ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/327630853

BSCGAN: Deep Background Subtraction with Conditional Generative Adversarial Networks

Conference Paper · September 2018

DOI: 10.1109/ICIP.2018.8451603

CITATION
1

READS
188

6 authors, including:

Mohamed Chafik Bakkay
Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'H…
16 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Hatem Rashwan
Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'H…
42 PUBLICATIONS **147** CITATIONS
SEE PROFILE

Houssam Salmane
Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'amén…
17 PUBLICATIONS **69** CITATIONS
SEE PROFILE

Louahdi Khoudour
Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'amén…
80 PUBLICATIONS **465** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

GAME-ABLING View project

Aeolix H2020 View project

All content following this page was uploaded by Mohamed Chafik Bakkay on 13 September 2018.

The user has requested enhancement of the downloaded file.

BScGAN: DEEP BACKGROUND SUBTRACTION WITH CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS

M. C. Bakkay[†], *H. A. Rashwan*[††], *H. Salmane*[†], *L. Khoudour*[†], *D. Puig*[††], *Y. Ruichek*[†††]

[†]CEREMA, Equipe-projet STI, Toulouse, France,
[††]DEIM, Universitat Rovira i Virgili, Tarragona, Spain,
[†††] CNRS-IRTES, Université de Technologie de Belfort, Montbéliard, France.

ABSTRACT

This paper proposes a deep background subtraction method based on conditional Generative Adversarial Network (cGAN). The proposed model consists of two successive networks: generator and discriminator. The generator learns the mapping from the observing input (i.e., image and background), to the output (i.e., foreground mask). Then, the discriminator learns a loss function to train this mapping by comparing real foreground (i.e., ground-truth) and fake foreground (i.e., predicted output) with observing the input image and background. Evaluating the model performance with two public datasets, CDnet 2014 and BMC, shows that the proposed model outperforms the state-of-the-art methods.

Index Terms— Background subtraction, Change detection, Generative Adversarial Networks, deep learning.

1. INTRODUCTION

Background subtraction (BS) is a crucial task for many applications, like tracking and video surveillance. BS techniques aim at detecting foreground objects from a stationary background by building a model of the static scene (background modeling) and comparing this model to input images (subtraction operation). However, BS faces many challenges such as dynamic background, shadows and illumination changes.

The classical methods focused on background modeling by generating a background model [1, 2, 3, 4]. These methods attempted to generate a background model and detect the pixels related to the foreground based on this model. Wren *et al.* [4] modeled every pixel by a single unimodal distribution. In addition, GMM [1] modeled every pixel with a mixture of k Gaussian to handle multiple backgrounds. To deal with fast variations backgrounds, a non-parametric technique KDE [2] was also proposed. Moreover, Kim *et al.* [3], with Codebook algorithm, summarized each background pixel by one or more codewords to cope with illumination changes. To take into account neighborhood texture description, robust descriptors, such as LBP, are used for describing background pixels [5].

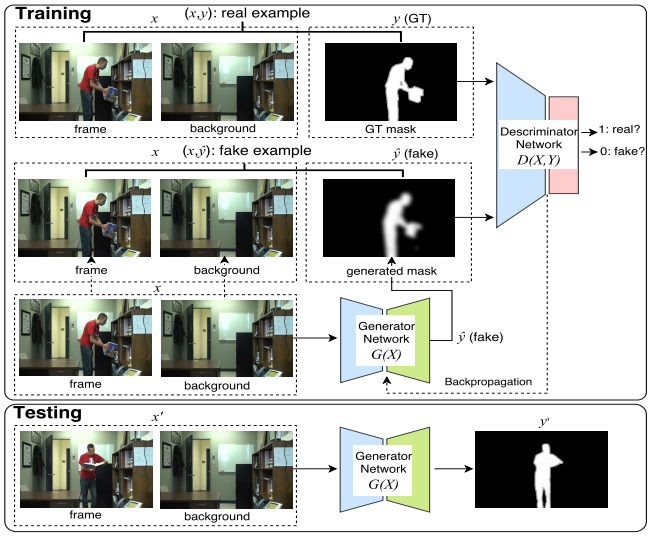


Fig. 1. Overview of BScGAN. For training, G generates a fake foreground mask $\hat{y}$ from each input x (i.e., an image plus a background). Then, D learns to discriminate between a fake example $\hat{y}$ and a real example y under condition of x. Backpropagating D and G leads to generate better masks. For testing, G outputs a foreground mask y' from each input x'.

Recently, deep learning methods that can directly learn features from raw images to detect foreground objects have been proposed [6, 7, 8]. For instance, Xu *et al.* [6] addressed the challenge of modeling background using two cascaded convolutional neural networks (CNN). However, instead of focusing on a complex background modeling strategy, other methods [7, 8] explored the subtraction operation itself based on a pre-trained CNN by computing the foreground probability for each pixel. Braham *et al.* [7] compared two small patches (centered on this pixel) extracted respectively from the input and background images. These methods [7, 8] considered the subtraction operation like a classification problem.

On the other hand, other deep learning approaches [9, 10]

This research was carried out at CEREMA. The authors are grateful for funding received for SAFER-LC project from the European Union's Horizon 2020 Framework program under Grant Agreement number: 723205-SAFER-LC-H2020-MG-2016-2017/H2020-MG-2016-Two-Stages.

addressed the problem of detecting foreground objects as a segmentation problem. Instead of using a background model, they used a single image as an input to predict the category (foreground/background) of each pixel. These methods used generally deep encoder-decoder networks (i.e., generator networks). However, the classical generator networks produce blurry foreground regions and such networks can not preserve the objects edges. Since they minimize the classical loss functions (e.g., Euclidean distance) between the predicted output and the ground-truth [11].

In this paper, we propose a deep background subtraction model using conditional Generative Adversarial Network (cGAN) [11], so-called 'BScGAN'. cGANs are deep learning models that can learn the statistical invariant features (texture, color, ...) of input images and then generate nearly synthetic images which look like the input images. The cGANs networks consist of two successive networks: generator and discriminator. The generator network learns the mapping from the input to the output, while the discriminator learns a loss function to train this mapping by comparing the ground-truth and the predicted output. Finally, the cGAN network optimizes a loss function that combines a conventional binary cross-entropy loss with an adversarial term. The adversarial term (i.e., discriminator) encourages the generator to produce output that cannot be distinguished from ground-truth ones.

Our cGAN model combines both subtraction and segmentation operations (Figure 1). Foreground detection is addressed as a segmentation problem instead of classification one; this segmentation is carried out by the generator network. At the same time, the cGAN network observes the original and background images concatenated as a condition for improving the network optimization. Furthermore, in order to reduce the processing time, we process entire images instead of dividing them into patches [7]. Thus, the contribution is twofold: 1) We present, to the best of our knowledge, the first application of the generative adversarial training for background subtraction. 2) The adversarial term yields more accurate foreground detection than the state-of-the-art methods, without adding algorithmic complexity to the model, since the output of a single network (generator) is only used instead of utilizing cascade networks as proposed in [6].

2. PROPOSED CGAN MODEL FOR BACKGROUND SUBTRACTION

The architecture of our model, BScGAN, is based on two modules as shown in figure 2: the generator and discriminator networks combine their "efforts" to predict foreground objects for a given image and a given background. This section provides details on the structure of two modules: considered loss functions and initialization.

2.1. Generator network

BScGAN follows an encoder-decoder architecture of Unet network with skip connections [12], where the encoder part includes downsampling layers that decrease the size of the feature maps followed by convolutional filters, while the decoder part uses upsampling layers followed by deconvolutional filters to construct an output image with the same resolution of the input one.

In BScGAN, (Figure 2), the encoder consists of 8 convolutional layers as proposed in [11]. The first layer uses 7×7 convolution to generate 64 feature maps. The 8th layer generates 512 feature maps with a 1×1 size. Their weights are randomly initialized. Furthermore, the middle six convolutional layers are six ResNet blocks that are initialized with the weights of a ResNet-101 model [12]. In all encoder layers, Leaky-ReLU non-linearities are used.

The decoder architecture is structured in the same way as the encoder one and includes 8 deconvolutional (e.g., Transpose Convolution) layers, but with a reverse layers ordering, and with downsampling layers being replaced by upsampling layers. The weights of the decoder layers are randomly initialized. All the deconvolutional layers use ReLU functions except the 8th 1×1 deconvolution layer that use Tanh activation to produce the final foreground objects binary mask.

2.2. Discriminator network

The discriminator network is composed of 4 convolutional and downsampling layers (Figure 2). The first layer generates 64 feature maps. Moreover, the 4th layer generates 512 feature maps with a 30×30 size. All convolutions are 3×3 spatial filters applied with a value of 2 for stride parameter. Their weights are randomly initialized and they use leaky-ReLU functions as activations. The last convolutional layer is followed by one fully connected (FC) layer. This FC layer is applied to transform the features map in a 1 dimensional vector and followed by a Sigmoid function.

2.3. Training

The BScGAN model has been trained over a loss [11] function resulting from combining a content and an adversarial losses. The content loss follows a classical approach in which the predicted foreground mask is pixel-wise compared with the corresponding one from ground-truth. In turn, the adversarial loss depends of the real/fake prediction of the discriminator over the ground-truth and the predicted foreground mask with observing the input image as a condition.

Given an input x (an image and a background), the generator G represents the predicted foreground mask $\hat{y}$ as a vector of probabilities of each pixel. The content loss function $\ell_{MSE}(G)$ is computed between $\hat{y}$ and its corresponding ground-truth y. Since mean squared error (MSE) aims at

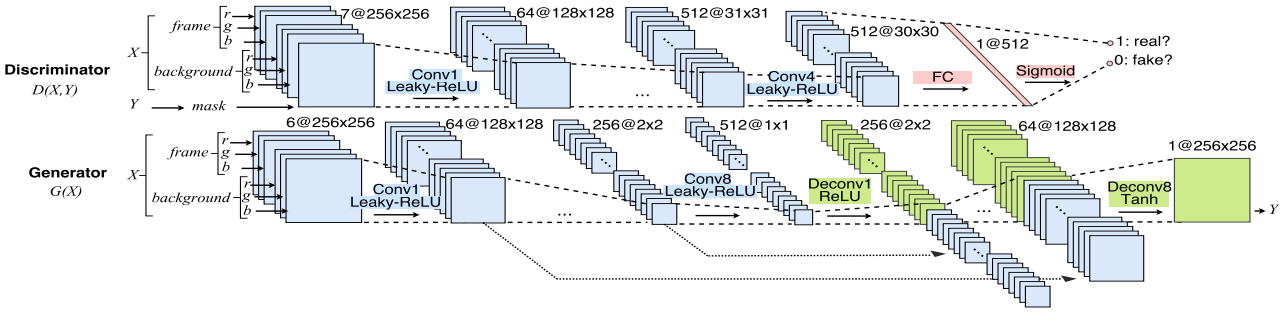


Fig. 2. Architecture of the proposed cGAN model for background subtraction.

maximizing the peak signal-to-noise ratio (PSNR), MSE is used as a content loss in our model that is defined as:

$$\ell_{MSE}(G) = \frac{1}{N}\sum_{j=1}^{N} \|y - \hat{y}\|_2, \tag{1}$$

where N is the number of the pixels per input image.

Otherwise, the adversarial loss used in cGAN models) [13] is the binary cross entropy (BCE) defined as:

$$\ell_{cGAN}(D,G) = \mathbb{E}_{x,y}[\log(D(x,y)] + \mathbb{E}_{x,\hat{y}}[\log(1 - D(x,\hat{y})], \tag{2}$$

where the first term is the prediction entropy of the discriminator D with the real data (i.e., ground-truth y), and the second term is the prediction entropy of D with the fake data (i.e., predicted mask $\hat{y}$). Both predictions are under observation of input images x.

The foreground detection problem has some differences from the cGAN scenario proposed [13]. First, the objective of our model is to fit a deterministic function that generates realistic foreground values from images, rather than realistic images from random noise. In addition, in our case, the input to G does not observe random noise, however, it observes two images (i.e., an image and a background). Second, it is clear that knowledge of the output, in our case, is a foreground mask that is essential to evaluate its quality. We therefore include both input and background images with the ground-truth mask as a real data to D. On the other hand, we include input and background images with the predicted foreground mask as a fake data to D. Training proceeds alternating between training D and training G, by keeping the weights of D constant and back-propagating the error through D to update the weights of G. In such case, when updating the parameters of G, we found that using the loss function, that is a combination of the error from D and the cross entropy with respect to the ground truth, improved the stability and convergence rate of the adversarial training. Thus, the loss function of G during adversarial training is formulated as:

$$\ell_{GD}(D,G) = \ell_{cGAN}(D,G) + \lambda \ell_{MSE}(G), \tag{3}$$

where, $\lambda = 10$. During training, D tries to maximize our function $\ell_{GD}(D,G)$, while the task of G is exactly the opposite that tries to minimize the function $\ell_{GD}(D,G)$.

$$G^* = arg \min_G \max_D \ell_{GD}(D,G), \tag{4}$$

The Adam solver [17] is used in optimizing the proposed model, with learning rate 0.0002, and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Batch size equals 4 and the number of epochs for training is 200 with random jitter and mirroring.

Regarding to the testing phase, each observation input x (i.e., an image, a background) is given to G which will generate an output y for a foreground mask.

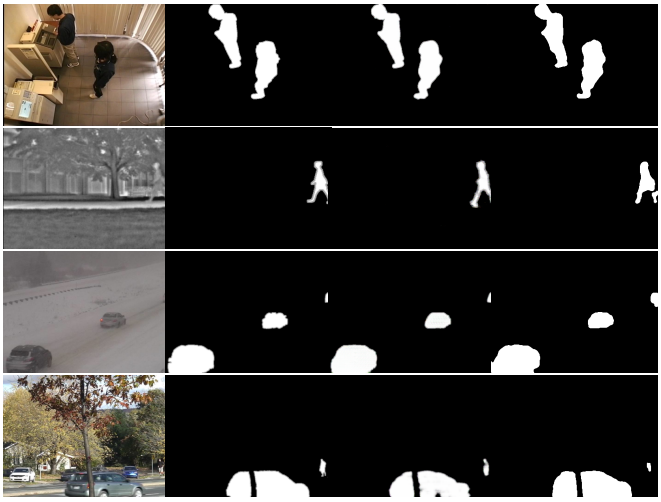
3. EXPERIMENTAL RESULTS

We evaluated our method on two public datasets: CDnet 2014 [18] and BMC [19] datasets. For CDnet 2014, we computed three performance measures: **recall**, **precision**, and **F-measure**.Regarding to BMC dataset, we computed **recall**, **precision**, Peak signal-to-noise ratio (**Psnr**), **F-measure**, **D-score** and **SSIM** detailed in [19]. Since the input images are coming from different sources, we decided to resize every image into (256×256) pixels. For the background images, we have taken into account the median value for all the considered background images for a given sequence [7].

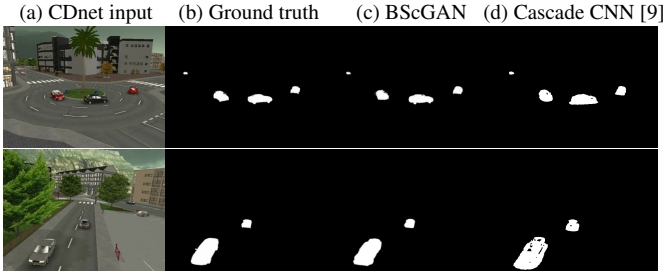
CDnet 2014 dataset: It includes 11 video categories. These categories correspond to different challenging situations (camera jitter, background motion, night videos...). For each sequence, the first half of the images is used as a training set, while the second half is used as a testing set. The proposed method is compared to six background subtraction methods [7, 9, 16, 8, 14, 15]. Three of them [7, 9, 8] are based on deep learning models. The other three methods are considered for the comparison, since they provide high accuracy with CDnet 2014 benchmark. For each video category, the F-measure values for all the methods are reported in Table 1. In this table, we can notice that BScGAN outperforms the evaluated methods with an average F-measure around 0.97%. This is particularly true for baseline category (0.9930). Qualitative results of our method and the Cascaded CNN method

Table 1. Overall and per-category F-measures for different methods on CDnet 2014 dataset (best accuracies are in **bold**).

Method	overall	baseline	jitter	intermittent	dynamic	shadows	thermal	badWeather	lowFramerate	night	turbulence
BScGAN	**0.9763**	**0.9930**	**0.9770**	**0.9623**	**0.9784**	**0.9828**	**0.9612**	**0.9796**	**0.9918**	**0.9661**	**0.9712**
Cascade CNN [9]	0,9213	0.9786	0.9758	0.8505	0.9658	0.9593	0.8958	0.9431	0.8370	0.8965	0.9108
Braham *et al.* [7]	0.9046	0.9813	0.9020	-	0.8845	0.9454	0.8543	0.9264	0.9612	0.7565	0.9297
DeepBS [8]	0,7891	0.9580	0.8990	0.6098	0.8761	0.9304	0.7583	0.8301	0.6002	0.5835	0.8455
SuBSENSE [14]	0,7801	0.9503	0.8152	0.6569	0.8177	0.8986	0.8171	0.8619	0.6445	0.5599	0.7792
PAWCS [15]	0,7682	0.9397	0.8137	0.7764	0.8938	0.8913	0.8324	0.8152	0.6588	0.4152	0.6450
IUTIS-5 [16]	0,8060	0.9567	0.8332	0.7296	0.8902	0.9084	0.8303	0.8248	0.7743	0.5290	0.7836



(a) CDnet input (b) Ground truth (c) BScGAN (d) Cascade CNN [9]



(a) BMC input (b) Ground truth (c) BScGAN (d) Hofmann *et al.* [20]

Fig. 3. Results with sequences of CDnet and BMC datasets.

proposed in [20] are shown on Figure 3 with four categories: from up to down of the four first rows: shadow, thermal, bad weather and dynamic background. As shown, the small objects and the boundaries are properly detected.

BMC dataset: It is a benchmark dataset and evaluation that contains synthetic videos representing urban scenes acquired from a static camera. It focuses on outdoor situations with weather variations such as wind, sun or rain. This dataset is composed of 20 synthetic urban video sequences (10 sequences for training and 10 sequences for testing). In Table 2, the quantitative results of the proposed model BScGAN with ten testing sequences are compared to four recent state of the art methods [4, 20, 21, 22] which provide the best results according to [23]. That is why we compare our method to these four other methods. In Table 2, we can notice that our method outperformed the four methods for recall, precision, F-measure, Psnr, D-Score, Ssim. This is particularly true for F-measure score. In addition, visualized results of two examples of synthetic images are shown in Figure 3. Again, with the BMC dataset, the small objects in the scenes are also well detected showing the ability of BScGAN to properly detect foreground objects in difficult situations.

Table 2. Global score of some methods evaluated on BMC data set (best accuracies are in **bold**).

Method	Recall	Precision	F-mesure	Psnr	D-Score	Ssim
BScGAN	**0.926**	**0.965**	**0.945**	**52.313**	**0.0007**	**0.996**
Hofmann *et al.* [20]	0.923	0.852	0.885	49.412	0.002	0.994
Yao *et al.* [21]	0.893	0.863	0.875	49.398	0.001	0.993
Maddalena *et al.* [22]	0.838	0.907	0.867	50.553	0.001	0.992
Wren *et al.* [4]	0.795	0.922	0.853	51.394	0.001	0.993

The model was implemented with PYTORCH on a GPU GeForce GTX 1080 with $8GB$ memory. BScGAN can achieve 400 images per second running on the GPU, while 10 images per second on a CPU Intel $7700HQ$ @ $3.60GHz$ with $32GB$ memory. Thus, it is obvious that BScGAN can improve foreground detection with a low computing time.

4. CONCLUSION

This paper has proposed a novel deep learning background subtraction model. This model is based on training a conditional Generative Adversarial Network (cGAN) that consists of two networks: generator and discriminator. This approach outperforms, in terms of background subtraction, several well known methods of the literature like cascaded neural networks without adding algorithmic complexity. The results show that BScGAN is robust to many challenges like dynamic background, shadows and illumination changes. In addition, the quantitative evaluation with two public datasets (CDnet 2014 and BMC) shows promising results. As a perspective work, the proposed model will be applied to life situations scenarios in the framework of SAFER-LC European project, dealing with safety at level crossings.

5. REFERENCES

[1] T. Bouwmans, F. El Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection-a survey," *Recent Patents on Computer Science*, vol. 1, no. 3, pp. 219–237, 2008.

[2] Ahmed Elgammal, David Harwood, and Larry Davis, "Non-parametric model for background subtraction," in *ECCV*. Springer, 2000, pp. 751–767.

[3] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Background modeling and subtraction by codebook construction," in *ICIP*. IEEE, 2004, vol. 5, pp. 3061–3064.

[4] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *PAMI*, vol. 19, no. 7, pp. 780–785, 1997.

[5] B Vishnyakov, V Gorbatsevich, S Sidyakin, Y Vizilter, I Malin, and A Egorov, "Fast moving objects detection using ilbp background model," *ISPRS*, vol. 40, no. 3, pp. 347, 2014.

[6] Pei Xu, Mao Ye, Xue Li, Qihe Liu, Yi Yang, and Jian Ding, "Dynamic background learning through deep auto-encoder networks," in *ACM Multimedia*. ACM, 2014, pp. 107–116.

[7] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *IWSSIP*, Bratislava, Slovakia, May 2016, pp. 1–4.

[8] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.

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

[10] Long Ang Lim and Hacer Yalim Keles, "Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding," *arXiv preprint arXiv:1801.02225*, 2018.

[11] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[14] P. St-Charles, G. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *TIP*, vol. 24, no. 1, pp. 359–373, 2015.

[15] P. St-Charles, G. Bilodeau, and R. Bergevin, "A self-adjusting approach to change detection based on background word consensus," in *WACV*. IEEE, 2015, pp. 990–997.

[16] S. Bianco, G. Ciocca, and R. Schettini, "How far can you get by combining change detection algorithms?," in *ICIAP*. Springer, 2017, pp. 96–107.

[17] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] Y. Wang, . Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P.h Ishwar, "Cdnet 2014: an expanded change detection benchmark dataset," in *CVPR-Workshops*, 2014, pp. 387–394.

[19] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequièvre, "A benchmark dataset for outdoor foreground/background extraction," in *ACCV*. Springer, 2012, pp. 291–300.

[20] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *CVPR-Workshops*. IEEE, 2012, pp. 38–43.

[21] J. Yao and J. Odobez, "Multi-layer background subtraction based on color and texture," in *CVPR*. IEEE, 2007, pp. 1–8.

[22] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *TIP*, vol. 17, no. 7, pp. 1168–1177, 2008.

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

View publication stats