Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Illumination-aware Multi-task GANs for Foreground Segmentation

DIMITRIOS SAKKOS[1], EDMOND S. L. HO [2], AND HUBERT P. H. SHUM.[3], (Member, IEEE)
[1]Northumbria University, NE1 8ST UK (e-mail: dksakkos@gmail.com)
[2]Northumbria University, NE1 8ST UK (e-mail: e.ho@northumbria.ac.uk)
[3]Northumbria University, NE1 8ST UK (e-mail: hubert.shum@northumbria.ac.uk)

Corresponding author: Edmond S. L. Ho

ABSTRACT Foreground-background segmentation has been an active research area over the years. However, conventional models fail to produce accurate results when challenged with videos of challenging illumination conditions. In this paper, we present a robust model that allows accurately extracting the foreground even in exceptionally dark or bright scenes, as well as continuously varying illumination in a video sequence. This is accomplished by a triple multi-task generative adversarial network (TMT-GAN) that effectively models the semantic relationship between dark and bright images, and performs binary segmentation end-to-end. Our contribution is two-fold: First, we show that by jointly optimising the GAN loss and the segmentation loss, our network simultaneously learns both tasks that mutually benefit each other. Secondly, fusing features of images with varying illumination into the segmentation branch vastly improves the performance of the network. Comparative evaluations on highly challenging real and synthetic benchmark datasets (ESI, SABS) demonstrate the robustness of TMT-GAN and its superiority over state-of-the-art approaches.

INDEX TERMS Background Subtraction, Multi-task Learning, Generative Adversarial Networks, Video Segmentation, Illumination-aware

I. INTRODUCTION

BACKGROUND subtraction (BGS) aims at segmenting the foreground objects from its surroundings in a given image. Unlike object detection, the task lies on a pixel-wise level, and therefore being inherently more challenging. It is commonly considered as the first step of many real-world applications, such as person re-identification [1], object tracking [2], gesture recognition [3], vehicle tracking [4], crowd analysis [5] and even use cases of the medical domain [6]. Thus, the development of robust BGS methods is of paramount importance.

With the recent success of Deep Learning in image segmentation [7, 8], high accuracy in BGS can be achieved in controlled environments, which can be either videos with minimal change in the background or images with adequate illumination and high contrast. However, it is a much harder problem in real-world scenarios, in which the illumination changes in the scene may cast shadows, cause reflection and even alter the color of objects. In unexpected, but not uncommon, scenarios such as a street light being suddenly switched off at night, the effect can be dramatic and unmanageable by existing models (Figure 5 and 6).

In this paper, we tackle the aforementioned problems by proposing a Triple Multi-Task Generative Adversarial Network (TMT-GAN) for background subtraction; comprised of three separate GANs, each solving a different task. GANs are deep learning models that are comprised of two distinct networks: the generator and the discriminator. The generator is trained to produce new samples of the same distribution as the input, while the discriminator's task is to classify the generated image as a fake/real sample [9]. These networks have been proven to be very successful in not only generic image-to-image translation [10, 11, 12] but also illumination-specific image editing [13, 14, 15], thus being an excellent method for our task. A naive approach would suggest using a single GAN to normalize the illumination of the input image and then perform background subtraction in a two-step manner. However, in this case the segmentation accuracy would be very sensitive to the reconstruction abilities of the GAN and would fail completely if the generated image is even slightly inaccurate. Our proposed TMT-GAN is specifically designed to solve two problems at the same time in a multi-task, end-to-end manner: decoding the illumination

of the scene and performing background subtraction. This is accomplished by generating a pair of low/high brightness images and using GANs to reconstruct each of them with the brightness level of the other. The foreground segmentation is then performed using multi-scale features extracted from different layers of the generators. The result is a unified system for robust BGS that addresses the weaknesses of existing approaches in videos featuring drastic illumination changes. Experimental results indicate the robustness of our proposed framework on benchmark datasets with significant change in illumination that outperforms state-of-the-art approaches.

The main contributions of this paper are summarized as follows:

- We propose a novel end-to-end architecture based on a triple multi-task generative adversarial network (TMT-GAN) for background-foreground segmentation on videos with significant changes in illumination.
- We construct the supervision of the generators in a manner that increases the contrast between foreground and background and facilitates illumination-aware BGS. We jointly optimise the GAN loss and the segmentation loss to obtain optimal results.
- To the best of our knowledge, this is the first multi-task GAN with inputs of different degrees of brightness. We show that fusing features of images with varying illumination into the segmentation branch vastly improves the performance.

The rest of the paper is structured as follows: Section III explains our methodology in detail and provides technical information. In Section IV, we introduce the datasets used in this study and present the experimental results. A summary and future work are discussed in Section V.

II. RELATED WORK

A. GAUSSIAN MIXTURE MODELS

Early approaches used Gaussian Mixture Models (GMMs) for BGS, in an attempt to represent the data distribution as a mixture of gaussians [16]. Recently, there have been a plethora of papers inspired by GMMs. Siva et al. [17] extend the work of Zivkovic et al. [18] and combine a GMM with a conditional probabilistic function which attempts to model the pixel intensity values affected by sudden local illumination change. Boulmerka and Allili [19] combine a GMM with inter-frame correlation analysis and histogram matching. Akilan et al. [20] enhance the results of a GMM model by fusing features of color similarity, color distortion, and illumination measures. Chen et al. [21] 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. [2] 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. Although GMM-based methods perform well on videos with minimal or gradual illumination changes, they fail when challenged with rapid variations of illumination [22].

B. PRINCIPAL COMPONENT ANALYSIS

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 [23]. Vosters et al. [24] extend this work by introducing a statistical illumination model similar to Pillet et al. [25]. Specifically, they introduce a spatial likelihood model for modelling the relationship of neighboring pixels which is updated as new frames are analysed. Candès et al. [26] 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 [27] 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. Xin et al. [28] also extended RPCA by utilising contextual information of the foreground pixels with the generalized fused lasso regularization which was originally proposed in [29]. Although they report the SABS dataset, only the *basic* video sequence was used, which has negligible changes in illumination. The pixels of the shadow were incorrectly classified as foreground, as the difference in the illumination makes them darker. While PCA-based methods are more robust to illumination changes than GMMs, they are limited by the lack of semantic knowledge in the scene.

C. DEEP LEARNING

Deep learning approaches use variants of the fully convolutional network (FCN) proposed by Long et al. [30]. 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 BGS methods follow the trend of recent generic image segmentation networks [7, 8, 31, 32, 33] and treat videos as a collection of images while disregarding the temporal information. Those approaches focus on improving the foreground objects boundaries. Following the success of earlier approaches in object detection [34], image dehazing [35], segmentation [36] etc., Lim and Keles [37] and Zeng and Zhu [38] attempt to improve their binary maps by employing multi-scale feature aggregation. While [38] realise this idea simply by concatenating features from different layers, [37] employ multi-scale inputs. Wang et al. [39] also adopt the same preprocessing, but they refine the original CNN output by feeding it into another CNN.

D. SPATIO-TEMPORAL MODELS

On the other hand, there exist methods which attempt to model the temporal information. Sakkos et al. [40] employ 3D convolutions on a cube of 10 consecutive frames to exploit the relationship between them. Berjón et al. [41] use information from previous frames in order to update the background model of their Kernel Density Estimation-based system. Liu et al [42] proposed a method based on sparse signal recovery which exploited group property information in both spatial and temporal domains. Javed et al. [43] improved [27] by incorporating spatio-temporal constraints and reported better performance.

III. METHODOLOGY

In this section, we introduce our proposed BGS framework and the proposed architecture, which is illustrated in Fig. 1. Given a video sequence, the proposed framework takes each individual frame as the input. Each image is first edited by increasing and decreasing the gamma value to create a pair of images with extreme illuminations (Section III-A). Then, each edited image is fed into the VGG16 [44] network for extracting the deep features. Next, our framework learns a robust representation between the paired images. Motivated by the success of double GAN for image-to-image translation between different domains [10, 11, 45], we employ it for illumination decomposition by learning the differences between exceedingly bright/dark images. This is accomplished by reconstructing an image of one domain with characteristics of the other. Furthermore, we extend these methods by appending an extra GAN for binary segmentation on the classes of foreground/background.

Overall, we use one encoder E_1 for general feature extraction and three generators G_b, G_d, G_s along with three discriminators D_b, D_d, D_s for the domain of bright images, dark images and binary segmentation respectively. Skip connections between layers are employed to all generators to aid the preservation of high-level information and edge alignment. In more detail, the output of three different layers of VGG16 is extracted and each is concatenated with that layer of the generators G_b and G_d, which is of the same resolution.

Basically, we divide our approach into three distinct parts: pre-processing, feature extraction and finally background subtraction. Each part is discussed in the following subsections.

A. PRE-PROCESSING WITH GAMMA CORRECTION

To ensure the robustness of our model against illumination changes, we create multi-scale inputs in regard to luminance. Specifically, given an image, we alter its brightness using gamma correction [46]. According to this approach, the intensity value of every pixel *p* is first normalised to the range [0,1] and then raised to the power of γ:

$$p_i^{\text{out}} = (p_i^{\text{in}}/255)^{\gamma} \quad i \in 1, \ldots, N, \tag{1}$$

where p^{out} and p^{in} are the pixels of the output and input image respectively and N is the total number of pixels. By setting $\gamma < 1$, the image becomes darker. Conversely, for $\gamma > 1$, the brightness is increased. As the γ value diverges from 1, the phenomenon becomes more extreme. Therefore, it is possible to generate an exceptionally bright/dark pair $\{I_b, I_d\}$ for each image regardless of its original brightness. The optimal value of γ is calculated adaptively and according to the average pixel intensity of the input image:

- $\gamma_b = 2.5$, $\gamma_d = 0.7$, if $\hat{p} \in \{0, \ldots, 95\}$
- $\gamma_b = 2.0$, $\gamma_d = 0.5$, if $\hat{p} \in \{96, \ldots, 120\}$
- $\gamma_b = 1.4$, $\gamma_d = 0.3$, if $\hat{p} \in \{121, \ldots, 150\}$
- $\gamma_b = 1.0$, $\gamma_d = 0.1$, if $\hat{p} \in \{151, \ldots, 255\}$

where $\hat{p}$ denotes the mean intensity values of the input image's pixels and γ_b, γ_d are the values of γ applied to the original input image for generating the input pair $\{I_b, I_d\}$.

B. FEATURE EXTRACTION

1) Transfer learning

As reported in recent work such as [37, 38], using a pre-trained CNN leads to an increased accuracy since it helps the model to converge to a better local minimum. In the proposed framework, we use the pre-trained VGG16 [44] as the encoder. Because it is only used for feature extraction, we keep the first four blocks and discard the rest as in previous work [37].

2) Pipeline

Once the input pair is obtained, each image is individually fed to VGG16 for general feature extraction. Consequently, each set of features forms the input of the corresponding generator, which is of an identical structure. The ultimate goal of the generators is to learn the illumination of the image by altering its brightness, not unlike transforming an image from day to night [12] and vice versa. At the same time, the generators need to focus on the foreground and separate it from the background. This can be divided into two objectives:

- Brightness alteration to the extremes
- Foreground object identification

We can simultaneously optimise both tasks with a single loss function by constructing the supervision as follows: the illumination decomposition is supervised by altering the intensity of the pixel values as described in section III-A. We can then train the network to detect the foreground objects by creating a large contrast between the foreground and the background pixels. To this end, the foreground pixels are assigned the value of $f_p = 0$ for dark supervision images. Conversely, for bright images, we set $f_p = 255$. The process is supervised by the corresponding discriminators, which ensure that the distribution of the generated images matches that of the input images.

C. BACKGROUND SUBTRACTION

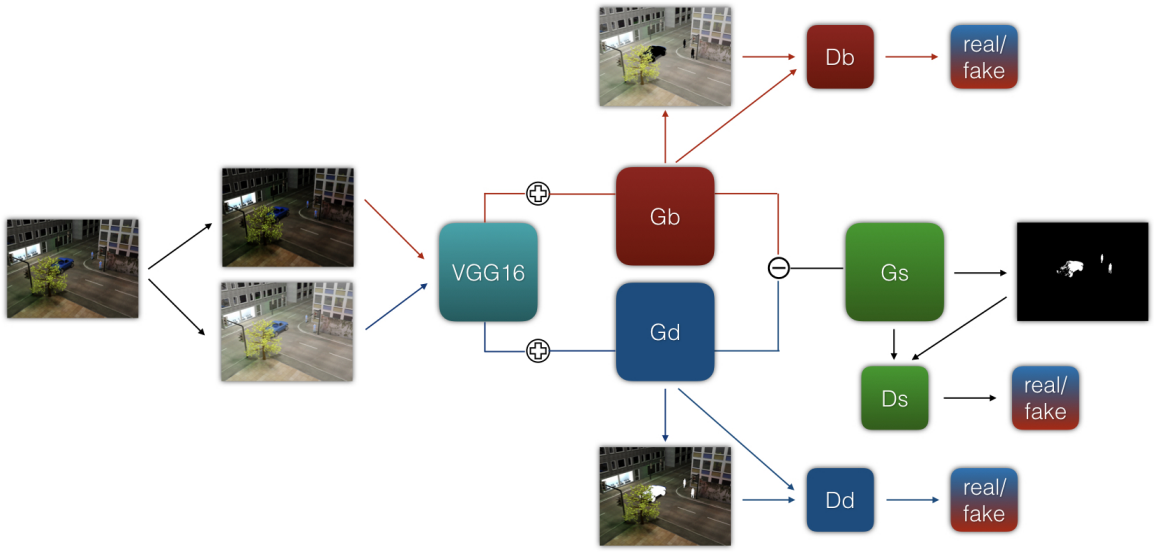


FIGURE 1: The architecture of TMT-GAN. Given an image, a low/high brightness image pair is generated with the gamma function. Both images individually undergo feature extraction by $VGG16$, followed by image generation by G_b and G_d. Red arrows show the path for dark images and blue arrows for bright images. Multi-scale features are extracted by G_b and G_d and incorporated into the foreground segmentator G_s with pixel-wise subtraction. D_b, D_d and D_s are the discriminators of the respective generators and ensure the distribution matching between the real and generated images. The plus and minus signs indicate feature concatenation and element-wise subtraction respectively.

1) Multi-scale feature fusion

Essentially, the deep features of G_b and G_d encode both illumination and saliency information. Moreover, they project the foreground object to two opposing extremities of the RGB spectrum. Therefore, the selection of the appropriate fusion mechanism now becomes apparent: subtraction. More specifically, we extract features of different resolution from three layers of G_b and G_d, namely, G_{bi} and G_{di} respectively, where $i \in \{2, 4, 8\}$ denotes the downsampling ratio. To obtain the final features, we perform element-wise subtraction and scaling by applying the hyperbolic tangent function: $G_{si} = \tanh(G_{di} - G_{bi})$. Finally, the foreground segmentation generator G_s accepts the features G_{si} as input and provides the final segmentation mask.

2) Attention

Attention-based CNNs have gathered intense interest among researchers recently [47, 48]. Intuitively, the attention mechanism is used for teaching the model to focus on specific parts of the input. Due to this feature, such a module is directly relevant to background subtraction, as it can potentially assist the model to focus on the foreground.

In the task of image segmentation, visual attention can be categorised into two parts: soft attention and hard attention. While the implementation varies wildly, generally hard attention samples one region of the image at a time and is not differentiable; on the other hand, soft attention is, as it creates a probability map which is used to assign different weights to each pixel according to its significance on the task [49]. In addition, attention can be both supervised [50] and unsupervised [51].

We employ a self-supervised, soft attention mechanism. Inspired by Li et al. [52], we divide soft attention into spatial and channel attention, each of which is modelled with a separate stream of a similar structure. In particular, each stream consists of an average pooling layer and two convolutional layers. The average pooling layer is used to compress information across the feature maps, thereby generating a single-channel map with the most consistent activations. The first convolutional layer is adding the attention, while the second one is used for scaling. The two attention streams are fused with tensor multiplication.

Three attention modules are embedded into the segmentation generator G_s. The modules operate in different resolution and accept as input the subtracted features of the corresponding layers of G_b and G_d, after being convolved with a 5x5 filter and concatenated with the output of the previous layer. Therefore, they utilise information from different

sources and resolutions to provide the refined features.

3) Foreground Segmentation Discriminator

To further increase the segmentation accuracy of the model, we append a discriminator, D_s, to the output of G_s. Basically, D_s discriminates between the generated mask of G_s and ground truth. In most cases, the foreground mask consists of a small number of objects of similarly defined boundaries and is easily discernible to noise. Therefore, D_s can lead G_s to generate higher quality segmentation masks in two ways: firstly, by reducing false-positive noise in the background and secondly, by ensuring that the foreground blobs are smooth and consistent without false-negative areas in their interior.

D. TRAINING

To optimise the task of domain translation, we use a loss that combines the optimisation of the generators for image reconstruction and the discriminators for ensuring the generated image is as natural as possible:

$$\begin{aligned} L_t = & D_d(G_d(x_d), t_d) + \alpha \left\| G_d - t_d \right\| \\ & + D_b(G_b(x_b), t_b) + \alpha \left\| G_b - t_b \right\|, \end{aligned} \tag{2}$$

where x_b, x_d and t_b, t_d are the input image and the supervision of bright and dark images respectively, and α is a hyper-parameter. In our experiments, we set $\alpha = 20$.

In the task of foreground segmentation, the classes of foreground and background are usually heavily imbalanced. To address this issue, we use the weighted cross-entropy loss, which is formally defined as follows:

$$G_s = wt[-\log \sigma(x)] + (1-t)[-\log(1-\sigma(x))], \tag{3}$$

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. To punish false negatives in the loss function and balance the two classes, we set $w = 5$.

IV. RESULTS

To evaluate the effectiveness of our proposed method, we compare our method with the following state-of-the-art approaches:

- OSVOS [53], the creators of the Davis dataset [54] for video segmentation,
- FgSegNet [37], the best performer on the benchmark dataset CDnet2014 [55], and
- CascadeCNN [39], the third best performer on CDnet2014 and the second best performer on CDnet2014 with source code open to the public

on two challenging datasets with a strong focus on intense illumination changes to demonstrate the robustness of our method. In particular, the Stuttgart Artificial Background Subtraction dataset (SABS) [56] and ESI [24] datasets are used.

To ensure a fair comparison between all models, we use the same training hyper-parameters. In more detail, we set *batch size* $= 1$ and *training epochs* $= 15$. When obtaining the binary segmentation map, we select the threshold that maximises the F-Measure. We also use pre-trained models to initialise the model parameters when applicable. Finally, all models are trained with the same training/testing split.

A. EVALUATION METRICS

We evaluate the models using a very wide variety of metrics which are commonly used in the task of BGS [55]: *Recall, Specificity, Precision, False Positive Rate, False Negative Rate, Percentage of Wrong Classifications, F-Measure* and *Intersection over Union*. The scientific formulation of these metrics is given below:

- $Recall = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$
- $Precision = \frac{TP}{TP+FP}$
- $FPR = \frac{FP}{FP+TN}$
- $FNR = \frac{FN}{TP+FN}$
- $PWC = \frac{FN+FP}{TP+FN+FP+TN} \times 100$
- $FM = \frac{2 \times Precision \times Recall}{Precision+Recall}$
- $IoU = \frac{TP}{TP+FP+FN}$,

where *TP, TN, FP, FN* are the true positive, true negative, false positive and false negative predicted pixels, respectively.

B. IMPLEMENTATION DETAILS

In the experiments, our proposed framework is implemented in Tensorflow with a single NVIDIA GeForce GTX 1060 GPU. Training was completed in approximately 6 hours. In terms of data pre-processing, all images are resized to 240x320 and normalised to $[-1, 1]$. Shuffling is also applied, however, there is no data augmentation with image crop/mirroring.

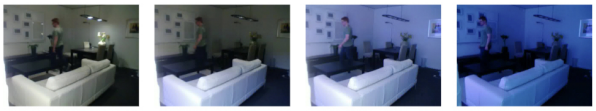
As mentioned before, we employ transfer learning by using the first four blocks of the pre-trained network VGG16 [44]. Computational efficiency is achieved with minimal loss of accuracy by freezing the first two blocks and only training the rest. Dropout is employed at the last block with keep probability $p_k = 0.5$.

C. SABS DATASET

The SABS dataset [56] contains 9 synthetic video sequences. Although the foreground movements are the same in every sequence, the illumination is changing over time. In addition, different videos have very different lighting conditions, such as day-time and night scenes. In particular, 3 (i.e. *Darkening*, *No Foreground Night* and *Light Switch*) out of the 9 video sequences are used in this experiment. To clearly demonstrate the robustness of our method, the most challenging video: *Light Switch* is used in the comparison as the testing data. An example of 3 consecutive frames is illustrated in Figure 2b. Explicit details of the training/testing data split are stated in

Scene	Frame indices
Training	
Darkening	1-800 (whole video)
No Foreground Night	1-801 (whole video)
Testing	
Light Switch	1-600 (whole video)

TABLE 1: Scenes and frame indices used in training and testing on the SABS dataset



(a) *Walking* video of ESI

(b) *Light Switch* video of SABS

FIGURE 2: Consecutive frames in the SABS and ESI datasets featuring sudden illumination changes

Table 1. For the training data, the *Darkening* scene is selected because it is a night scene which is similar to the testing video sequence and *No Foreground Night* is chosen to keep the balance between background and foreground examples in the training set. The rest of the video sequence in the SABS dataset [56] are not used because they are either day-time scenes and/or do not have significant illumination changes over time. The results are presented in Table 2.

The F-Measure indicates the average accuracy of the BGS. While the OSVOS [53], FgSegNet [37] and CascadeCNN [39] are having similar performance, our proposed method significantly outperforms the existing methods and achieved a much higher F-Measure value. This highlights the effectiveness of our method.

To evaluate the performance qualitatively, some examples of the BGS results are illustrated in Figure 5. Three different scenarios are presented in Figure 5, namely *normal* (row 1 and 4), *occlusion* (row 2) and *light off* (row 3). *Normal* scenes are having normal illumination in which the scene is bright in general and BGS can be done more easily. In *occlusion* scenes, some foreground objects are occluded by static objects which make the segmentation task more difficult. *Light off* scenes are those with the lights being switched off and results in significant illumination change over consecutive frames, which amounts to a very challenging situation.

From the results, it is demonstrated that the foreground masks (coloured in white) obtained using our method (Figure 5, rightmost column) are less noisy than those obtained using OSVOS [53], FgSegNet [37] and CascadeCNN [39]. Also, our results are closest to the ground truth in this test, which aligns well with the quantitative results presented in Table 2.

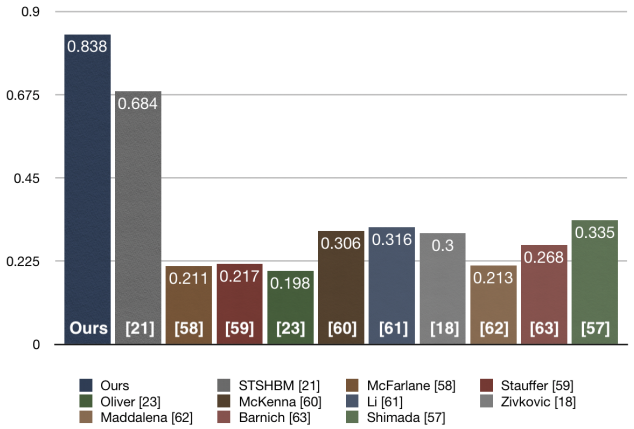


FIGURE 3: Comparison of F-Measure values with state-of-the-art models on the *Light Switch* sequence of SABS. Statistics are taken from Shimada and Taniguchi [57].

In particular, our method significantly outperformed others in the more challenging (i.e. *occlusion* and *light off*) scenes (Figure 5 2nd and 3rd rows). Even in the normal scenario, the superiority of our method is evidenced by the well-defined boundaries, such as the light post being correctly classified as background in the first row and the shape of the wheels in the last row of Figure 5. On the other hand, OSVOS [53] had trouble adjusting to dynamic environments, as it incorrectly classified many pixels of the tree as foreground. For FgSegNet [37] and CascadeCNN [39] similar results were obtained, as they failed to accurately segment both of the cars, including the non-occluded car. In the *normal* scenes (1st and 4th rows of Figure 5), all methods performed well, and CascadeCNN [39] achieved comparable performance with our method. However, CascadeCNN [39] tends to create masks with blurry edges/boundaries as depicted by the shape of the wheels in Figure 5.

In addition to the state-of-the-art approaches, we also compared our method with other existing methods and the results (F-measure) are illustrated in Figure 3 and 4. Again, our method outperformed the other methods. Note that the statistics are obtained from [57] and [43], and all the results are showing the performance (F-measure) on the same testing video sequence *Light Switch*.

D. ESI DATASET

We further evaluate our method by the ESI [24] dataset which contains 8 video sequences filmed indoor, 3 of which being background-only. Throughout all the videos, various sources of light are being switched on and off which causes drastic changes to the illumination of the room. Examples of these changes are shown in Figure 2a. The data used for training and testing the models is listed in Table 3. The split between training and testing sets is specifically performed in a way that it separates the two without allowing the models to take a glimpse into the future. Instead of manually selecting repre-

Model	Recall	Specificity	FPR	FNR	PWC	FM	Precision	IoU
OSVOS [53]	0.60402	0.99141	0.00859	0.39598	1.76350	0.61536	0.62714	0.44442
FgSegNet [37]	0.56973	0.99675	0.00325	0.43027	1.32193	0.66812	0.80758	0.50163
CascadeCNN [39]	0.56743	0.99515	0.00485	0.43257	1.48428	0.64102	0.73654	0.47169
TMT-GAN (ours)	0.87491	0.99488	0.00512	0.12509	0.79216	0.83783	0.80376	0.72091

TABLE 2: Results on the SABS (Light Switch) dataset

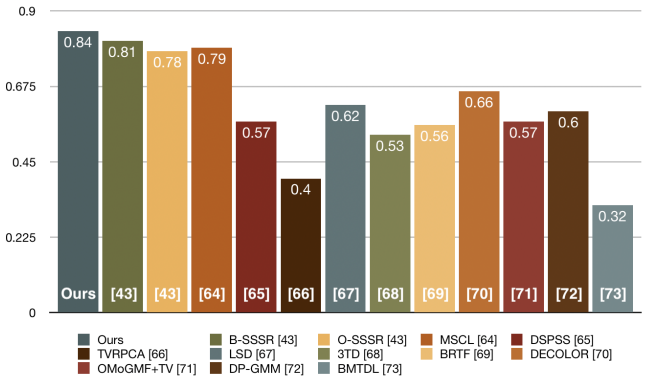


FIGURE 4: Comparison of F-Measure values with state-of-the-art models on the *Light Switch* sequence of SABS. Statistics are taken from Javed et al. [43].

Scene	Frame indices
Training	
Background scene 1	1-1375
Background scene 2	1-1093
house	1-401
walking	24-757
scene1	599-1238
scene2	1768-1836
Testing	
chair	90-663
scene1	489-589
scene2	1846-1921
sofa	34-418

TABLE 3: Scenes and frame indices used in training and testing on the ESI dataset

sentative frames from each video sequence as in [37, 39], we either select different videos for training and testing or split a video into two continuous parts, depending on whether they share the same background. Therefore, *scene1* and *scene2* are divided into two parts consisted of consecutive frames, the first being reserved for training and the latter for testing, since some foreground of these scenes has to be included in the training data. Nevertheless, the testing sequences are unseen data as indicated by the frame indices for *scene1* and *scene2* in Table 3 and the movement of the person (foreground) is completely different between the training and testing scenes. On the other hand, *chair* and *sofa* can be used for testing in their entirety, since they share the same background with *walking*. The results are presented in Table 4 and 5.

Again, the F-Measure indicates our method significantly outperforms the OSVOS [53], FgSegNet [37] and CascadeCNN [39]. This highlights the consistency and robustness of our method.

The qualitative results are illustrated in Figure 6. To clearly show the way each method handles sudden illumination changes, we provide the segmentation maps on consecutive frames of each video sequence of the testing set where the illumination of the scene changes drastically. Among the 4 videos, *Scene2* features the slightest change in illumination, which explains why all methods are performed well (Row 3-4 in Figure 6).

It can be seen that state-of-the-art models have considerable noise in low-light frames. FgSegNet [37] has very few false positives. However, it comes with a cost of a large number of false negatives. OSVOS [53] and CascadeCNN [39] on the other hand, provide better person silhouettes but also have many false positives. All in all, our method (Figure 6 rightmost column) achieves accurate segmentation maps even in images of low brightness with very minimal false positives and negatives, as seen in comparison to the ground truth.

E. ABLATION STUDIES

In this subsection, we justify the decisions we made in the proposed framework by conducting a series of ablation tests. In particular, we evaluate the performance of the proposed model by testing the effect on removing individual components on foreground-background segmentation tasks. The results are shown in Table 6.

From the results, it can be seen that all modules improve the performance in both datasets. First of all, it is shown that removing the weighted loss function and training with regular cross-entropy significantly affects the performance in the SABS dataset, but has a lower impact on the ESI dataset. This is because there is a higher imbalance on the foreground/background classes in the SABS dataset, as the foreground objects are -mostly- smaller in size. Similarly, the attention module has a larger contribution on SABS than ESI because it is generally a more challenging dataset with smaller foreground objects, thus the effect is more profound. Finally, adding a discriminator on the G_s module also has a beneficial effect, as it forces G_s to create better quality masks.

V. CONCLUSION

In this paper, we propose a robust background subtraction method based on adversarial learning and feature fusion. Extensive experimental results demonstrate the superiority of the proposed method against state-of-the-art approaches in both normal and challenging scenarios and show its robustness in handling sudden illumination changes. As future

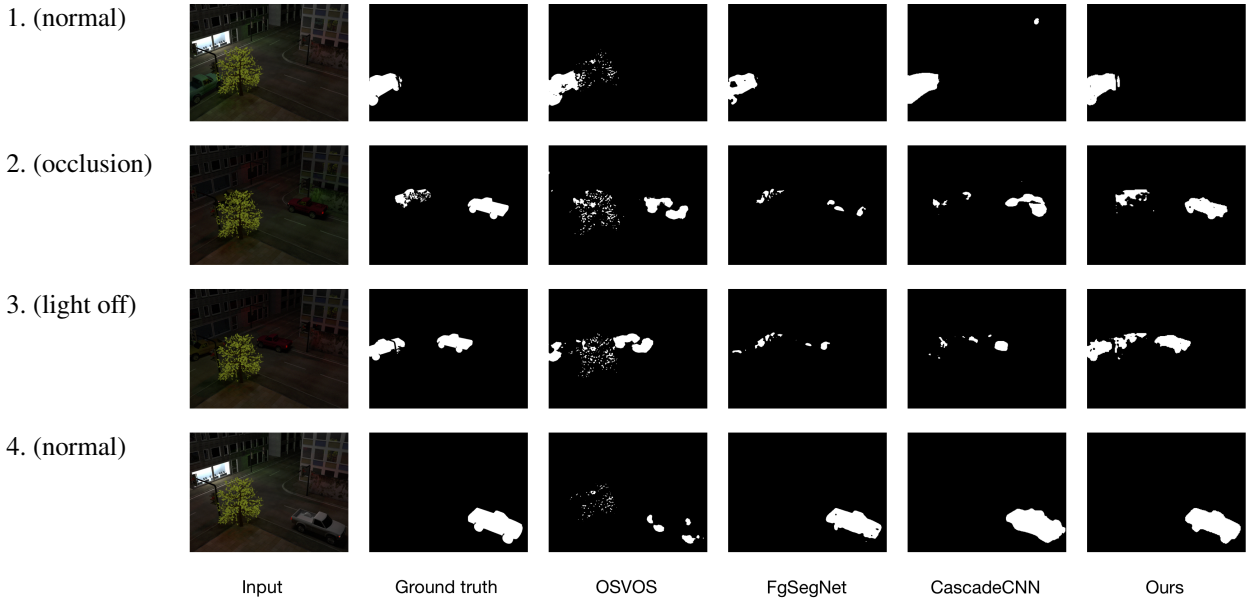


FIGURE 5: Qualitative results on the SABS dataset

Method	**Category**							
	chair		*scene1*		*scene2*		*sofa*	
	F-measure	PWC	F-measure	PWC	F-measure	PWC	F-measure	PWC
OSVOS [53]	0.74020	1.48911	0.87111	3.03099	0.90901	1.83846	0.63719	4.72149
FgSegNet [37]	0.80933	0.97998	0.90373	2.05771	0.93176	1.41115	0.70716	3.93419
CascadeCNN [39]	0.73602	1.44233	0.76520	4.83995	0.89543	2.24383	0.60473	5.58378
TMT-GAN (ours)	0.83427	0.84147	0.90956	1.94898	0.93394	1.34220	0.76900	3.33227

TABLE 4: Results on the ESI dataset by category

Model	**Recall**	**Specificity**	**FPR**	**FNR**	**PWC**	**FM**	**Precision**	**IoU**
OSVOS [53]	0.77036	0.98878	0.01122	0.22964	2.32949	0.78519	0.80060	0.64635
FgSegNet [37]	0.81815	0.99209	0.00791	0.18185	1.75072	0.83763	0.85806	0.72062
CascadeCNN [39]	0.75095	0.98382	0.01618	0.24905	2.90498	0.74075	0.73083	0.58825
TMT-GAN (ours)	0.90152	0.99235	0.00765	0.09848	1.26758	0.88723	0.87338	0.79731

TABLE 5: Results on the ESI dataset

	SABS (Light Switch)			ESI		
Removed Component	**F-Measure**	**Recall**	**Precision**	**F-Measure**	**Recall**	**Precision**
Attention	0.80048	0.79081	0.81039	0.88311	0.90675	0.86067
D_f	0.82734	0.78464	0.87496	0.85124	0.87532	0.82845
Weighted loss	0.72439	0.62910	0.85369	0.83729	0.81345	0.86257
All	**0.83783**	0.87291	0.80376	**0.88723**	0.90152	0.87338

TABLE 6: F-Measure values of the model if a component is removed

work, we intend to incorporate semantic and temporal information to the model.

ACKNOWLEDGEMENT

This project was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) (Ref: EP/M002632/1) and the Royal Society (Ref: IES\R2\181024).

REFERENCES

[1] L. Bazzani, M. Cristani, V. Murino, Symmetry-driven accumulation of local features for human characterization and re-identification, Computer Vision and Image Understanding 117 (2013) 130–144.

[2] Y. Shen, W. Hu, M. Yang, J. Liu, B. Wei, S. Lucey, C. Chou, Real-time and robust compressive background subtraction for embedded camera networks, IEEE TRANSACTIONS ON MOBILE COMPUTING 15 (2016) 406 – 418.

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

[4] N. Sirikuntamat, S. Satoh, T. H. Chalidabhongse, Ve-

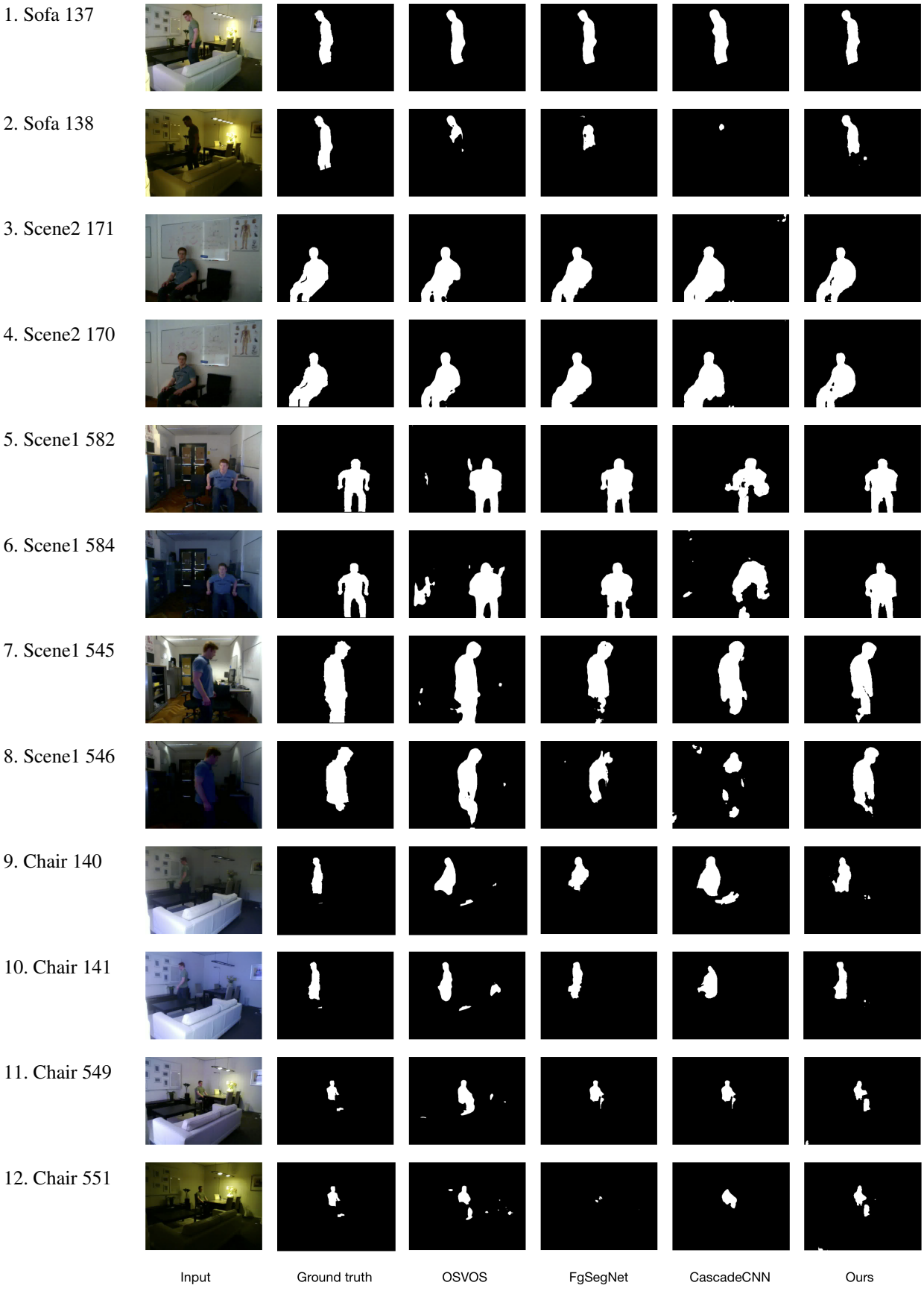


FIGURE 6: Qualitative results on the ESI dataset

hicle tracking in low hue contrast based on camshift and background subtraction, in: 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 58–62.

[5] X. Wang, C. Lu, J. Jia, H. Li, l_0 regularized stationary-time estimation for crowd analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 981–994.

[6] K. Them, M. G. Kaul, C. Jung, M. Hofmann, T. Mummert, F. Werner, , T. Knopp, Sensitivity enhancement in magnetic particle imaging by background subtraction, IEEE TRANSACTIONS ON MEDICAL IMAGING 35 (2016).

[7] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask r-cnn, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2980–2988.

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

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680.

[10] J. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251.

[11] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: Unsupervised Dual Learning for Image-to-Image Translation, in: Proceedings of the IEEE International Conference on Computer Vision, volume 2017-October, pp. 2868–2876.

[12] P. Isola, J. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976.

[13] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, L. V. Gool, Night-to-day image translation for retrieval-based localization, arXiv:1809.09767v1 (2018).

[14] X. Wang, A. Gupta, Generative image modeling using style and structure adversarial networks, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 318–335.

[15] C. Liu, X. Wu, X. Shu, Learning-based dequantization for image restoration against extremely poor illumination, arXiv:1803.01532v2 (2018).

[16] Z. Zivkovic, Improved adaptive gaussian mixture model for background subtraction, Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (2004).

[17] P. Siva, M. J. Shafiee, F. Li, 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).

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

[19] A. Boulmerka, M. S. Allili, Foreground segmentation in videos combining general gaussian mixture modeling and spatial information, IEEE Transactions on Circuits and Systems for Video Technology 28 (2018) 1330–1345.

[20] T. Akilan, Q. J. Wu, Y. Yang, Fusion-based foreground enhancement for background subtraction using multivariate multi-model gaussian distribution, Information Sciences 430-431 (2018) 414–431.

[21] M. Chen, X. Wei, Q. Yang, Q. Li, G. Wang, M.-H. Yang, Spatiotemporal gmm for background subtraction with superpixel hierarchy, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018) 1518–1525.

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

[23] B. R. N. Oliver, A. Pentland, A bayesian computer vision system for modeling human interactions, IEEE Trans. Pattern Anal. Mach. Intell. 22 (Aug. 2000) 831–843.

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

[25] J. Pilet, C. Strecha, P. Fua, Making background subtraction robust to sudden illumination changes, European Conference on Computer Vision (2008).

[26] E. J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM 58 (2011) 1–37.

[27] S. E. Ebadi, E. Izquierdo, Foreground segmentation with tree-structured sparse rpca, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018) 2273–2280.

[28] B. Xin, Y. Tian, Y. Wang, W. Gao, Background subtraction via generalized fused lasso foreground modeling, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015).

[29] B. Xin, Y. Kawahara, Y. Wang, L. Hu, W. Gao, Efficient generalized fused lasso and its applications to the diagnosis of alzheimers disease, ACM Transactions on Intelligent Systems and Technology 7 (2016) 1–22.

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

(2015).

[31] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239.

[32] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 636–644.

[33] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. W. Cottrell, Understanding convolution for semantic segmentation, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018) 1451–1460.

[34] Z. Cai, Q. Fan, R. S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 354–370.

[35] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, M.-H. Yang, Single image dehazing via multi-scale convolutional neural networks, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 154–169.

[36] A. Roy, S. Todorovic, A multi-scale cnn for affordance segmentation in rgb images, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 186–201.

[37] L. A. Lim, H. Y. Keles, Foreground segmentation using convolutional neural networks for multiscale feature encoding, Pattern Recognition Letters 112 (2018) 256 – 262.

[38] D. Zeng, M. Zhu, Background subtraction using multiscale fully convolutional network, IEEE Access 6 (2018) 16010–16021.

[39] Y. Wang, Z. Luo, P.-M. Jodoin, Interactive deep learning method for segmenting moving objects, Pattern Recognition Letters 96 (2017) 66–75.

[40] D. Sakkos, H. Liu, J. Han, L. Shao, End-to-end video background subtraction with 3d convolutional neural networks, Multimedia Tools and Applications 77 (2018) 23023–23041.

[41] D. Berjón, C. Cuevas, F. Morán, N. García, Real-time nonparametric background subtraction with tracking-based foreground update, Pattern Recognition 74 (2018) 156–170.

[42] X. Liu, J. Yao, X. Hong, X. Huang, Z. Zhou, C. Qi, G. Zhao, Background subtraction using spatio-temporal group sparsity recovery, IEEE Transactions on Circuits and Systems for Video Technology 28 (2018) 1737–1751.

[43] S. Javed, A. Mahmood, S. Al-Maadeed, T. Bouwmans, S. K. Jung, Moving object detection in complex scene using spatiotemporal structured-sparse rpca, IEEE Transactions on Image Processing 28 (2018) 1007–1022.

[44] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, International Conference on Learning Representations (ICRL) (2015) 1–14.

[45] X. Huang, M.-Y. Liu, S. J. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, CoRR abs/1804.04732 (2018).

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

[47] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 451–466.

[48] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 2048–2057.

[49] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, T. M. Hospedales, Deep spatial-semantic attention for fine-grained sketch-based image retrieval, 2017 IEEE International Conference on Computer Vision (ICCV) (2017).

[50] L. Liu, M. Utiyama, A. Finch, E. Sumita, Neural machine translation with supervised attention, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, 2016, pp. 3093–3102.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[52] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, CoRR abs/1802.08122 (2018).

[53] S. Caelles, K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. V. Gool, One-shot video object segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5320–5329.

[54] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 724–732.

[55] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, Change detection benchmark website, http://jacarini.dinf.usherbrooke.ca/results2014/, 2018. Accessed: 2018-10-30.

[56] S. Brutzer, B. Hoferlin, G. Heidemann, Evaluation of

background subtraction techniques for video surveillance, CVPR (2011).
[57] H. N. A. Shimada, R. Taniguchi, Background modeling based on bidirectional analysis, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2013) 1979–1986.
[58] N. J. B. McFarlane, C. P. Schofield, Segmentation and tracking of piglets in images, Machine Vision and Applications 8 (1995) 187–193.
[59] C. Stauffer, W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), volume 2, pp. 246–252 Vol. 2.
[60] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, H. Wechsler, Tracking groups of people, Comput. Vis. Image Understanding 80 (2000) 42–56.
[61] I. G. Li, W. Huang, Q. Tian, Foreground object detection from videos containing complex background, Proc. 11th ACM Int. Conf. Multimedia (2003) 2–10.
[62] L. Maddalena, A. Petrosino, A self-organizing approach to background subtraction for visual surveillance applications, IEEE Trans. Image Process. 17 (2008) 1168–1177.
[63] O. Barnich, M. V. Droogenbroeck, Vibe: A universal background subtraction algorithm for video sequences, IEEE Trans. Image Process. 20 (2011) 1709–1724.
[64] S. Javed, A. Mahmood, T. Bouwmans, K. Soon, Background-foreground modeling based on spatiotemporal sparse subspace clustering, IEEE T-IP (2017).
[65] S. E. Ebadi, E. Izquierdo, Foreground segmentation with tree-structured sparse rpca, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018) 2273–2280.
[66] X. Cao, L. Yang, X. Guo, Total variation regularized rpca for irregularly moving object detection under dynamic background, IEEE Transactions on Cybernetics 46 (2016) 1014–1027.
[67] X. Liu, G. Zhao, J. Yao, C. Qi, Background subtraction based on low-rank and structured sparse decomposition, IEEE T-IP 24 (2015) 2502–2514.
[68] O. Oreifej, X. Li, M. Shah, Simultaneous video stabilization and moving object detection in turbulence, IEEE T-PAMI 35 (2013) 450–462.
[69] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, S.-I. Amari, Bayesian robust tensor factorization for incomplete multiway data, IEEE T-NNLS PP (2016) 1–1.
[70] X. Zhou, C. Yang, W. Yu, Moving object detection by detecting contiguous outliers in the low-rank representation, IEEE T-PAMI 35 (2013) 597–610.
[71] H. Yong, D. Meng, W. Zuo, L. Zhang, Robust online matrix factorization for dynamic background subtraction, IEEE T-PAMI PP (2017).
[72] T. S. Haines, T. Xiang, Background subtraction with dirichletprocess mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 670–683.
[73] A. Stagliano, N. Noceti, A. Verri, F. Odone, Subsense: A universal change detection method with local adaptive sensitivity, IEEE T-IP 24 (2015) 2415–2428.

DIMITRIOS SAKKOS received the B.Sc. degree in Mathematics from the Aristotle University of Thessaloniki, Greece, in 2012 and the M.Sc. degree in Computer Science from the University of Birmingham, United Kingdom, in 2013. He is currently a PhD candidate at Northumbria University in Newcastle Upon Tyne, United Kingdom.

His research interests lie in the field of image and video segmentation.

EDMOND. S. L. HO received the B.Sc. degree in Computer Science from the Hong Kong Baptist University, in 2003, the M.Phil. degree in Computer Science from the University of Edinburgh, Scotland, in 2006 and the Ph.D. degree in Informatics from the University of Edinburgh, Scotland, in 2011. He is currently a Senior Lecturer with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom. Prior to joining Northumbria University, he was a Research Assistant Professor with the Department of Computer Science, Hong Kong Baptist University in 2011-2016.

His current research interests include computer graphics and animation, computer vision, machine learning, and robotics.

HUBERT P. H. SHUM is an Associate Professor (Reader) at Northumbria University, U.K., as well as the Director of Research and Innovation of the Computer and Information Sciences Department. Before this, he worked as a Senior Lecturer at Northumbria University, U.K., a Lecturer in the University of Worcester, U.K., a post-doctoral researcher in RIKEN, Japan, as well as a research assistant in the City University of Hong Kong. He received his Ph.D. degree from the School of Informatics in the University of Edinburgh, U.K. His research interests include computer graphics, computer vision, motion analysis and machine learning.