTLX-1000140725).

References

[1] Olivier Barnich and Marc Van Droogenbroeck. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):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*, 117(2):130–144, 2013.

[3] Y. Shen, W. Hu, M. Yang, J. Liu, B. Wei, S. Lucey, and C.T. Chou. Real-time and robust compressive background subtraction for embedded camera networks. *IEEE TRANSACTIONS ON MOBILE COMPUTING*, 15(2):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.

[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)*, pages 58–62, July 2015.

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

[7] Edmond S. L. Ho, Jacky C. P. Chan, Donald C. K. Chan, Hubert P. H. Shum, Yiu ming Cheung, and Pong C. Yuen. Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments. *Computer Vision and Image Understanding*, pages –, 2016.

[8] 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*, 39(5):981–994, 2017.

[9] 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*, 35(3), 2016.

[10] K. D. McCay, E. S. L. Ho, C. Marcroft, and N. D. Embleton. Establishing pose based features using histograms for the detection of abnormal infant movements. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5469–5472, July 2019.

[11] K. D. McCay, E. S. L. Ho, H. P. H. Shum, G. Fehringer, C. Marcroft, and N. D. Embleton. Abnormal infant movements classification with deep learning on pose-based features. *IEEE Access*, 2020.

[12] D. Sakkos, E. S. L. Ho, and H. P. H. Shum. Illumination-based data augmentation for robust background subtraction. In *International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 2019.

[13] Sebastian Brutzer, Benjamin Hoferlin, and Gunther Heidemann. Evaluation of background subtraction techniques for video surveillance. *CVPR*, 2011.

[14] Parthipan Siva, Mohammad Javad Shafiee, Francis Li, and Alexander Wong. Pirm: Fast background subtraction under sudden, local illumination changes via probabilistic illumination range modelling. *2015 IEEE International Conference on Image Processing (ICIP)*, 2015.

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

[16] Aissa Boulmerka and Mohand Said Allili. Foreground segmentation in videos combining general gaussian mixture modeling and spatial information. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1330–1345, 2018.

[17] Mingliang Chen, Xing Wei, Qingxiong Yang, Qing Li, Gang Wang, and Ming-Hsuan Yang. Spatiotemporal gmm for background subtraction with superpixel hierarchy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1518–1525, 2018.

[18] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):831–843, Aug. 2000.

[19] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[20] Salehe Erfanian Ebadi and Ebroul Izquierdo. Foreground segmentation with tree-structured sparse rpca. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2273–2280, 2018.

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

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[24] Chen L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Symmetry-driven accumulation of local features for human characterization and re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834 – 848, 2018.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, July 2017.

[26] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 636–644, July 2017.

[27] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison W. Cottrell. Understanding convolution for semantic segmentation. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460, 2018.

[28] Zhaowei Cai, Quanfu Fan, Rogerio S. Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 354–370, Cham, 2016. Springer International Publishing.

[29] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 154–169, Cham, 2016. Springer International Publishing.

[30] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 186–201, Cham, 2016. Springer International Publishing.

[31] Long Ang Lim and Hacer Yalim Keles. Foreground segmentation using convolutional neural networks for multiscale feature encoding. *Pattern Recognition Letters*, 112:256 – 262, 2018.

[32] Dongdong Zeng and Ming Zhu. Background subtraction using multiscale fully convolutional network. *IEEE Access*, 6:16010–16021, 2018.

[33] Xiqun Lu. A multiscale spatio-temporal background model for motion detection. *2014 IEEE International Conference on Image Processing (ICIP)*, 2014.

[34] Yi Wang, Zhiming Luo, and Pierre-Marc Jodoin. Interactive deep learning method for segmenting moving objects. *Pattern Recognition Letters*, 96:66–75, 2017.

[35] Dimitrios Sakkos, Heng Liu, Jungong Han, and Ling Shao. End-to-end video background subtraction with 3d convolutional neural networks. *Multimedia Tools and Applications*, 77(17):23023–23041, Sep 2018.

[36] Daniel Berjón, Carlos Cuevas, Francisco Morán, and Narciso García. Real-time nonparametric background subtraction with tracking-based foreground update. *Pattern Recognition*, 74:156–170, 2018.

[37] 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*, 28(8):1737–1751, Aug 2018.

[38] 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.

[39] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 36(6), November 2017.

[40] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[41] Marc-Andre Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagne, and Jean-Francois Lalonde. Deep parametric indoor lighting estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[42] Dimitrios Sakkos, Edmond S. L. Ho, and Hubert P. H. Shum. Illumination-aware multi-task gans for foreground segmentation. *IEEE Access*, 7(1):10976–10986, 2019.

[43] C. A. Poynton. Gamma and its disguises: the nonlinear mappings of intensity in perception, crts, film, and video. *SMPTE*, 102:1099–1108, December 1993.

[44] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[45] Michael Goesele, Xavier Granier, Wolfgang Heidrich, and Hans-Peter Seidel. Accurate light source acquisition and rendering. In *Proc. of SIGGRAPH '03 (Special issue of ACM Transactions on Graphics)*, pages 621–630, July 2003.

[46] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICRL)*, pages 1–14, 2015.

[47] Andrew S. Glassner, editor. *An Introduction to Ray Tracing*. Academic Press Ltd., GBR, 1989.

[48] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980, Dec 2014.

[49] François Chollet et al. Keras. `https://keras.io`, 2015.

[50] Pavel Yakubovskiy. Segmentation models. `https://github.com/qubvel/segmentation_models`, 2019.