

Deep Autoencoders with Value-at-Risk Thresholding for Unsupervised Anomaly Detection

Albert Akhriev and Jakub Marecek

Abstract—Many real-world monitoring and surveillance applications require non-trivial anomaly detection to be run in the streaming model. We consider an incremental-learning approach, wherein a deep-autoencoding (DAE) model of what is normal is trained and used to detect anomalies at the same time. In the detection of anomalies, we utilise a novel thresholding mechanism, based on value at risk (VaR). We compare the resulting convolutional neural network (CNN) against a number of subspace methods, and present results on changedetection.net.

Index Terms—convolutional autoencoder, incremental training, background subtraction

I. INTRODUCTION

Consider the problem, where starting from high-dimensional streamed data, one should like to distinguish between “normal” input of time-varying nature (also known as background) and “anomalies” (events, or foreground). This problem is known as anomaly or event or outlier detection in Data Engineering [1], [2] and as background subtraction in Computer Vision [3].

Early methods studied the data (block-)coordinate-wise. For example for image data, these methods worked pixel-by-pixel [4]–[7], where each pixel is a block of three coordinates. Subsequently, the view of anomaly detection as a low-rank matrix-completion problem has become popular. A typical approach stacks a number of recent flattened observations, e.g. video frames, into rows of a data matrix, which is then approximated via a low-rank matrix. The low-rank model corresponds to the background, and anomalies are outside of the low-rank subspace [8]–[12]. In Signal Processing [13]–[15], there is much related work on robust principal component analysis (RPCA) and subspace tracking. Extensive literature surveys can be found in [16]–[18].

More recently, deep-learning techniques have been developed, based on the matrix-completion view of anomaly detection [2], [19]. Notably, autoencoder architecture have proven successful, in practice. A multi-scale framework, proposed in [20], encodes the input images by means of pre-trained VGG-16 followed by a sub-net that pools features at multiple scales before feeding them into decoder. Authors claim robustness against camera jitter and shadows despite a very limited number of labelled images used for training. In [21] authors extended ideas of [4], [5] by training a CNN on image patches. Subsequently, the trained CNN was applied to assemble foreground mask from patches on previously unseen images. Likewise a patch-based CNN in [22] learns to output foreground probability on a small number of labelled training examples. Double-autoencoder network, introduced in [23], reconstructs background in two stages. It requires some initial training but afterwards can proceed in online fashion.

There are also sparse [24], robust [25] variants and variants combining auto-encoders with Gaussian mixture models [26]. For a broader view of recent advances in deep architectures for background modelling, reader is advised to consult the comprehensive surveys [3], [27].

A key challenge across subspace and deep-learning approaches is the amount of supervision and tuning. Interestingly, many approaches therein rely on very simple thresholding mechanisms that require extensive tuning and output quite noisy foreground masks. In particular, across both traditional methods [4], [5], methods based on matrix-completion [28], and autoencoders [26], the use of Gaussian mixture models (GMM) is the state of the art. While there are plausible alternatives [29, e.g.], the typical use of GMM involves the use of expectation-maximization (EM) heuristics, which suffer from a host of issues, including the sensitivity to noise and sensitivity to balance in the mixing coefficients, as well as getting stuck in arbitrarily poor local optima. One would hence like to obtain an unsupervised approach, without the GMM assumptions.

Our main contribution is a technique for unsupervised use of deep autoencoders:

- the use of a thresholding mechanism based on value at risk (VaR), which can be computed exactly in time required to sort the incoming data.
- a novel weighing (pre-processing) of the input to the autoencoder.
- a numerical study of deep-autoencoders and matrix-completion methods with a variety of thresholding methods on changedetection.net.

The use of VaR-based thresholding makes it possible to adapt deep-learning approaches into unsupervised methods without data-dependent tuning.

The paper is organised as follows: in Section II, we describe a variant of background subtraction algorithm based on a convolutional autoencoder. In Section IV, it is compared against selected low-rank approximation methods. Both approaches can be seen as non-linear subspace tracking [30], which motivated our choice of comparison candidates. Section V summarises our vision of future efforts.

II. THE UNSUPERVISED CONVOLUTIONAL AUTOENCODER

Since having labelled data is unaffordable in many real-world applications, an unsupervised approach to anomaly detection is desirable. Our approach uses a convolutional neural networks (convolutional autoencoder) without an explicit training phase for anomaly detection in streamed data. In particular, it uses incremental training of a model of what is normal

(background), without any supervised data, and concurrent use of the model, to estimate what is normal (background).

A baseline deep convolutional autoencoder (DAE) could be seen as a generalisation of low-rank approximation methods. Upon arrival of a new frame I of a stream, we do exactly one forward/backward iteration in order to train the autoencoder (*update phase*), and then draw an estimate of background model B as an output of the autoencoder network (*reconstruction phase*). L_1 -norm loss function minimises the difference between I and B , effectively ignoring the outliers (here, points of moving objects).

Figure 1 depicts the architecture of our autoencoder. Except the first and the last layer, all other convolution layers have 64 filters of size 5×5 (input and output number of channels equals to 64), interleaved with non-linear activations. Experimentally, we found that hyperbolic tangent function (\tanh) works better than logistic function or rectified linear unit.

In a pre-processing step, the input is flattened, block-coordinate to a single coordinate (e.g., red-green-blue to grayscale), before being fed into the autoencoder. (Note that at night time, there are no colours in video data anyway.) Each frame is then reshaped into a “standard” layout ($1 \times 576 \times 704$) and the values are normalised to $[-0.5 \dots 0.5]$ range. Reconstructed background and foreground mask undergo the inverse transformation. The bottleneck layer is represented by 1D tensor of size 2048, surrounded by fully-connected layers. Encoder’s layers are shrunk by a factor of 2 (stride 2), as the data propagate from input to output, and decoder’s (transposed) layers are expanded by the same factor of 2 (stride 2), respectively. Here we try to balance between network depth (that increases computational burden) and background reconstruction quality.

In early experiments, we used single image as an input and output respectively. This approach demonstrated solid results compared to other methods. Nonetheless, dynamic background case can be handled better, if we admit multiple images in the process of outlier thresholding. We have at least two options here. First, accumulate a short history of recent frames (50 in our experiments, or 2 seconds of video). In each iteration, pick up a subset of 10 images uniformly spread over this short history, and stack them into $10 \times H \times W$ tensor, where W, H stand for width and height respectively. This tensor is fed into an autoencoder as a multi-channel image $I(c, i)$ with “colour channels” $c = 1 \dots 10$ and points $i = 1 \dots (H \times W)$. The reconstructed background $B(c, i)$ has the same layout. The goal is to make training procedure less prone to overfitting.

Multi-channel input image increases the number of weights in the first and the last convolution layers making forward/backward passed more compute-intensive. The second option, also adopted in this study, presents a trade-off between CPU/GPU load and foreground detection quality. Namely, we use approach similar to described above except a multi-channel input is replaced by a mini-batch of 10 input/output images. This reduces the number of weights in convolutional layers but, in theory, also lessens the flexibility of autoencoder network. In practice, we found no difference and both options produce very similar results up to minor variations attributed

to randomization in Xavier’s initialization of network weights.

Algorithm 1 summarizes the main steps of proposed approach 1) transform a new frame into a “standard” form; 2) update a time window of a history of 50 recent frames; 3) compute weights using the optic-flow algorithm (2) and plug them into the loss function (1); 4) make one forward and one backward step in training the autoencoder (cf. Figure 1); 5) reconstruct background model B , and compute the residuals; 6) estimate the optimum threshold using Value at Risk, and apply it to the residuals; 7) output a binary mask outlier/background reshaped back to the original size.

Algorithm 1 A single step of the incremental autoencoder training and anomaly detection

- 1: **Input:** Next element from a data stream, e.g., one frame of video data.
 - 2: **Output:** a binary mask suggesting what is an anomaly.
 - 3: Reshape the data to suit the convolutional network and normalise the values, e.g., to an image in $1 \times 576 \times 704$ resolution, linearly transformed to $[-0.5 \dots 0.5]$ range.
 - 4: Update a time window considered, e.g., a history of 50 recent images.
 - 5: Compute weights (2) using the optic flow and plug them into the loss function (1).
 - 6: Make one forward and one backward step in training the autoencoder.
 - 7: Reconstruct a model of what is normal (background) B , and compute residuals r_i .
 - 8: **return** `threshold(r_i)` of Algorithm 2, which estimates the optimum threshold, and applies it to the residuals.
-

A. DIS-Based Weights

As pointed out in [23], the robustness of the loss function L can be improved by introducing point-wise weights:

$$L = \sum_{c,i} w_i |I(c, i) - B(c, i)|, \quad (1)$$

where index i runs through all the pixels, and w_i is close to 0 for a moving point, and close to 1 otherwise. Because filters in CNN are shared by all the points, few remaining outliers would not harm the training process and will be suppressed by L_1 -norm loss function. One possible option for weight computation is to decrease the weights using the DIS algorithm [31]:

$$w_i = \exp(-v_i^2 / (2 \cdot \text{median_over_image}(v^2))), \quad (2)$$

where v_i is a velocity at point obtained by optic-flow computation from a pair of consecutive input frames $I^{(t-1)}$ and $I^{(t)}$, at times $t-1, t$. The scaling factor inside the exponent is computed as a median squared velocity over all points. In this way, the moving points are effectively down-weighted during the training phase. Alternatively, the masks output by fast and reliable algorithms introduced in [6], [7] could be used as the weights. Note that it would be sufficient to mask out the majority of foreground points. Because filters in a convolutional network are shared by all the points, few remaining outliers would not be harmful for the training process and will be suppressed by L_1 -norm loss function.

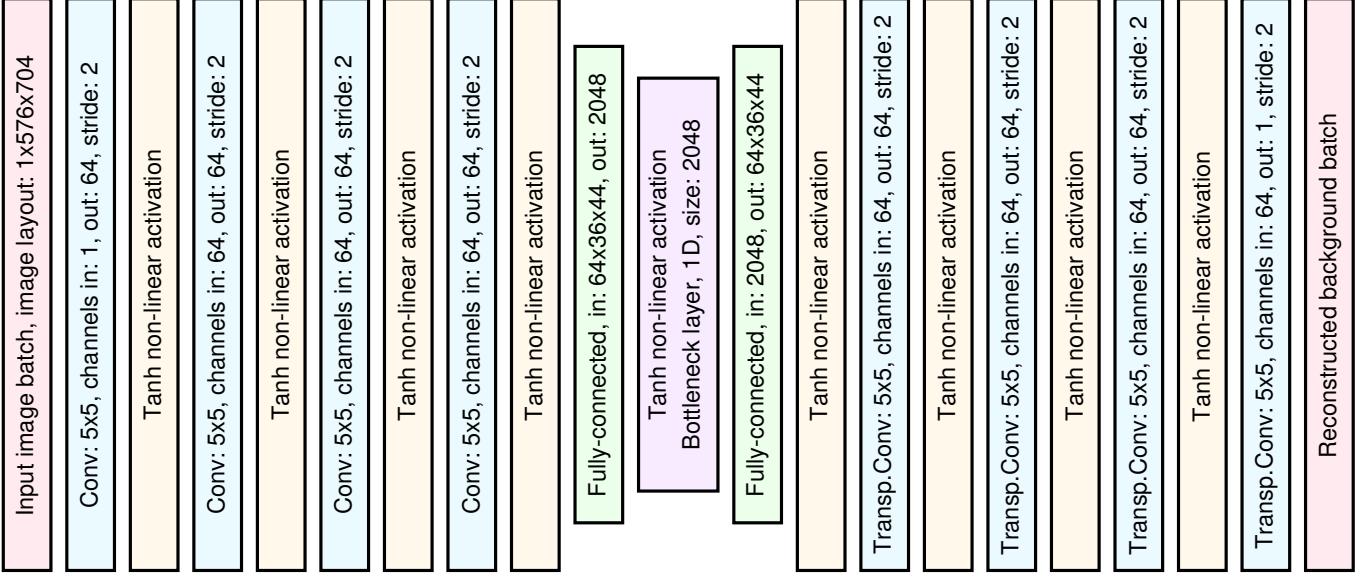


Fig. 1. The architecture of the convolutional autoencoder. Note that after the second fully-connected layer, we reshape the intermediate tensor to $64 \times 36 \times 44$ for subsequent decoding, where the number of channels (64) comes first in PyTorch convention. Also, parameter “padding=2” is specified in all convolutional layers and additionally “output_padding=1” in transposed convolutional layers (Transp.Conv) respectively.

III. THE VAR-BASED THRESHOLDING

A key step in the use of low-rank or autoencoding approaches is thresholding. Therein, one considers the so-called *residual map*, which is an array of the same dimensions as the input I and background B . In this study, the residual r_i at i th point is defined by:

$$r_i = \min_{k \in \mathcal{N}_i, c=1 \dots 10} |I(1, i) - B(c, i)|, \quad (3)$$

wherein \mathcal{N}_i are the points in 3×3 vicinity of (central) point i , c ranges over 10 recently reconstructed backgrounds, and a variety of norms (distance functions) can be used, outside of the absolute value of (3). Based on the residual map, the thresholding produces a binary-per-block-coordinate array $r_i > t$, which suggest what are the anomalies (moving objects) and what constitutes the normal background process.

We should like to stress that thresholding is a vast area by itself, even when restricted to anomaly detection [32]. Although locally adapted threshold, e.g. [33], [34], may work best, it is quite common to choose a single, even hard-coded, threshold for each frame. We follow the same practice by choosing automatically estimated, global threshold for every block-coordinate in residual map. Among the techniques for automatic threshold selection, we found that classical methods and their variations such like Otsu’s one [35], where threshold is determined by minimizing intra-class intensity variance, or maximum entropy method in [36] do not produce convincing foreground/background segmentation. The comprehensive surveys [32], [37], [38] give a good insight into the related methods.

Apparently, a threshold estimator should take some spatial information into consideration in order to make an “optimal” decision. At a high level, we seek a threshold for the *highest*

sensitivity, when isolated noisy points “just” show up. To this end, we need to define isolation and sensitivity.

In defining isolation, we have adopted a simple, yet efficient, mechanism first proposed by Malistov [39]. He estimated the probability of formation of a false “object” when 4 or more adjacent points exceeded a threshold. From that consideration, he deduced an optimum threshold value. We extended the idea by testing spatial neighborhood of a pixel and examining 3×3 contiguous patches that have 1 or 2 pixels, including the center, marked as an anomaly. This is illustrated in Figure 2.

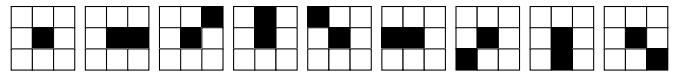


Fig. 2. The configurations of the 3×3 contiguous patches, whose fraction within all the 3×3 contiguous patches is sought.

For a given threshold, one analyses 3×3 patches centered at each point in the residual map. There are several cases how residual value r_i at the central point relates to its neighbours. Let us consider one example. Suppose that the central value r_1 is the largest one and we pick up the second r_2 and the third r_3 largest ones from the 3×3 patch, $r_3 \leq r_2 \leq r_1$. Suppose that all the values are integral, as usual in computer-vision applications. If a threshold happens to lie in the interval $[r_3 + 1 \dots r_1]$, then one of the patterns depicted on Figure 2 will show up after thresholding. As such, this particular central point “votes” for the range $[r_3 + 1 \dots r_1]$.

One can view the “votes” as forming a probability mass function of a random variable supported on the range of possible values of the thresholds, e.g., $[0, 255]$. Figure 3 suggests to see this as an *histogram of thresholds*. In particular for the center point of the previous paragraph, we would

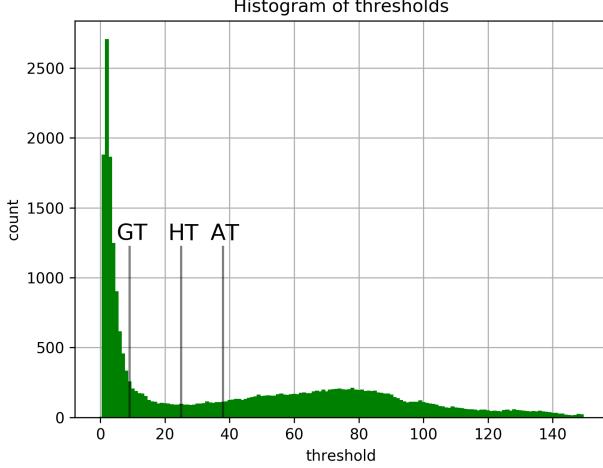


Fig. 3. An example of the histogram of thresholds. The histogram was truncated from the original 255 size to allow for some clarity of presentation. Three threshold values read as follows: GT – “ground-truth” threshold obtained by minimizing a mismatch between binarized image of residuals and the ground-truth foreground mask provided in cdnet 2014; HT – “hard” threshold, here 25; AT – “automatic” threshold obtained by Algorithm 2.

increment counters in the bins corresponding to threshold values $r_3 + 1$ to r_1 in the histogram. Repeating the process for all the points and all combinations of 3 largest residuals r_1, r_2, r_3 , we arrive at the information as to the number $H_T(t)$ of (central) points of one of the patterns of Figure 2 would be observed if the threshold t were used. In other words, $H_T(t)$ gives a rate of appearance of patterns in Figure 2 as a function of selected threshold t . Figures 3 to 5 give an example of a histogram of thresholds, a histogram of residuals and a *log*-transformed histogram of residuals respectively.

In defining sensitivity, consider a discrete random variable X , and risk measures thereof. While it is customary to analyze the histogram of residuals, e.g. [32], [35]–[37], [40], [41], all our calculations are done over the histogram of thresholds introduced above. As risk measures, we consider the value at risk (VaR):

$$\text{VaR}_\alpha(X) := \min\{c : P(X \leq c) \geq \alpha\}, \quad (4)$$

and the conditional value at risk (CVaR):

$$\text{CVaR}_\alpha(X) := \mathbb{E}[X : X \geq \text{VaR}_\alpha(X)]. \quad (5)$$

CVaR is also known as the Average Value-at-Risk, Expected Shortfall, and Tail Conditional Expectation, in various communities within computational finance, where the random variable usually models the loss associated with an asset or a collection of assets.

For comparison purposes, we introduce two more types of thresholds in this study. The first one is called an “ground-truth” threshold (GT), which is obtained by minimizing a mismatch between thresholded residuals and the ground-truth foreground mask provided in cdnet 2014 dataset. A “hard” threshold (HT) is a hard-coded value of 25. (Notice that a “hard” threshold is quite common in state-of-the-art implementations, cf. [42].) For the sample frame in Figure 6,

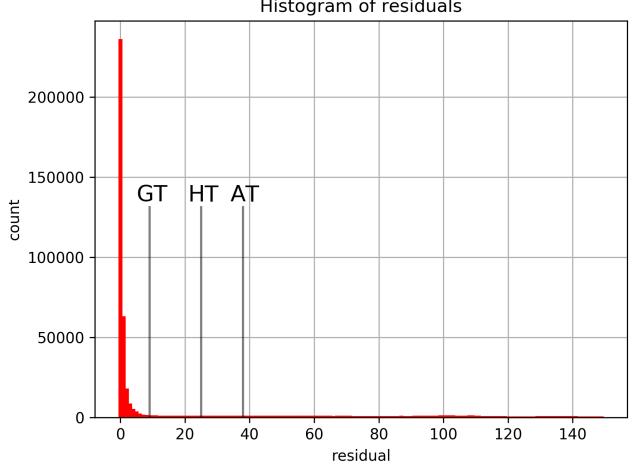


Fig. 4. An example of the histogram of residuals. The histogram was truncated from the original 255 size to allow for some clarity of presentation. Three threshold values read as follows: GT – “ground-truth” threshold obtained by minimizing a mismatch between binarized image of residuals and the ground-truth foreground mask provided in cdnet 2014; HT – “hard” threshold, here 25; AT – “automatic” threshold obtained by Algorithm 2.

our automatic threshold is $\text{VaR}_\alpha^{(at)} = 38$ and the ground-truth threshold is $\text{VaR}_\alpha^{(gt)} = 9$. The corresponding parameters are $\alpha_{gt} = 0.835$ and $\alpha_{at} = 0.876$ respectively. Loosely speaking, this means the following: Had we selected the above VaRs as thresholds, a pixel with the residual exceeding the values of 83.5% and 87.6% of the smallest residuals should be considered as an outlier (a point of a moving object in this context). For the same parameters α , the corresponding CVaR values are $\text{CVaR}_\alpha^{(gt)} = 70.1$ and $\text{CVaR}_\alpha^{(at)} = 85.8$ respectively. The big difference between VaR and CVaR values (for the same α) can be explained by slow decay of the residual distribution, see Figure 5, which contains not only the regular background samples, but also the outliers belonging to moving objects.

Figure 6 shows the result of thresholding by either “ground-truth” or by our “automatic” threshold. Clearly, the ground-truth threshold produces a solid mask, but it captures the background noise as well. In some cases, it could be not trivial to eliminate that noise. On the other hand, the “automatic” threshold gives less amount of clutter in background, but more “holes” in moving objects. Note, we do not use any mask post-processing in this study.

Algorithm 2 summarizes all the steps. It returns maximum of the four values: (1) hard threshold T_h ; (2) value exceeding $2/3$ of the smallest residuals; (3) right-hand side margin of the smallest interval that contains 50% of histogram area; (4) automatic VaR threshold as described above.

Finally, note that assuming the cardinality of the finite discrete range for the threshold values is a constant, the histogram of thresholds could be built in linear time, that is, the computational complexity is linear in the number of central points. In computer-vision applications, residuals often take integral values between 0 and 255 inclusive, and the same is true for the thresholds, so the assumption is easily satisfied.

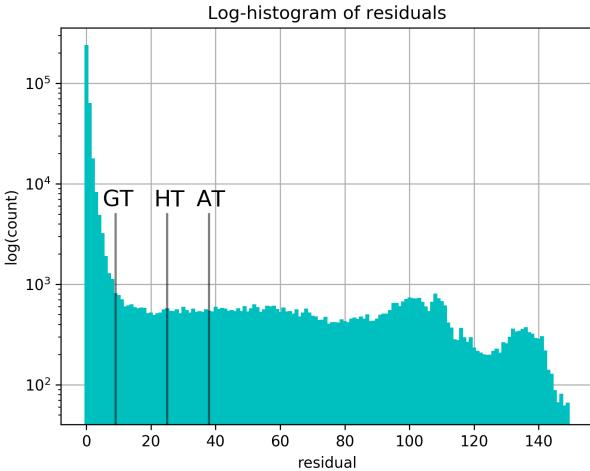


Fig. 5. An example of the histogram of residuals. The histogram was truncated from the original 255 size to allow for some clarity of presentation. Also, the vertical axes was *log*-transformed making tail values better visible. Three threshold values read as follows: GT – “ground-truth” threshold obtained by minimizing a mismatch between binarized image of residuals and the ground truth foreground mask provided in cdnet 2014; HT – “hard” threshold, here 25; AT – “automatic” threshold obtained by Algorithm 2.

Algorithm 2 Computation of Value-at-Risk threshold

```

1: Input:
    • image of integral residuals  $r_i \in [0, 255]$ ,
    • rate of noisy patterns in Figure 2,  $R = 0.0025$ ,
    • hard threshold  $T_h = 25$ .
2: Output: threshold  $t$ .
3: Initialise threshold  $t$  to a lower bound, which is just above the values of  $2/3$  of the smallest residuals.
4: Build the histogram of thresholds  $H_T(t)$  from  $r_i$ .
5: Compute the smallest interval that contains  $50\%$  of the histogram area, and get the right margin of this interval  $m_{right}$ .
6: Adjust the initial threshold:  $t \leftarrow \max(t, T_h, m_{right})$ .
7: for  $t \leq 255$  do
8:   Explore the range  $x \in [t-5 \dots t+5]$ .
9:   if all  $H_T(x) \leq R \cdot (\text{number of pixels})$  then
10:    break
11:   else
12:     $t \leftarrow t + 1$ 
13:   end if
14: end for
15: return  $t$ .
```

A. Extensions

Intuitively, it is clear that we are interested in the study of the tail of the random variable, whose histogram of thresholds we have constructed. That is: The region around the mode of the histogram (approximately 50% of its area) mostly contains noise. We could search for the optimum threshold from the right-hand side of the region around the mode of the histogram of thresholds $H_T(t)$, until the rate of noisy patterns in Figure 2 falls below certain value.

One could also require the threshold to be above $2/3$ of residual values in the image. This requirement is typically well-matched with expected amount of moving points in surveillance video.

One can consider other notions of isolation, which may have further benefits. In particular, we have considered 5×5 vicinity of each points and account only for those configurations where internal 3×3 pattern from Figure 2 is completely isolated from the pixels on outer border of 5×5 neighbourhood with values exceeding a threshold. In other words, one-pixel thin lines should not influence threshold estimation. However, the implementation would be more complicated in this case. We hence consider only the 3×3 patches.

IV. EXPERIMENTAL EVALUATION

To demonstrate our approach, we have implemented the autoencoder in Python 3 using PyTorch, the deep-learning engine. For evaluation and comparison against other methods, we present results on changedetection.net (cdnet 2014), a well-established benchmark of [43]. Although cdnet-2014 is not the benchmark where the convolutional autoencoder would work best, considering that the video sequences in cdnet 2014 are quite short (1,200 to 9,000 frames), we perform three passes over each sequence, in order to circumvent the shortage of data. Evaluation and computation of the foreground mask is performed only during the final pass. Note that we do not use any labelling information in the training at all, in sharp contrast to some other authors testing deep-learning approaches on cdnet-2014. For scoring, we utilize cdnet 2014 evaluation software on categories: “badWeather”, “dynamicBackground”, “cameraJitter”, “baseline”, “nightVideos”, “shadow”. Each category contains several videos. Since our autoencoder expects a standard image size of 576×704 , we resize input image, if necessary, and the computed foreground mask is resized back to the original size. This procedure affects both quality and performance, but seems unavoidable because network architecture is determined by the image size.

For comparison, we have chosen several methods from LRSlibrary, an excellent toolbox developed by A. Sobral and co-authors [42], [44]. Considering that our approach can be seen as a low-rank approximation method, we have focussed on five well-performing methods considering the “low-rank and sparse” model for background modeling and subtraction in videos. The five methods were: LRR_FastLADMAP [8], MC_GROUSE [9], RPCA_FPCP [10], ST_GRASTA [11], TTD_3WD [12], when considering both statistical performance and the run-time per frame. We also used the recent and state-of-the-art algorithm OM_OGMF, proposed in [28], [45], as implemented in Matlab by the authors. OM_OGMF was identified as a top performer in our experiments. The only caveat to keep in mind is that OM_OGMF algorithm outputs black foreground mask after about 2,500 frames. As a workaround, we discard the foreground masks with all black pixels from the scoring process. Considering neither of the methods used for comparison requires a fixed frame size, we do not rescale the frames.

The data matrix has been built from 50 most recent frames, which is the default setting used across both LRSlibrary

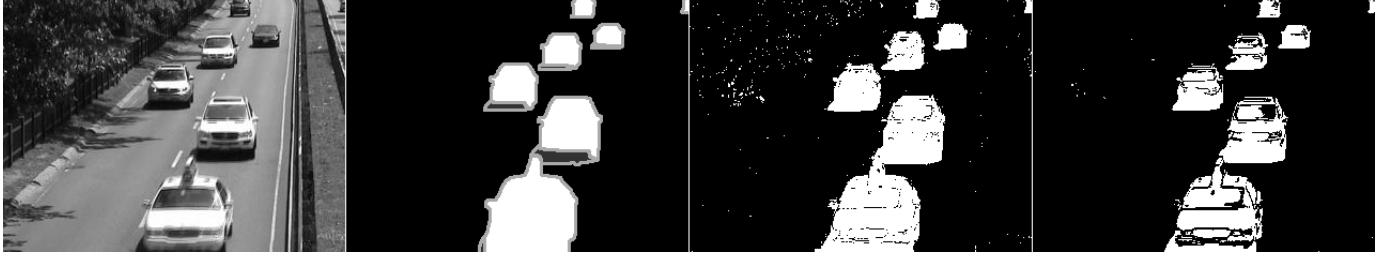


Fig. 6. Sample frame from the video-sequence *highway* (left), corresponding ground-truth foreground mask provided by cdnet 2014 (middle-left), result of binarization of residual image by the “ground-truth” threshold obtained by minimizing a mismatch between binarized image and the ground-truth (middle-right), result of binarization of residual image by the “automatic” threshold obtained by Algorithm 2 (right).

<i>Method</i>	<i>Recall</i>	<i>Specificity</i>	<i>FPR</i>	<i>FNR</i>	<i>Precision</i>	<i>F1</i>	Run-time [s/frame]
LRR_FastLADMAP [8]	0.74694	0.93980	0.06020	0.25306	0.28039	0.36194	4.611
MC_GROUSE [9]	0.65640	0.89692	0.10308	0.34360	0.25425	0.31495	10.621
OMoGMF [28], [45]	0.89943	0.98289	0.01711	0.10057	0.62033	0.72611	0.123
RPCA_FPCP [10]	0.73848	0.94733	0.05267	0.26152	0.29994	0.37900	0.504
ST_GRASTA [11]	0.45340	0.98205	0.01795	0.54660	0.44009	0.42367	3.266
TTD_3WD [12]	0.61103	0.97117	0.02883	0.38897	0.35557	0.40297	10.343
Autoencoder, 5×5 min. thr.	0.65676	0.99360	0.00640	0.34324	0.77756	0.70593	0.57

TABLE I

PERFORMANCE RESULTS ON “BASELINE” VIDEO-CATEGORY FROM [HTTP://CHANGEDETECTION.NET](http://CHANGEDETECTION.NET). THE EXECUTION TIME IN SECONDS PER FRAME IS GIVEN FOR “HIGHWAY” VIDEO-SEQUENCE WITH 240×320 IMAGES. THE LAST LINE REFLECTS ONLY FORWARD/BACKWARD SINGLE ITERATION TIME.

and OMoGMF. Upon arrival of a new image, it was inserted at the end of the queue keeping the time-ordering, which some algorithms might be sensitive to.

A. Statistical Performance

Results for the “baseline” category, along with timing information in the last column for “Highway” video, are presented in Table I.

Table II summarizes scoring results obtained on 6 video-categories for all the methods including OMoGMF, which was identified as the best subspace method in this study, and our convolutional autoencoder. Note, there are two sub-tables for autoencoder results. In the first case (“min. threshold”) we derive the optimum threshold from a distribution of minimum residuals across a batch of reconstructed backgrounds, but *without* looking at points’ neighbourhoods. In the second case (“5×5 min. threshold”), threshold was computed with the extensions described in the previous section, formula (3), and that leads to an improvement in the overall *F*-measure, especially for dynamic-background category.

Figures 7 and 8 present selected results obtained by Autoencoder and OMoGMF methods. Here we provide some typical cases, where advantages and disadvantages of both approaches can be clearly seen. A few observations deserve attention. First, OMoGMF performs really well in general. It integrates a flexible Gaussian Mixture Model (GMM) and produces a solid motion mask, but it is less resistant to noise and non-stationary background than our autoencoder. Second, our autoencoder yields a good background estimation and copes better with dynamic background than many other methods. While a single choice of a threshold is not flexible enough, as it produces “holes” in the motion mask and worse scores, our VaR-based method seems more robust than the alternatives.

Third, foreground sometimes “leaks” into background, as for example in the “canoe” result in Figure 8. Partially, this can be explained by videos being too short for proper training of autoencoder, which tends to memorize images — the known problem. Also note that autoencoder operates on grayscale images, unlike other methods. By taking advantage of colour images the detectability of moving objects can be improved, whereas realistic scenarios include night-time videos, where colour information is not available anyway.

B. Runtime

All the methods used for comparison, except our autoencoder, are implemented in Matlab. It took several weeks to process selected videos on Intel Core i7-4800MQ, 4-core CPU, 16 Gb, 2.70 GHz workstation powered by RedHat 7.6/64 Linux and Matlab 2018a. To make use of the general-purpose graphics processing unit (GPGPU), we have utilised a different machine to run the autoencoder. This machine was equipped with Intel Xeon E5-2699 CPU at 2.20 GHz and Tesla K40c GPGPU with 12 GB of on-board memory, and ran by RedHat Linux 7.5. With the GPGPU, it took about 2 days to run on the benchmark.

Since we are using different hardware for our autoencoder and OMoGMF, it is not straightforward to compare their computational speed. In general, our autoencoder utilizes a single CPU core and the GPGPU specified above. OMoGMF utilizes 4 CPU cores and takes about 0.123 seconds per a 240×320 frame (video-sequence “highway”). In the case of our autoencoder, the DIS algorithm followed by forward/backward steps takes about 0.57 sec. per a 576×704 frame. Thresholding and storing on hard-disk (for subsequent scoring) take about 1 second. Our thresholding procedure has very simple code and can be easily improved to real-time performance. The

actual bottleneck resides in neural network implementation. Considering 576×704 images and at least linear dependency of processing time on the number of pixels, the processing times of OM_OGMF and the main part of our autoencoder should be comparable.

V. A SUMMARY AND DISCUSSION

We have presented an algorithm for tracking of time-varying low-rank background models of time-varying matrices, using a continuously trained and applied convolutional autoencoder. Our approach displays solid performance overall, and seems comparable to the best subspace methods. Three features may be worth highlighting. First, incremental training makes the network adaptive to gradual scene changes, which always happen in reality. Second, no labelled data is needed, in contrast to typical deep-learning approaches. Instead, we down-weighting the moving points using a rough estimation of the foreground mask. The training is primarily driven by background points and robust to outliers. Third, we compute foreground mask of moving objects by considering the spatial neighbourhood of each pixel and VaR-based thresholding.

Low-rank methods have made remarkable progress in recent years, but still demonstrate certain limitations. As it turns out, all the methods considered in this study have difficulties in producing a convincing foreground motion mask in the case of fast-varying background. Clearly, there are good statistical and complexity-theoretic reasons [46], [47] for any method to have difficulties in fast-changing environments, but there may be a scope for improvement. Improving upon thresholding technique can alleviate the problem of a poor foreground mask, to some extent.

Originally, we designed the algorithm to handle video streams, collected from a network of CCTV cameras in Dublin, Ireland, which posed a serious challenge to any algorithm we tested: camera jitter, compression artefacts, poor image quality, adverse weather condition, night videos and so on. Usually, it takes approximately 20 minutes (30,000 video-frames) for autoencoder to build up a good background model. However, the standard video sequences, adopted by the community for benchmarking, are typically very short – few thousands frames. This rises a question about the scoring process, particularly when algorithms based on deep learning architecture are involved.

It seems that the next generation of benchmarks should be designed. This can be, for example, a collection of compressed, one-day long videos with few thousand check-point images evenly scattered across the sequence and manually labelled. When video-processing reaches the next check-point frame, the scoring procedure is applied to collect and update a performance statistics. That would offer a more realistic comparison protocol.

An important avenue for further efforts, as can be seen in the last column of Table I, is the speed-up of the autoencoder. This is a principal limitation of any deep-learning architecture, overall. Although hardware advances lessen the cost and increases the performance of GPGPUs at remarkable pace, having a GPGPU per camera may still be too expensive for

many applications. One possible solution is to run forward steps often and backward (training) ones rarely. This will reduce the rate of adaptation, but also decrease the run-time.

REFERENCES

- [1] M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, and Y. Manolopoulos, “Continuous monitoring of distance-based outliers over data streams,” in *2011 IEEE 27th International Conference on Data Engineering*, pp. 135–146, April 2011.
- [2] L. Cao, Y. Yan, C. Kuhlman, Q. Wang, E. A. Rundensteiner, and M. Eltabakh, “Multi-tactic distance-based outlier detection,” in *IEEE 33rd Int. Conf. on Data Engineering (ICDE)*, pp. 959–970, 2017.
- [3] T. Bouwmans and B. Garcia-Garcia, “Background subtraction in real applications: Challenges, current models and future directions,” *CoRR*, vol. abs/1901.03577, 2019.
- [4] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” *Proc. 1999 IEEE Computer Society Conf. on Comp. Vision and Pattern Recognition*, vol. 2, pp. 246–252 Vol. 2, 1999.
- [5] Z. Zivkovic and F. van der Heijden, “Efficient adaptive density estimation per image pixel for the task of background subtraction,” *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773 – 780, 2006.
- [6] O. Barnich and M. Van Droogenbroeck, “Vibe: A powerful random technique to estimate the background in video sequences,” in *2009 IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 945–948, 2009.
- [7] P. St-Charles, G. Bilodeau, and R. Bergevin, “Subsense: A universal change detection method with local adaptive sensitivity,” *IEEE Transactions on Image Processing*, vol. 24, pp. 359–373, Jan 2015.
- [8] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low-rank representation,” in *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds.), pp. 612–620, Curran Associates, Inc., 2011.
- [9] L. Balzano and S. J. Wright, “On GROUSE and incremental SVD,” in *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 1–4, Dec 2013.
- [10] P. Rodriguez and B. Wohlberg, “Fast principal component pursuit via alternating minimization,” in *2013 IEEE International Conference on Image Processing*, pp. 69–73, Sep. 2013.
- [11] J. He, L. Balzano, and J. C. S. Lui, “Online Robust Subspace Tracking from Partial Information,” *arXiv e-prints*, p. arXiv:1109.3827, Sep 2011.
- [12] O. Oreifej, X. Li, and M. Shah, “Simultaneous video stabilization and moving object detection in turbulence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 450–462, Feb 2013.
- [13] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayananurthy, “Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery,” *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 32–55, 2018.
- [14] Y. Chen and Y. Chi, “Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization,” *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14–31, 2018.
- [15] G. Lerman and T. Maunu, “An overview of robust subspace recovery,” *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1380–1410, 2018.
- [16] S. Ma and N. S. Aybat, “Efficient optimization algorithms for robust principal component analysis and its variants,” *Proceedings of the IEEE*, vol. 106, pp. 1411–1426, Aug 2018.
- [17] T. Bouwmans, S. Javed, H. Zhang, Z. Lin, and R. Otazo, “On the applications of robust pca in image and video processing,” *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1427–1457, 2018.
- [18] L. Balzano, Y. Chi, and Y. M. Lu, “Streaming pca and subspace tracking: The missing data case,” *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1293–1310, 2018.
- [19] X. Zhang, W. Dou, Q. He, R. Zhou, C. Leckie, R. Kotagiri, and Z. Salcic, “Lshiforest: A generic framework for fast tree isolation based ensemble anomaly analysis,” in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 983–994, April 2017.
- [20] L. A. Lim and H. Y. Keles, “Learning multi-scale features for foreground segmentation,” *CoRR*, vol. abs/1808.01477, 2018.
- [21] M. Babaee, D. T. Dinh, and G. Rigoll, “A deep convolutional neural network for background subtraction,” *CoRR*, vol. abs/1702.01731, 2017.
- [22] M. Braham and M. Van Droogenbroeck, “Deep background subtraction with scene-specific convolutional neural networks,” in *Int. Conf. on Systems, Signals and Image Proc. (IWSSIP)*, pp. 1–4, May 2016.

- [23] P. Xu, M. Ye, X. Li, Q. Liu, Y. Yang, and J. Ding, "Dynamic background learning through deep auto-encoder networks," in *Proceedings of the 22Nd ACM International Conference on Multimedia, MM '14*, (NY, USA), pp. 107–116, 2014.
- [24] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in neural information processing systems*, pp. 801–808, 2007.
- [25] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674, ACM, 2017.
- [26] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations*, 2018.
- [27] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *CoRR*, vol. abs/1811.05255, 2018.
- [28] H. Yong, D. Meng, W. Zuo, and L. Zhang, "Robust online matrix factorization for dynamic background subtraction," *IEEE Trans. on Pat. Analysis and Mach. Intelligence*, vol. 40, pp. 1726–1740, July 2018.
- [29] J. Xu and J. Marecek, "Parameter estimation in gaussian mixture models with malicious noise, without balanced mixing coefficients," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 446–453, Oct 2018.
- [30] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [31] T. Kroeger, R. Timofte, D. Dai, and L. J. V. Gool, "Fast optical flow using dense inverse search," *CoRR*, vol. abs/1603.03590, 2016.
- [32] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, pp. 146–168, Jan 2004.
- [33] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *J. Graphics Tools*, vol. 12, pp. 13–21, 01 2007.
- [34] W. Tao, H. Jin, Y. Zhang, L. Liu, and D. Wang, "Image thresholding using graph cuts," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, pp. 1181 – 1195, 10 2008.
- [35] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 62–66, Jan 1979.
- [36] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273–285, 1985.
- [37] P. Rosin, "Thresholding for change detection," in *Proc. ICCV*, pp. 274–279, 02 1998.
- [38] C. Chang, Y. Du, J. Wang, S. Guo, and P. D. Thouin, "Survey and comparative analysis of entropy and relative entropy thresholding techniques," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 153, pp. 837–850, Dec 2006.
- [39] A. Malistov, "Estimation of background noise in traffic conditions and selection of a threshold for selecting mobile objects," *Actual problems of modern science*, vol. 4, 2014. In Russian. In addition, see the abstract of PhD Thesis, pp. 12-13 in http://www.malistov.ru/docs/dissertation/abstract_malistov.pdf.
- [40] C. Conaire, N. O'Connor, E. Cooke, and A. Smeaton, "Detection thresholding using mutual information," *VISAPP – International Conference on Computer Vision Theory and Applications*, Feb 2006.
- [41] X. Fu-song, "Survey over image thresholding techniques based on entropy," in *2014 International Conference on Information Science, Electronics and Electrical Engineering*, vol. 2, pp. 1330–1334, April 2014.
- [42] A. Sobral, T. Bouwmans, and E.-h. Zahzah, "Lrslibrary: Low-rank and sparse tools for background modeling and subtraction in videos," in *Robust Low-Rank and Sparse Matrix Decomposition: Appl. in Image and Video Proc.*, CRC Press, Taylor and Francis Group., 2015.
- [43] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changefetection.net: A new change detection benchmark dataset," in *in Proc. IEEE Workshop on Change Detection (CDW-2012) at CVPR-2012, Providence, RI*, Jun 2012.
- [44] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-h. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *CoRR*, vol. abs/1511.01245, 2015.
- [45] D. Meng and F. D. L. Torre, "Robust matrix factorization with unknown noise," in *Proc. of the 2013 IEEE Int. Conf. on Computer Vision, ICCV'13*, pp. 1337–1344, IEEE Computer Society, 2013.
- [46] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.
- [47] J. C. Gittins, K. D. Glazebrook, R. Weber, and R. Weber, *Multi-armed bandit allocation indices*, vol. 25. Wiley Online Library, 1989.

<i>Video</i>	<i>Recall</i>	<i>Specificity</i>	<i>FPR</i>	<i>FNR</i>	<i>Precision</i>	<i>F1</i>
LRR_FastLADMAP [8]:						
badWeather	0.82941	0.82644	0.17356	0.17059	0.09239	0.15580
baseline	0.74694	0.93980	0.06020	0.25306	0.28039	0.36194
cameraJitter	0.75423	0.83766	0.16234	0.24577	0.18119	0.28715
dynamicBackground	0.69953	0.79853	0.20147	0.30047	0.03828	0.06968
nightVideos	0.80056	0.84435	0.15565	0.19944	0.11062	0.18503
shadow	0.72950	0.88521	0.11479	0.27050	0.23030	0.32793
Overall	0.76003	0.85533	0.14467	0.23997	0.15553	0.23125
ST_GRASTA [11]:						
badWeather	0.26555	0.98971	0.01029	0.73445	0.45526	0.30498
baseline	0.45340	0.98205	0.01795	0.54660	0.44009	0.42367
cameraJitter	0.51138	0.91313	0.08687	0.48862	0.23995	0.31572
dynamicBackground	0.41411	0.94755	0.05245	0.58589	0.08732	0.13736
nightVideos	0.42488	0.97224	0.02776	0.57512	0.24957	0.28154
shadow	0.44317	0.96681	0.03319	0.55683	0.42604	0.41515
Overall	0.41875	0.96192	0.03808	0.58125	0.31637	0.31307
RPCA_FPCP [10]:						
badWeather	0.82546	0.84424	0.15576	0.17454	0.09950	0.16687
baseline	0.73848	0.94733	0.05267	0.26152	0.29994	0.37900
cameraJitter	0.74452	0.84143	0.15857	0.25548	0.18436	0.29024
dynamicBackground	0.69491	0.80688	0.19312	0.30509	0.03928	0.07134
nightVideos	0.79284	0.85751	0.14249	0.20716	0.11797	0.19497
shadow	0.72132	0.90454	0.09546	0.27868	0.26474	0.36814
Overall	0.75292	0.86699	0.13301	0.24708	0.16763	0.24509
OMoGMF [28], [45]:						
badWeather	0.86871	0.98939	0.01061	0.13129	0.57917	0.67214
baseline	0.89943	0.98289	0.01711	0.10057	0.62033	0.72611
cameraJitter	0.85954	0.90739	0.09261	0.14046	0.30566	0.44235
dynamicBackground	0.87655	0.86383	0.13617	0.12345	0.08601	0.15012
nightVideos	0.75607	0.92372	0.07628	0.24393	0.23252	0.31336
shadow	0.55771	0.80276	0.03057	0.27562	0.40539	0.37449
Overall	0.80300	0.91166	0.06056	0.16922	0.37151	0.44643
Autoencoder, min. threshold:						
badWeather	0.83978	0.97169	0.02831	0.16022	0.46847	0.57446
baseline	0.72216	0.98873	0.01127	0.27784	0.62647	0.63878
cameraJitter	0.74510	0.92411	0.07589	0.25490	0.40775	0.50148
dynamicBackground	0.81096	0.86072	0.13928	0.18904	0.08430	0.14604
nightVideos	0.66925	0.95927	0.04073	0.33075	0.23837	0.34159
shadow	0.71458	0.98464	0.01536	0.28542	0.69253	0.68589
Overall	0.75031	0.94819	0.05181	0.24969	0.41965	0.48137
Autoencoder, 5×5 min. thr.:						
badWeather	0.64712	0.99898	0.00102	0.35288	0.92511	0.75683
baseline	0.65676	0.99360	0.00640	0.34324	0.77756	0.70593
cameraJitter	0.64940	0.95970	0.04030	0.35060	0.43331	0.51465
dynamicBackground	0.55339	0.97154	0.02846	0.44661	0.24313	0.30643
nightVideos	0.53979	0.97544	0.02456	0.46021	0.33687	0.39017
shadow	0.62316	0.98857	0.01143	0.37684	0.76103	0.67776
Overall	0.61160	0.98130	0.01870	0.38840	0.57950	0.55863

TABLE II

PERFORMANCE RESULTS ON 6 VIDEO-CATEGORIES FROM [HTTP://CHANGEDETECTION.NET](http://CHANGEDETECTION.NET). FOR AUTOENCODER WE PRESENT RESULTS FOR TWO THRESHOLDING STRATEGIES IN THE LAST TWO SUB-TABLES AS DETAILED IN SECTION II.

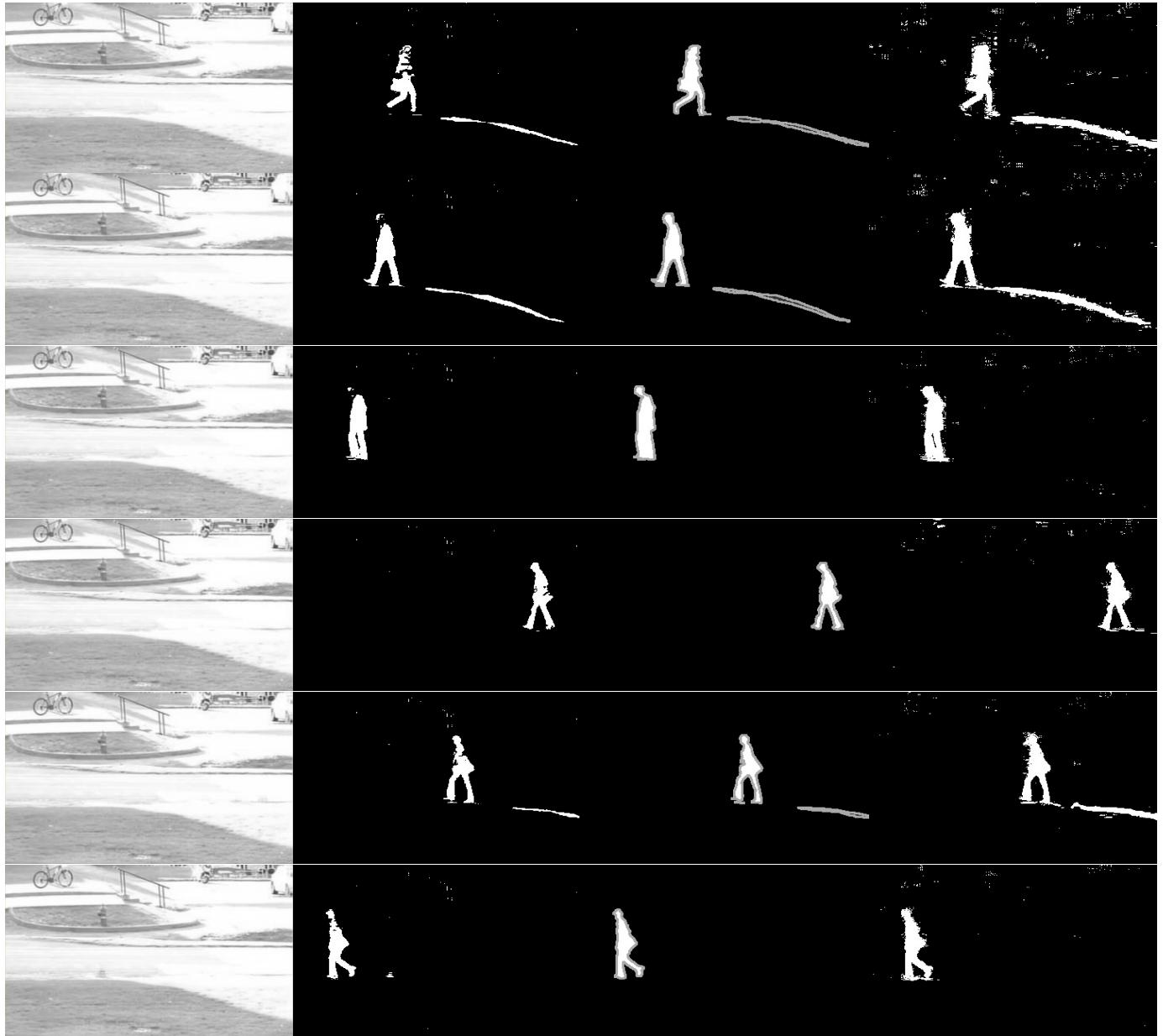


Fig. 7. “Pedestrians” video-sequence, cdnet 2014. *Left to right:* reconstructed background outputted by the Autoencoder, foreground mask by the Autoencoder, the ground-truth, and foreground mask obtained by OMOMF. Foreground mask obtained from autoencoder output does not undergo any post-processing.

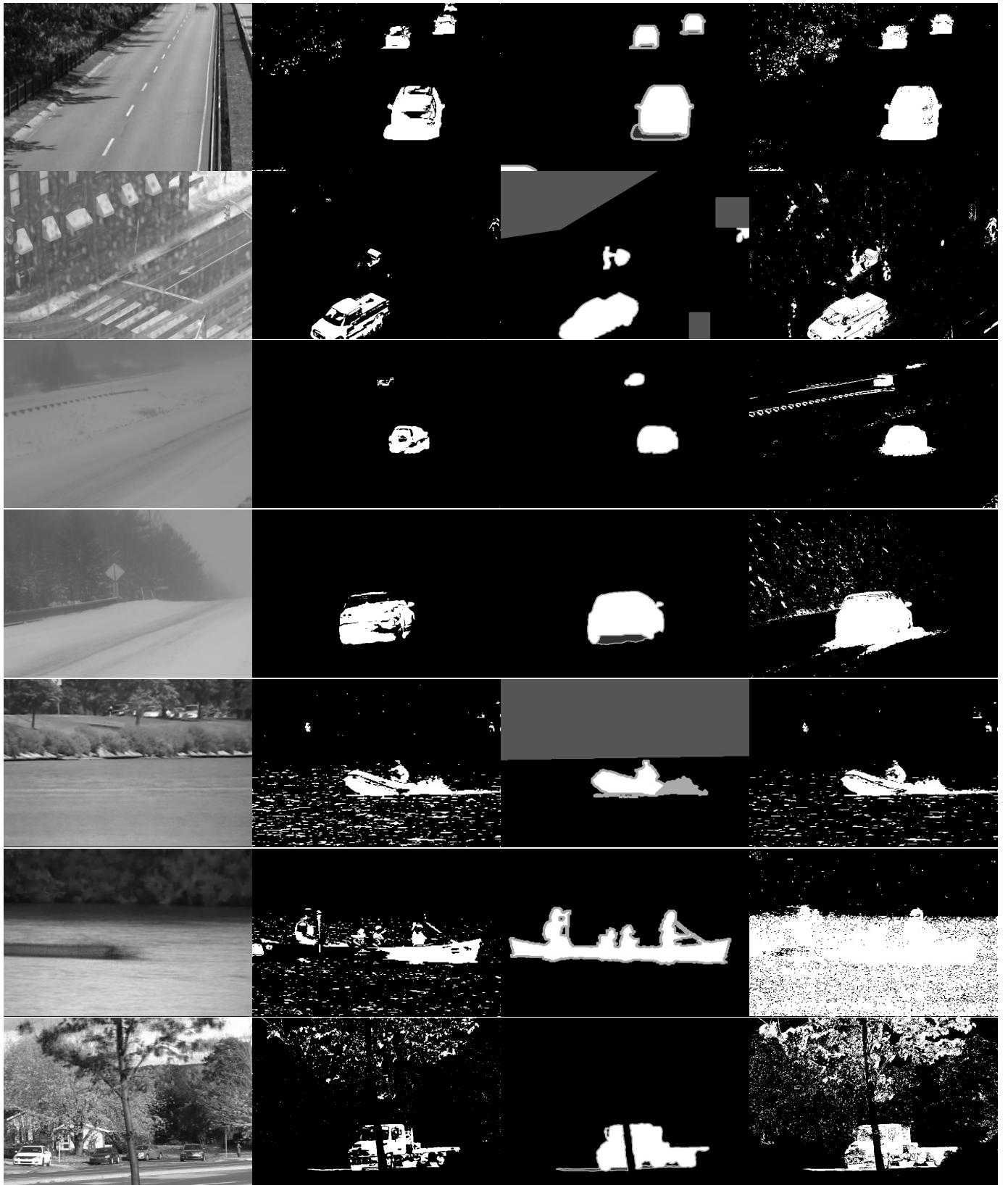


Fig. 8. Selected background subtraction results. *Left to right*: reconstructed background outputted by the Autoencoder, foreground mask by the Autoencoder, the ground-truth, and foreground mask obtained by OM_oGMF. Video-sequences, cdnet 2014, *top to bottom*: “highway”, “wetsnow”, “blizzard”, “snowfall”, “boats”, “canoe”, “fall”. Foreground mask obtained from autoencoder output does not undergo any post-processing.

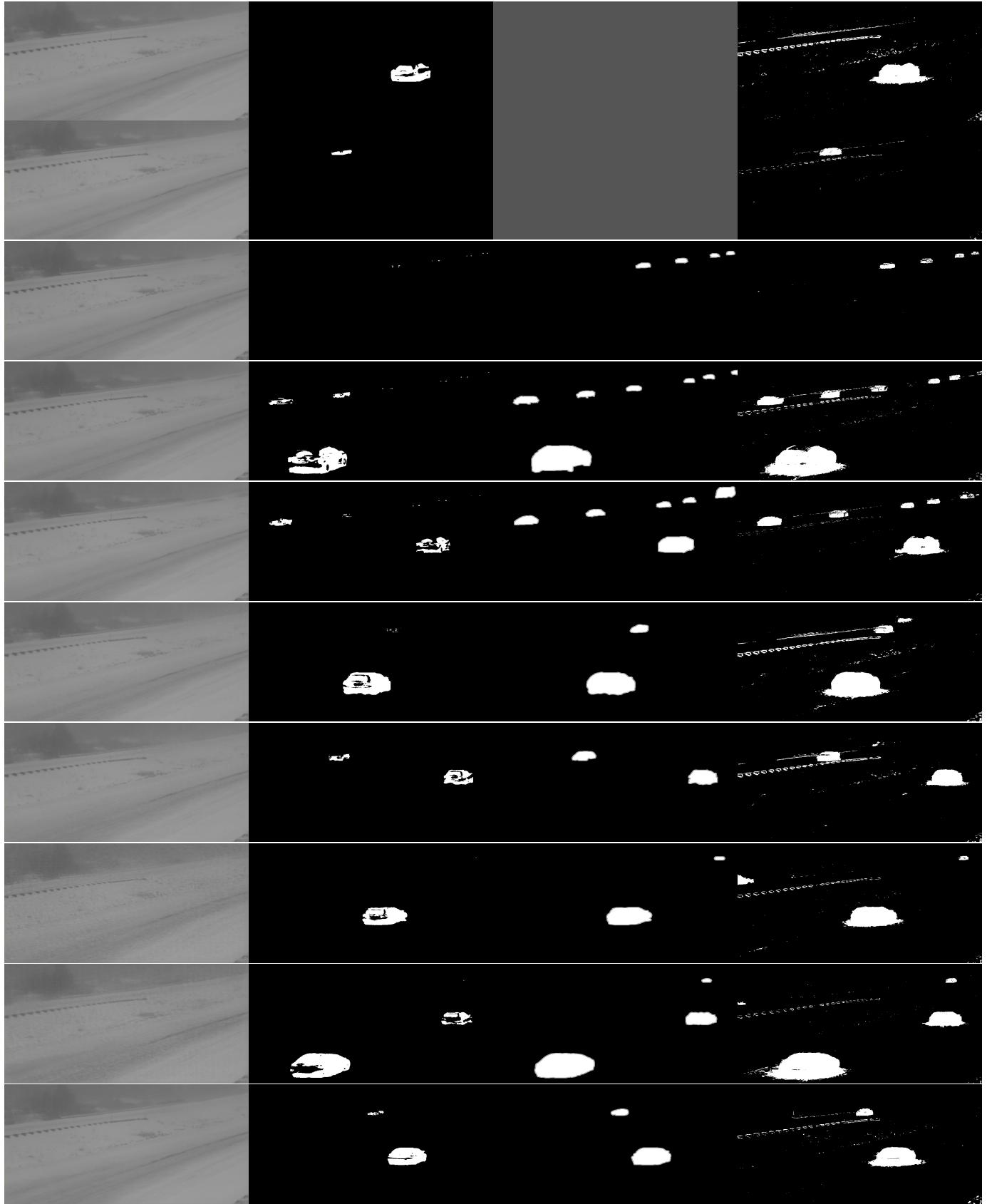


Fig. 9. “Blizzard” video-sequence, cdnet 2014. *Left to right*: reconstructed background outputted by the Autoencoder, foreground mask by the Autoencoder, the ground-truth, and foreground mask obtained by OMogmf. OMogmf is good in detecting small objects, but noise is also often acquired. Foreground mask obtained from autoencoder output does not undergo any post-processing.

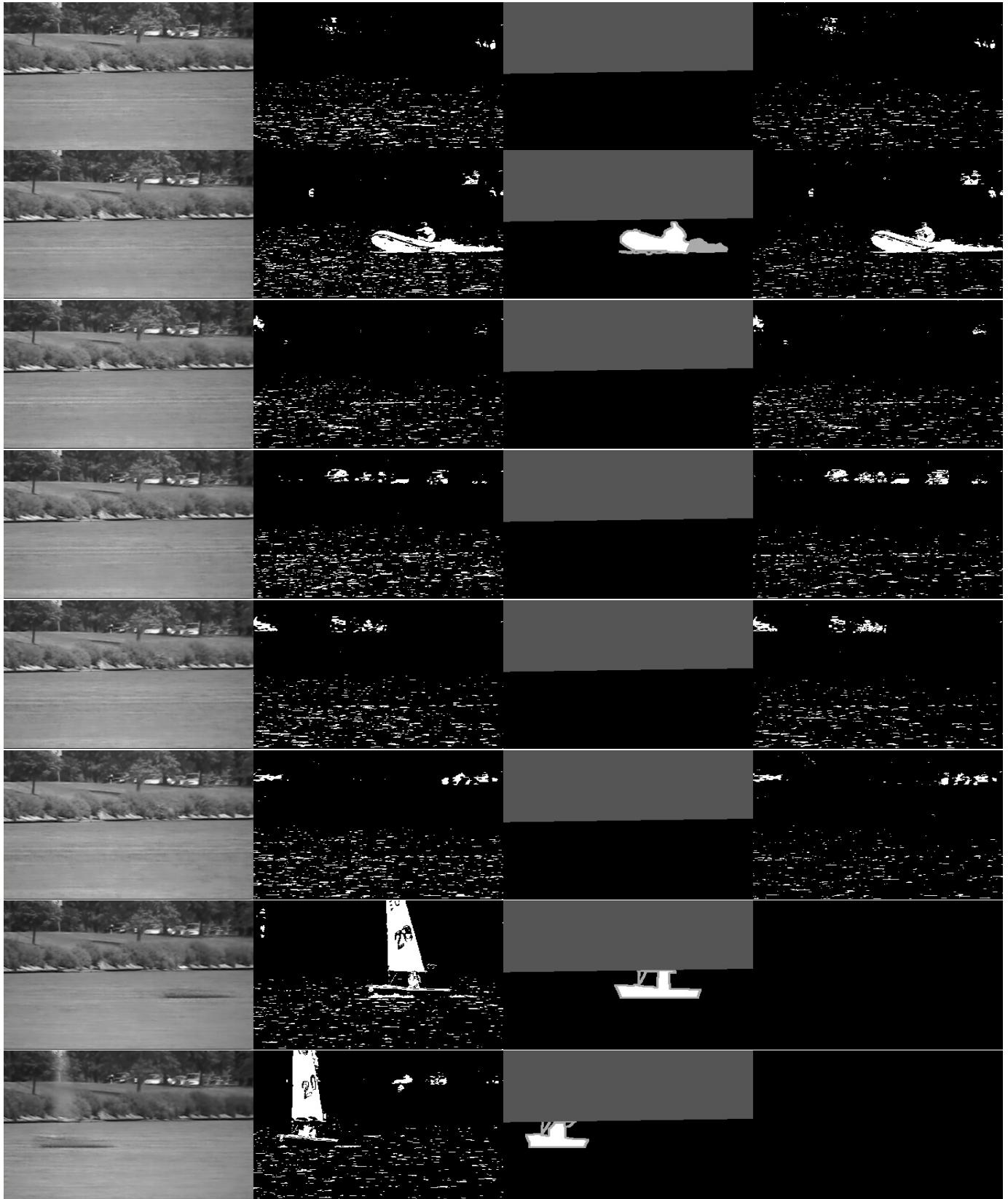


Fig. 10. “Boats” video-sequence, cdnet 2014. *Left to right*: reconstructed background outputted by the Autoencoder, foreground mask by the Autoencoder, the ground-truth, and foreground mask obtained by OMogmf. Non-stationary background is a particularly difficult case for modelling by any method mentioned in this study. Note that there is an issue in OMogmf code that results in black foreground mask after approximately 2,500 frames. Foreground mask obtained from autoencoder output does not undergo any post-processing. Some “leakage” of slowly moving objects into reconstructed background can be observed in the last two rows.

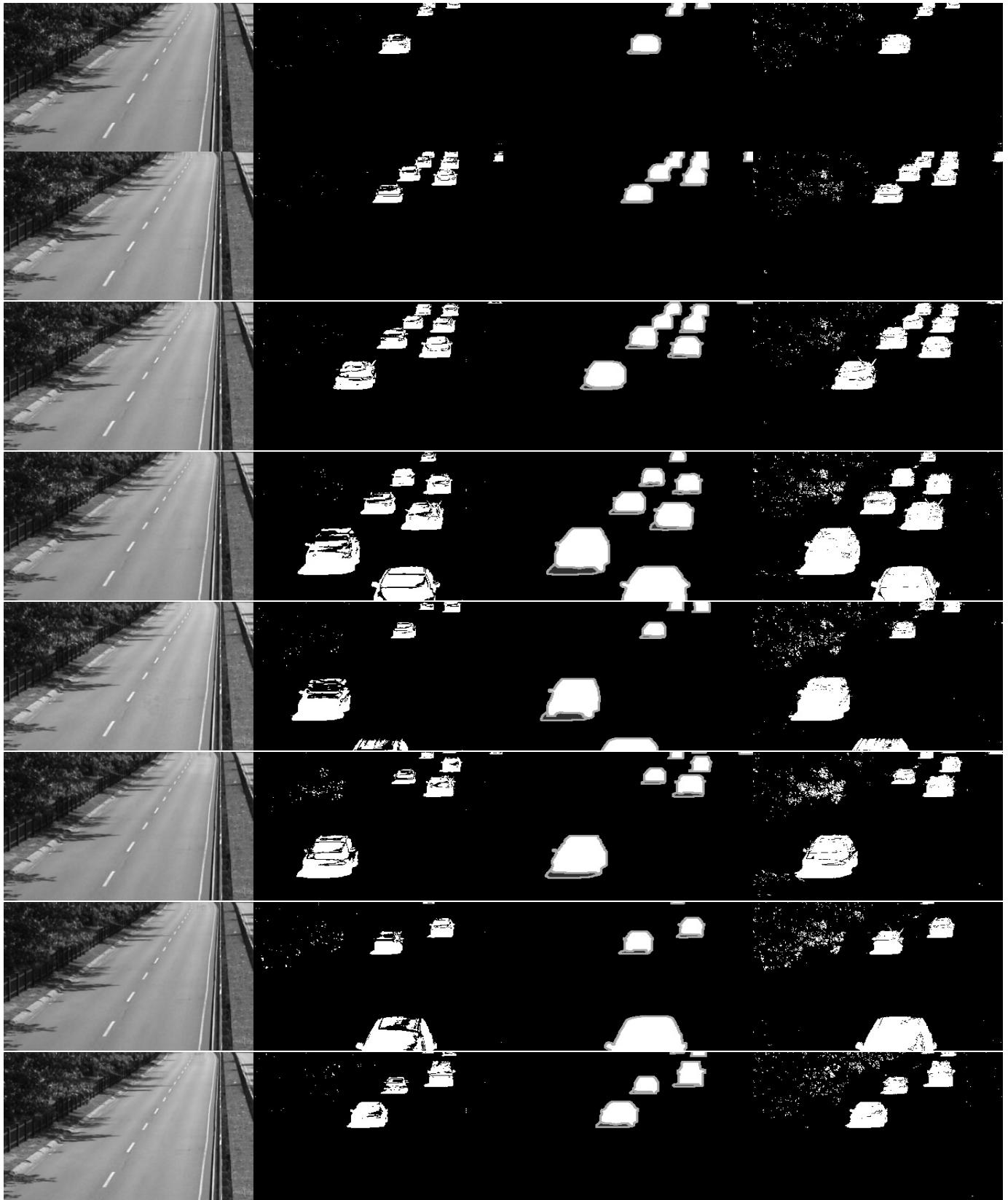


Fig. 11. “Highway” video-sequence, cdnet 2014. *Left to right:* reconstructed background outputted by the Autoencoder, foreground mask by the Autoencoder, the ground-truth, and foreground mask obtained by OMoGMF. Notice that the autoencoder is more resistant against dynamic background, at the expense of less solid motion mask. Foreground mask obtained from autoencoder output does not undergo any post-processing.

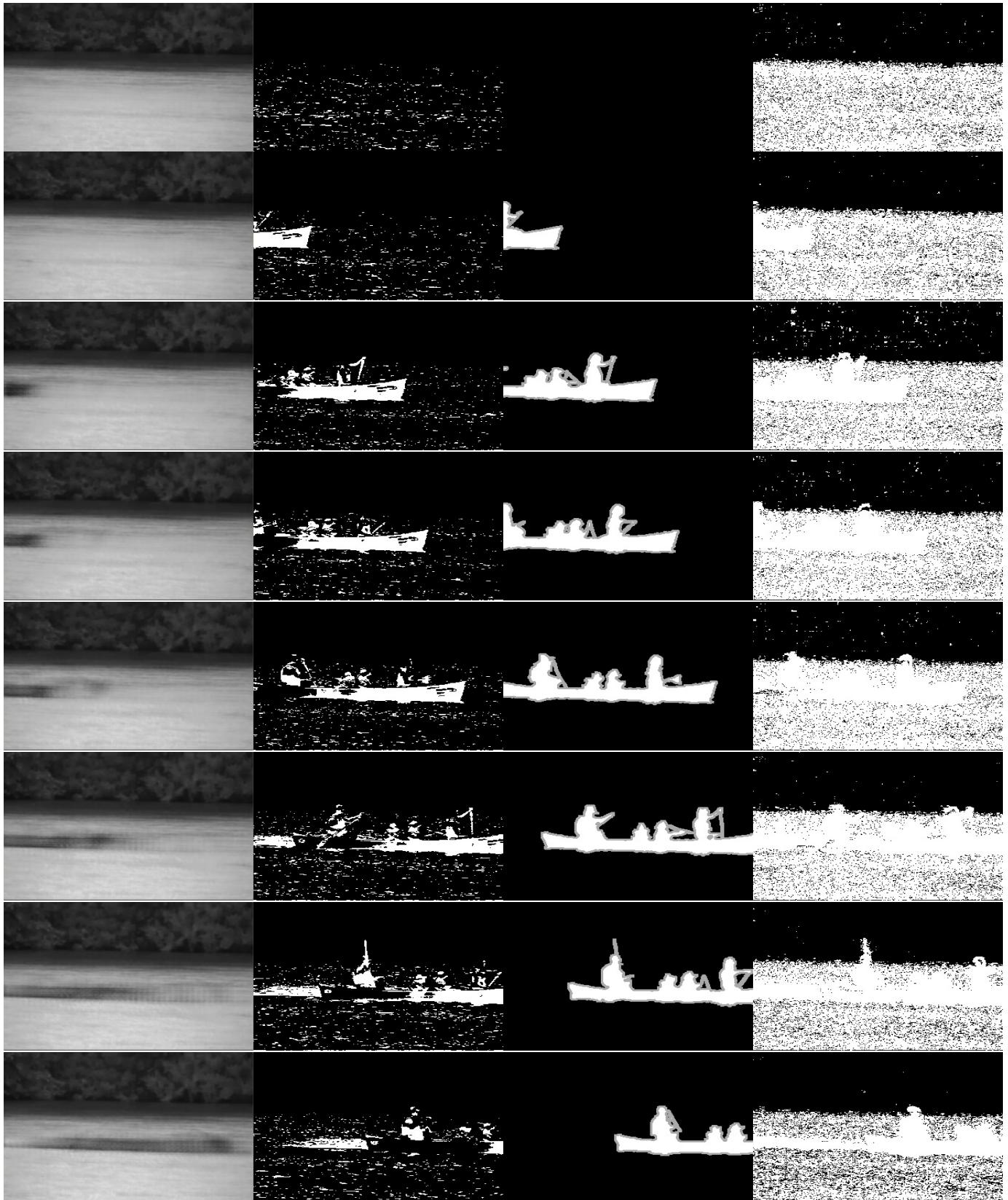


Fig. 12. “Canoe” video-sequence, cdnet 2014. *Left to right:* reconstructed background outputted by the Autoencoder, foreground mask by the Autoencoder, the ground-truth, and foreground mask obtained by OMogMF. This is another difficult case of non-stationary background. Notice the “leakage” of slowly moving object into reconstructed background. Foreground mask obtained from autoencoder output does not undergo any post-processing.