Background Modelling Based on Generative Unet

Ye Tao[1,2], Petar Palasek[2], Zhihao Ling[1], Ioannis Patras[2]
[1] East China University of Science and Technology, China
[2] Queen Mary University of London, UK
taoye1992@gmail.com, p.palasek@qmul.ac.uk, zhhling@ecust.edu.cn, i.patras@qmul.ac.uk

Abstract

Background Modelling is a crucial step in background/foreground detection which could be used in video analysis, such as surveillance, people counting, face detection and pose estimation. Most methods need to choose the hyper parameters manually or use ground truth background masks (GT). In this work, we present an unsupervised deep background (BG) modelling method called BM-Unet which is based on a generative architecture that given a certain frame as input it generates as output the corresponding background image - to be more precise, a probabilistic heat map of the colour values. Our method learns parameters automatically and an augmented version of it that utilises colour, intensity differences and optical flow between a reference and a target frame is robust to rapid illumination changes and camera jitter. Besides, it can be used on a new video sequence without the need of ground truth background/foreground masks for training. Experiment evaluations on challenging sequences in SBMnet data set demonstrate promising results over state-of-the-art methods.

1. Introduction

Background modelling is a stepping stone task in computer vision field. It is implemented in various video applications such as video surveillance [6], pose estimation [8] and face detection [3]. Background modelling methods can be utilised to remove the background from the video and preserve the foreground which is generally regarded important for further video analysis. Robust and accurate background modelling makes these high-level video analysis tasks more effective and efficient. However, robust background modelling in complex real scenarios is still a challenge and no approach has been able to robustly address the problem so far, even with human interaction. Most background modelling methods can use only one set of tuning parameters for all videos, but their hyper parameters are still needed to be manually fine-tuned for optimal results on different video sequences. Ideally, an adequate algorithm should be able to handle real world scene background modelling without user intervention, and also deal with challenges such as rapid illumination variations and camera jitter.

In this paper, we propose an unsupervised method for background modelling. The proposed solution has two main stages: 1) generate static backgrounds conditioned on the colour features of an input frame; 2) generate dynamic backgrounds further conditioned on the mean difference and dense optical flow [7] between two frames. The proposed method is based on a skip-pooling deep convolution neural network (DCNN) model, Unet [14]. The main reason for using DCNNs is their ability to learn features that perform better than hand-designed features at the given task, and they can also preserve the spatial relationships in the frames. Our method trains one model for one video, but it is still generic because it is user intervention free and unsupervised, which enables it to be easily adapted to new videos automatically.

The contributions of this paper are two fold:

1. At the best of our knowledge, we propose the first unsupervised deep convolutional neural network method for background modelling.
2. Our method is robust and highly adaptive to sudden illumination changes and camera jitter.

The paper is structured as follows. In Section 2, we will discuss the related works. In Section 3, we will present the proposed method in details. In Section 4, we will describe the experiments and discuss the results. Finally, Section 5 concludes the paper and presents potential future directions.

2. Related Work

Background modelling is generally the first step for the background/foreground detection task. The disparities between a reference background and the current frame usually indicate foreground objects. However, it is hard to find a good reference background from a complex real scene

978-1-5386-2939-0/17/$31.00 ©2017 IEEE

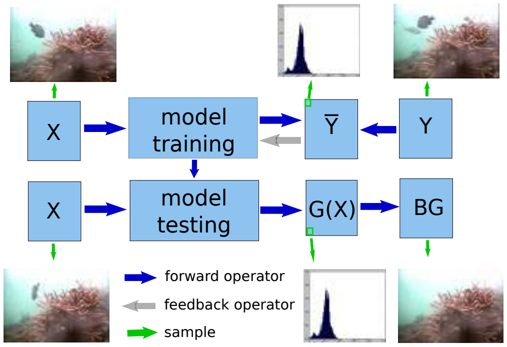

Figure 1. An illustration of our baseline method for background modelling, where X and Y are random frames from one sequence.

video. Besides, the background in real world is often dynamic.

Top performing algorithms for background modelling can be classified into four families: basic background modelling, statistical background modelling, subspace background modelling and neural network background modelling. Median based methods are typical basic methods that are simple and efficient but they fail to deal with the common condition when the background is visible for less than half of the time in the video. Two methods, LabGen [11] and LabGen-P [12] were proposed to handle this challenge. They utilise motion detection methods to select the background region or pixel and conduct the pixel-wise median filter. The drawback of these methods is that they depend too much on motion detection methods. Statistical background estimation methods assume that each pixel in the background is drawn from a Probability Distribution Function (PDF) which is modelled by a Gaussian Mixture Model (GMM) [4] or Kernel Density Estimation (KDE) [9]. They are robust to dynamic backgrounds, but they suffer from slow adaptation. Robust principal component analysis (RPCA) [2] based decompositions of matrices into low-rank and sparse components as subspace learning have also shown promising performance in background modelling, but they are sensitive to outliers. In contrast, weightless neural networks [5] and self-organizing maps [16] were also adapted to background modelling, achieving competitive performance. These methods are mostly pixel-wise and classification based and share the main drawback of losing spatial information.

Although there have been many methods based on DCNNs used in video analysis, to the best of our knowledge, no unsupervised generative deep background modelling methods have yet been proposed. Recently, deep generative models achieved a large breakthrough in the field of image generation and frame prediction, casting a light to generating background models based on unsupervised DCNNs. PixelCNN [15] introduced a filter mask and activation gate based CNN architecture for reconstruction of real world images. To deal with the resolution loss, PixelCNN applied a no-pooling architecture [10]. To both preserve the resolution and overcome the problem of short-range dependencies of convolutions, Unet [14] was proposed with a skip-pooling architecture which transmits some of the high resolution maps from shallow layers to deeper layers, skipping the pooling layers. Inspired by these methods, our BM-Unet is proposed to generate the background from frames. Specifically, the foregrounds in our method are regarded as 'noise' and we assume that the frames with foreground objects are in fact noisy background. Thus, background modelling deals with a similar problem as the image generation problem, which reconstructs images from destroyed ones.

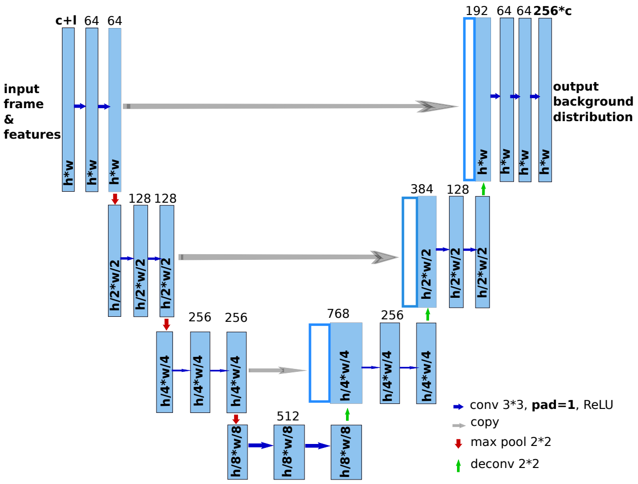

Figure 2. The architecture of our network is similar to [14], with the differences shown in **bold**. Each blue box corresponds to a multi-channel feature map. The number of the channels is denoted on top of the box. The h-w-size is provided at the lower center of the box. White boxes represent copied feature maps. The arrows denote different operations, defined in the bottom right corner.

3. Proposed Method

In this section we will describe the proposed method in detail. We will first present a simple baseline background modelling network, named Background Modelling Unet (BM-Unet), which takes as input a single frame and gives as the output the distribution of background. In Section 3.1, we will show how to train this network in an unsupervised manner. In Section 3.2, we will augment the network so that it can deal with the problem of jitter and illumination changes. In the heart of our baseline method is a network that takes an input image and then outputs a distribution of background. It is trained with random pairs of images, X and Y, from an image sequence. For illustration purposes, we show how the frame pairs are used for training and how the network generates the background distribution during testing in Figure 1.

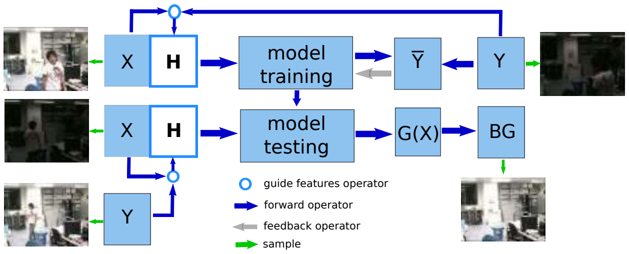

Figure 3. An illustration of our augmented method for background modelling.

3.1. BM-Unet

We regard the background modelling problem as a classification problem. We define $Y \in \mathbb{R}^{h \times w \times c}$ as the noisy target frame which we want to predict from the input frame $X \in \mathbb{R}^{h \times w \times c}$ taken from a video sequence, where c denotes the number of color channels (*e.g.*, gray image c=1, RGB image c=3), and h, w denote the height and width of frames, respectively. Our network, denoted as G, is based on Unet [14] which uses 'skip connections' to preserve the high frequency information, as illustrated in Figure 2. The network can be trained to predict the background from the frame X by minimizing the cross-entropy loss (L) between the approximated background distribution $G(X)$ and a noisy target frame distribution $\bar{Y}$, where $\bar{Y}$ is the one-hot encoded version of Y over 256 class labels, and $\bar{y}_{i,j,k}$ is the probability of j-th class in k-th color channel of i-th pixel in Y:

$$L(G(X), Y) = -\sum_{i=1}^{h \times w} \sum_{j=1}^{256} \sum_{k=1}^{c} \bar{y}_{i,j,k} \log(G(X)_{i,j,k}). \quad (1)$$

Our network can be seen as modelling the conditional distribution $p(Y|X)$ of pixels given the frame X.

In the case of unsupervised learning, no manually selected ground truth image is used in training. Instead, an input frame X and a noisy target frame Y are randomly chosen from the video sequence. Generally, X and Y are frames containing foreground objects, but the objects are at different locations. If we were to train the network on a specific pair of an input frame and a noisy target image, the network would over-fit to generate one frame from the other. Because of that we train the network on multiple random pairs of input frames and noisy target images. In this way, the network extracts common texture patterns that are more likely to belong to the background, under the assumption that the background is most often visible. At test time, the network generates the background distribution from the given random input frame.

3.2. Augmented BM-Unet

The network presented in the previous subsection can deal with sequences with a static background. However, videos of complex real world scenes usually contain more than one background due to illumination changes and/or camera jitter. Because of this, we introduce so called 'guide' features denoted by **H** which are used to guide the network to generate the background corresponding to the target frame. The modified version is named augmented BM-Unet. For illustration purposes, we show how the augmented network generates a background for a target frame in Figure 3.

Category	**videos**	
Dynamic Background	advertisementBoard	
Intermittent Motion	busStation	AVSS2007
Clutter	boulevardJam	Board
Very Short	CUHK_Square	DynamicBackground
Basic	Blurred	511
Illumination Change	CameraParameter	
Very Long	BusStopMorning	
Jitter	badminton	boulevard

Table 1. The categories of the sequences.

Given the low-level image description (features) represented as a latent matrix $\mathbf{H} = [H_1, \ldots, H_l] \in \mathbb{R}^{h \times w \times l}$, where l denotes the number of features, we seek to model the conditional distribution of backgrounds suiting this description, $p(Y|X, \mathbf{H})$. In order to achieve this, we replace the input X of BM-Unet in Equation 1 by the matrix X', where $X' \in \mathbb{R}^{h \times w \times (c+l)}$ is the concatenation of X and $\mathbf{H}$, where $\mathbf{H}$ represents the guiding features extracted from Y and X. In terms of illumination change, the mean intensity of an image can indicate its global illumination. Thus, the difference between the mean intensity of Y and X is denoted by h_1 and its corresponding latent matrix, which enables the background to follow the illumination of the target frame, is $H_1 = [h_1]$. In addition, dense optical flow [7], a feature used to detect camera motion, is also considered as a kind of a guiding feature. The two dimensions of the flow vector are denoted as H_2 and H_3 respectively, and are used to enable the background to follow the jitter of the target frame. At testing time, the network generates the background distribution given the random input frame and guiding features.

4. Experiments

Our method is evaluated on the SBMnet dataset which contains 79 different videos divided into 8 different challenge categories. The SceneBackgroundModeling.NET (SBMnet)[1] dataset was opened to the public by P. Jodoin *et al.* in 2016 as an affluent benchmarking framework, including the evaluation methodology and a set of relevant metrics. However, the ground truth images of the background are provided only for 13 sequences. The categories of these video sequences are shown in Table 1. For comparison purposes, SBMnet demands that only one background from a sequence can be used as the final result for comparing with

[1] http://scenebackgroundmodeling.net

the ground truth.

The proposed methods have been evaluated both quantitatively and qualitatively. The quantitative evaluation is conducted using the pipeline code provided by SBMnet, while the qualitative evaluations are shown in Figure 4. Our network is trained with a learning rate of 0.1 (which is decreased by half every 30 epochs) using the RMSProp optimization strategy and the training is stopped after 20,000 iterations. The output of our network is the background distribution. Thus, we obtain background images by computing the median of the first three possible values ranked by the probability of the pixel value. As our method generates the background from each single input frame, to compare with the ground truth provided by the data set, we take the pixel-wise median of all the computed backgrounds for each sequence, in a post-processing step.

4.1. Quantitative Evaluation Matrix

In terms of quantitative evaluation, the metrics are provided by the SBMnet dataset. Below, CB denotes computed background.

1. Average Gray-level Error (AGE) is the average of the gray-level absolute difference between the GT and CB images.
2. Percentage of Error Pixel (pEPs) is the percentage of EPs (number of pixels in CB whose value differs from the value of the corresponding pixel in GT by more than a threshold) with respect to the total number of pixels in the image.
3. Percentage of Clustered Error Pixels (pCEPS) is the percentage of CEPs (number of pixels whose 4-connected neighbors are also error pixels) with respect to the total number of pixels in the image.
4. MultiScale Structural Similarity Index (MSSI) is the estimate of the perceived visual distortion.
5. Peak-Signal-to-Noise-Ratio (PSNR) is the mounts to $10log_{10}((L-1)^2/MSE)$ where L is the maximum number of grey levels and MSE is the Mean Squared Error between GT and CB images.
6. Color image Quality Measure (CQM) is based on a reversible transformation of the YUV color space and on the PSNR computed in the single YUV bands. It assumes values in db and the higher the CQM value, the better is the background estimate.

For the metrics AGE, pEPs, and pCEPs the lower the better, for MSSI, PSNR, and CQM the higher the better.

4.2. Baseline BM-Unet

Our BM-Unet is evaluated on 13 image sequences provided in the SBMnet dataset, including the challenges of dynamic background, intermittent motion, clutter, jitter, illumination changes, very long and very short.

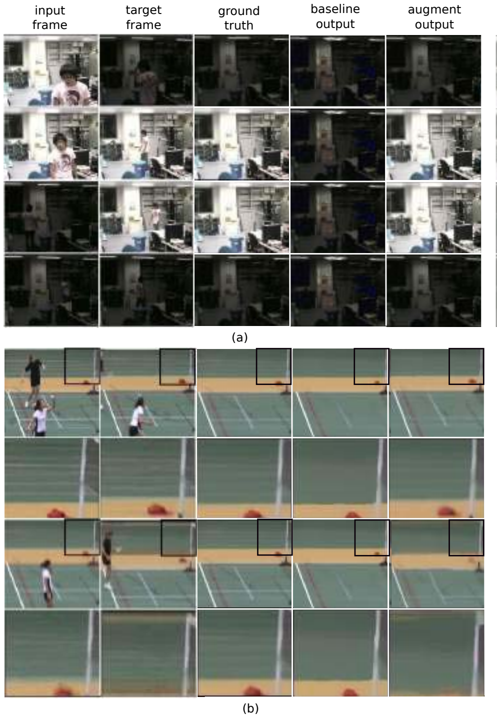

Figure 4. The qualitative results are sampled from the CameraParameter and badminton sequences. The first two columns are the input frames and target frames, respectively. The third column is the ground truth chosen manually from the dataset. The fourth column is the outputs of the BM-Unet baseline. The final column shows the backgrounds generated for the target frames from the input frames. (a) the sample results from CameraParameter; (b) the sample results from badminton, in which the second and fourth rows are the zoomed-in versions of the black box in previous rows, respectively.

For comparison, we use the results generated by five other BM methods, including Photomontage [1], BE-WiS [5], BE-AAPSA [13], LabGen-P [12] and FC-FlowNet. These are the top performing methods reported in the SBMnet dataset. Especially, FC-FlowNet is a convolutional neural network based method. For these methods we show the results published on the website SceneBackgroundModeling.Net, held in conjunction with ICPR 2016.

The comparison between our baseline method to other methods is shown in Table 2. The experimental results demonstrate that BM-Unet achieves very good results, comparable with the other considered methods. It is also possible to notice that no method is able to obtain the best performance over all the sequences. For 7 sequences, our method achieves the best result on most of the evaluation metrics. For other sequences it is not the best one, but it is still among the top of the considered methods. However, in terms of the qualitative evaluation, when the background is occluded by the same objects for a long period of time, our network can not generate the ground truth level results

Sequence	Method	AGE↓	pEPs↓	pCEPS ↓	MSSI↑	PSNR↑	CQM ↑
advertisement Board	Photomontage	2.4462	0.0039	0.0023	0.9844	34.9257	35.6554
	BEWiS	1.9999	0.0035	0.0017	0.9938	36.4428	37.0137
	BE-AAPSA	2.6457	0.0134	0.0084	0.9841	33.0200	33.6271
	FC-FlowNet	1.9866	0.0017	0.0002	0.9805	38.0131	37.8362
	LabGen-P	**1.6655**	**0.0001**	**0.0000**	**0.9966**	**40.7212**	**40.7527**
	BM-Unet	1.9261	0.0053	0.0018	0.9725	36.4918	36.9634
AVSS2007	Photomontage	12.0167	0.1047	0.0837	0.8400	19.2860	20.2173
	BEWiS	17.0903	0.1382	0.1128	0.7784	17.0107	18.0036
	BE-AAPSA	20.6172	0.1954	0.1630	0.7929	16.4960	17.5546
	FC-FlowNet	11.6751	0.1214	0.0966	0.8726	20.7442	21.7565
	LabGen-P	9.0258	0.0722	**0.0513**	0.8888	**22.1115**	**22.9735**
	BM-Unet	**8.2298**	**0.0696**	0.0515	**0.8892**	21.6466	22.4956
busStation	Photomontage	6.5309	0.0519	0.0380	0.8872	21.8651	22.8979
	BEWiS	3.5297	**0.0046**	**0.0013**	0.9826	**33.8541**	**34.4443**
	BE-AAPSA	4.5206	0.0334	0.0187	0.9621	30.0286	30.9833
	FC-FlowNet	4.3513	0.0253	0.0068	0.9622	31.1049	31.7573
	LabGen-P	**2.8988**	0.0083	0.0029	**0.985**1	33.1836	33.8982
	BM-Unet	4.2761	0.0270	0.0138	0.9501	26.8583	27.5910
boulevard Jam	Photomontage	12.1045	0.1546	0.0848	0.7604	20.9163	22.1436
	BEWiS	6.0621	0.0723	0.0400	0.8652	25.1959	26.3532
	BE-AAPSA	5.1418	0.0350	0.0090	0.9219	28.9114	30.0986
	FC-FlowNet	5.0200	0.0278	0.0087	0.8619	30.3476	31.4309
	LabGen-P	17.0916	0.2286	0.1255	0.5550	19.3248	20.5221
	BM-Unet	**4.3034**	**0.0173**	**0.0019**	**0.9679**	**31.4674**	**32.1057**
Board	Photomontage	13.4739	0.1098	0.0667	0.5029	18.8444	20.0911
	BEWiS	18.3758	0.2780	0.2382	0.6050	18.4726	19.5750
	BE-AAPSA	25.4532	0.3156	0.2498	0.7629	15.6631	16.9305
	FC-FlowNet	14.1523	0.1787	0.0999	0.8691	22.1587	23.2484
	LabGen-P	**6.4846**	**0.0398**	**0.0094**	**0.9077**	**28.4399**	**29.3575**
	BM-Unet	7.5180	0.0502	0.0114	0.8832	24.6694	25.3337
CUHK _Square	Photomontage	4.6470	0.0398	0.0022	0.9672	29.1463	29.6655
	BEWiS	5.1866	0.0384	0.0015	0.9628	29.3726	29.8880
	BE-AAPSA	12.1545	0.1463	0.0715	0.8273	20.5782	21.4139
	FC-FlowNet	8.1204	0.0917	0.0275	0.9190	26.2069	26.8343
	LabGen-P	4.9347	0.0330	0.0012	0.9655	29.8919	30.3907
	BM-Unet	**3.4368**	**0.0082**	**0.0002**	**0.9832**	**32.9033**	**33.3043**
Dynamic Background	Photomontage	10.4144	0.1345	0.0122	0.9268	24.5275	25.2235
	BEWiS	7.4114	0.0553	**0.0007**	0.9615	27.3600	28.0085
	BE-AAPSA	9.8202	0.1079	0.0135	0.9318	24.2536	24.9544
	FC-FlowNet	9.1948	0.0968	0.0197	0.9426	25.8003	26.5256
	LabGen-P	7.3852	**0.0539**	0.0010	0.9628	**27.5186**	**28.0590**
	BM-Unet	**7.3162**	**0.0539**	0.0012	**0.9631**	27.3917	27.9706
Blurred	Photomontage	2.0214	0.0003	**0.0000**	0.9941	38.2473	38.5613
	BEWiS	1.5186	**0.0000**	**0.0000**	**0.9976**	41.0599	41.2976
	BE-AAPSA	15.2057	0.2891	0.1978	0.8924	22.4556	23.3364
	FC-FlowNet	2.6962	0.0015	0.0002	0.9902	36.3751	36.8199
	LabGen-P	**1.4585**	0.0001	**0.0000**	0.9975	**41.3447**	**41.4575**
	BM-Unet	2.1951	0.0005	**0.0000**	0.9949	37.7622	38.0083
511	Photomontage	5.7977	0.0620	0.0021	0.9488	26.6706	28.7131
	BEWiS	3.5316	0.0242	0.0004	**0.9808**	**31.2115**	**33.0870**
	BE-AAPSA	4.0511	0.0303	0.0012	0.9744	30.0319	31.8292
	FC-FlowNet	3.9735	0.0321	0.0018	0.9735	30.8573	32.5541
	LabGen-P	4.7583	0.0511	0.0028	0.9496	27.8376	29.6666
	BM-Unet	**3.4354**	**0.0249**	**0.0003**	0.9796	30.3882	32.0455
Camera Parameter	Photomontage	**1.6291**	**0.0003**	**0.0000**	**0.9925**	**40.4538**	**40.7403**
	BEWiS	4.9567	0.0493	0.0306	0.9781	26.5173	27.2836
	BE-AAPSA	5.4333	0.0260	0.0131	0.9365	23.5066	24.6215
	FC-FlowNet	3.1284	0.0421	0.0305	0.9266	31.0808	32.0749
	LabGen-P	6.6294	0.0302	0.0262	0.9385	18.4163	20.1838
	BM-Unet	2.5533	0.0132	0.0016	0.9633	27.0713	27.3019
BusStop Morning	Photomontage	6.1219	0.0423	0.0007	0.9834	28.3680	29.0803
	BEWiS	6.0792	0.0233	0.0007	0.9857	29.2522	29.9708
	BE-AAPSA	5.7741	0.0317	0.0007	0.9836	29.4815	29.9870
	FC-FlowNet	5.6795	0.0265	0.0003	0.9833	29.8590	30.5611
	LabGen-P	5.8279	0.0240	0.0010	0.9849	29.3030	29.9402
	BM-Unet	**3.6418**	**0.0089**	**0.0002**	**0.9925**	**32.0084**	**32.5829**
badminton	Photomontage	4.2924	0.0392	0.0126	0.9237	29.6868	30.4911
	BEWiS	2.4827	0.0145	0.0114	0.9641	28.1868	29.3508
	BE-AAPSA	4.3975	0.0400	0.0239	0.9204	29.1352	29.9490
	FC-FlowNet	5.5368	0.0318	0.0123	0.9367	29.9097	30.7442
	LabGen-P	**2.2477**	**0.0097**	**0.0020**	**0.9798**	**34.9925**	**35.5817**
	BM-Unet	4.5987	0.0324	0.0103	0.9209	28.7756	29.5892
boulevard	Photomontage	9.7829	0.1309	0.0175	0.8995	21.6868	23.0513
	BEWiS	10.3209	0.1344	0.0198	0.8891	21.3146	22.8651
	BE-AAPSA	10.8262	0.1476	0.0297	0.8821	21.1393	22.5861
	FC-FlowNet	10.6830	0.1389	0.0261	0.8956	**22.5246**	**24.0208**
	LabGen-P	9.7258	0.1293	0.0202	0.8997	21.7984	23.2425
	BM-Unet	**8.6670**	**0.1040**	**0.0074**	**0.9087**	21.9716	23.2600

Table 2. Comparison of different results. ↓ denotes the lower the better, ↑ denotes the higher the better. The best results are shown in **bold**.

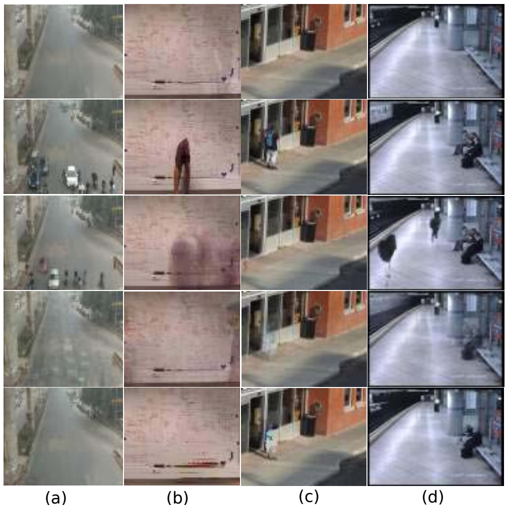

Figure 5. The qualitative results are sampled from (a) boulevard-Jam; (b) Board; (c) busStation; (d) AVSS2007. The first row is the ground truth; the following rows are the result of Photomontage, BEWis, FC-FlowNet and our BM-Unet, respectively.

but it still outperforms most existing methods, as shown in Figure 5.

It is also worth noticing that the baseline BM-Unet is robust to the length of the sequence, including very long and very short. Our network is trained individually from scratch for each sequence in an unsupervised manner. Thus, comparing with other deep methods, our method can be trained with only 7 frames while still giving promising results, which highly reduces the work needed for obtaining data.

4.3. Augmented BM-Unet

For illumination changes, we first use the augmented BM-Unet with $\mathbf{H} = [H_1]$, denoted by BM-Unet-H_1. For the jitter problem, we use the augmented BM-Unet with $\mathbf{H} = [H_1, H_2, H_3]$, denoted by BM-Unet-H. This experiment is conducted on 6 sequences, including sequences with/without the challenge of illumination changes or camera jitter. The quantitative comparison between the baseline and the augmented BM-Unet is shown in Table 3. From the quantitative evaluation for the illuminate change sequence 'CameraParameter', it can be seen that the augmented BM-Unet performs significantly better than the baseline, which can be explained by the advantage of the guiding feature H_1. In addition, we find that using the augmented BM-Unet still works well even when the sequences do not contain illumination changes or camera jitter. On the other hand, the results in the jitter category, such as 'boulevard', are not improved even with the guiding features H_2 and H_3. This is because there are multiple backgrounds in this sequence but only one ground truth background is provided. The qualitative evaluations which illustrate the effects of the guiding

Sequence	Method	AGE↓	pEPs↓	pCEPS ↓	MSSI↑	PSNR↑	CQM ↑
busStation	BM-Unet	4.2761	0.0270	0.0138	0.9501	26.8583	27.5910
	BM-Unet-H_1	**3.2852**	**0.0219**	**0.0102**	**0.9601**	**28.0970**	**28.9279**
	BM-Unet-H	5.5255	0.0465	0.0361	0.8958	22.3791	23.4362
Dynamic Background	BM-Unet	7.3162	0.0539	0.0012	0.9631	27.3917	27.9706
	BM-Unet-H_1	**7.1720**	0.0525	**0.0010**	0.9654	27.8452	28.3934
	BM-Unet-H	7.2555	**0.0520**	**0.0010**	**0.9664**	**27.8545**	**28.4228**
Blurred	BM-Unet	2.1951	**0.0005**	**0.0000**	**0.9949**	37.7622	38.0083
	BM-Unet-H_1	**2.1582**	0.0008	**0.0000**	0.9947	**37.7969**	**38.0203**
	BM-Unet-H	2.3079	0.0011	**0.0000**	0.9935	37.1141	37.3622
Camera Parameter	BM-Unet	2.5533	0.0132	0.0016	0.9633	27.0713	27.3019
	BM-Unet-H_1	**1.2100**	**0.0016**	**0.0000**	**0.9932**	**38.2071**	**38.9252**
	BM-Unet-H	1.5226	0.0029	0.0001	0.9918	37.694	38.0992
BusStop Morning	BM-Unet	3.6418	0.0089	0.0002	0.9925	32.0084	32.5829
	BM-Unet-H_1	3.7719	0.0103	0.0002	0.9925	31.6383	32.2838
	BM-Unet-H	**3.5764**	**0.0085**	**0.0002**	**0.9933**	**32.1029**	**32.7529**
boulevard	BM-Unet	**8.6670**	**0.1040**	**0.0074**	**0.9087**	**21.9716**	**23.2600**
	BM-Unet-H_1	9.4352	0.1171	0.0111	0.8988	21.4978	22.8129
	BM-Unet-H	9.6583	0.1259	0.0170	0.8939	21.4434	22.8259

Table 3. Comparison of baseline and augment BM-Unet. ↓ denotes the lower the better, ↑ denotes the higher the better. The best results are shown in **bold**.

features **H** are shown in Figure 4.

In order to show the performance in illumination changes, we show some results sampled from the sequence 'CameraParameter' in Figure 4.a. Figure 4.a shows that the baseline BM-Unet can only extract the most frequent background, the dark scene, while the augmented BM-Unet can generate the background corresponding to the target frame. Specifically, if the target frame is a light frame, then the generated background is a light background independent of input frame, and vice versa.

In order to show the performance in the jitter category, we show some results sampled from the sequence badminton in Figure 4.b. The augmented output matches the background of the target frame better, although the baseline output is more similar to the ground truth. The zoomed-in rows in Figure 4.b show clearly that the augmented output changes following the target frame.

5. Conclusions

In this paper, we have described an unsupervised deep background generative method, called BM-Unet. The key aspect of BM-Unet is its ability of generating the background of a sequence by unsupervised training. It can also generate the accurate background model even if no clear frames (*e.g.*, images without foreground objects) are present in the image sequence. We further proposed an augmented version of BM-Unet which uses guiding features to help the network deal with the illumination change and the camera jitter challenges.

The experiment results, obtained on the challenging sequences from the SBMnet dataset, demonstrate that BM-Unet's performance is robust to most of the challenges in all of the sequences, and that it is especially good in very long and very short sequences, quantitatively. Qualitatively, we have shown that augmented BM-Unet has a high ability of adapting to camera jitter and illumination changes and that it can correctly generate the corresponding background for the given target frame.

As future work, we intend to apply the BM-Unet to background subtraction by combining it with other subtraction methods.

References

[1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 294–302. ACM, 2004.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[3] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. S. Huang. Illumination normalization for face recognition and uneven background correction using total variation based image models. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 532–539. IEEE, 2005.

[4] Y. Chen, J. Wang, and H. Lu. Learning sharable models for robust background subtraction. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[5] M. De Gregorio and M. Giordano. Background modeling by weightless neural networks. In *International Conference on Image Analysis and Processing*, pages 493–501. Springer, 2015.

[6] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002.

[7] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image analysis*, pages 363–370, 2003.

[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[9] J. Hao, C. Li, Z. Xiong, and E. Hussain. A temporal-spatial background modeling of dynamic scenes. *Frontiers of Computer Science in China*, 5(3):290–299, 2011.

[10] V. Jain, J. F. Murray, F. Roth, S. Turaga, V. Zhigulin, K. L. Briggman, M. N. Helmstaedter, W. Denk, and H. S. Seung. Supervised learning of image restoration with convolutional networks. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[11] B. Laugraud, S. Piérard, M. Braham, and M. Van Droogenbroeck. Simple median-based method for stationary background generation using background subtraction algorithms. In *International Conference on Image Analysis and Processing*, pages 477–484. Springer, 2015.

[12] B. Laugraud, S. Pierard, and M. Van Droogenbroeck. Labgen-p: A pixel-level stationary background generation method based on labgen. In *2016 International Conference on Pattern Recognition Contest Proceedings*. IEEE, 2016.

[13] G. Ramirez-Alonso, J. A. Ramirez-Quintana, and M. I. Chacon-Murguia. Temporal weighted learning model for background estimation with an automatic re-initialization stage and adaptive parameters update. *Pattern Recognition Letters*, 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[15] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[16] Z. Zhao, X. Zhang, and Y. Fang. Stacked multilayer self-organizing map for background modeling. *IEEE Transactions on Image Processing*, 24(9):2841–2850, 2015.