Deep Fusion of Appearance and Frame Differencing for Motion Segmentation

Marc Ellenfeld[1,3], Sebastian Moosbauer[1], Ruben Cardenes[2], Ulrich Klauck[3,4], Michael Teutsch[1]

[1] Hensoldt Optronics GmbH, Germany [2] Hensoldt Analytics GmbH, Germany
{sebastian.moosbauer, ruben.cardenes, michael.teutsch}@hensoldt.net
[3] Aalen University of Applied Sciences, Germany [4] University of the Western Cape, South Africa
ulrich.klauck@hs-aalen.de

Abstract

Motion segmentation is a technique to detect and localize class-agnostic motion in videos. This motion is assumed to be relative to a stationary background and usually originates from objects such as vehicles or humans. When the camera moves, too, frame differencing approaches that do not have to model the stationary background over minutes, hours, or even days are more promising compared to background subtraction methods. In this paper, we propose a Deep Convolutional Neural Network (DCNN) for multi-modal motion segmentation: the current image contributes with appearance information to distinguish between relevant and irrelevant motion and frame differencing captures the temporal information, which is the scene's motion independent of the camera motion. We fuse this information to receive an effective and efficient approach for robust motion segmentation. The effectiveness is demonstrated using the multi-spectral CDNet-2014 dataset that we re-labeled for motion segmentation. We specifically show that we can detect tiny moving objects significantly better compared to methods based on optical flow.

1. Introduction

Motion segmentation is a technique to automatically detect and localize motion in videos that is independent of the camera motion. This motion can be assumed to be class-agnostic as in some applications we cannot expect to have an object appearance at sufficiently high resolution for a reliable classification. Video surveillance is a popular example application [47]: the distance between the camera and the observed scene can be large with several kilometers and thus objects appear at a low resolution in the image. Furthermore, the environment is usually not cooperative and objects try to occlude or camouflage themselves. An example image with related Ground Truth (GT) is shown in Fig. 1. So, we consider motion segmentation as a task for the detection of motion originating from an unknown source under adverse environmental conditions in real-time and on-line without any prior knowledge. This is in contrast to most existing literature, where either the classes are known and assumed to be observable [17], or the environment is assumed to be cooperative like in automotive scenarios [49], or the entire scene and motion history is assumed to be prior known [23], or time-consuming off-line processing is performed [37], or elaborate geometric modeling is performed to extract 3D information [64]. Actually, motion segmentation in our case is more related to tasks like change detection [41] and foreground-background segmentation [5, 7]: we consider motion segmentation as a binary segmentation task to separate moving and stationary pixels in the scene [3]. Popular approaches in this field are based on frame differencing [46], background subtraction [15], or optical flow [35]. Typical challenges are precise image alignment for moving cameras to avoid false positive detections due to alignment errors or parallax effects, distinguishing between relevant motion that originates from moving objects and irrelevant motion coming from moving vegetation, water, or clouds, and detecting moving objects in a large distance with a very tiny appearance in the image.

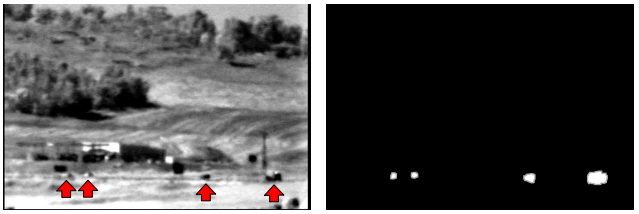

Figure 1. Example image (left) and related ground truth (right) taken from the CDNet-2014 dataset [59]. Moving objects in the scene are additionally indicated with red arrows.

In this paper, we propose a Deep Convolutional Neural Network (DCNN) for multi-modal motion segmentation: the current image contributes with appearance infor-

mation to distinguish between relevant and irrelevant motion and frame differencing captures the temporal information, which is the scene's motion independent of the camera motion. We fuse this information in the DCNN to receive an effective and efficient approach for robust motion segmentation. The proposed approach is scene-independent [33], which is non-compulsory in foreground-background segmentation [9, 25]. Our contributions are: (1) we present an efficient fusion DCNN architecture that uses a difference image as second modality in addition to the current image. In contrast to existing literature [56] this difference image does not come from background subtraction but from three-frame differencing, i.e. we just need two images in addition to the current image. (2) we re-labeled the CDNet-2014 dataset for the task of motion segmentation. The new labels are published here[1]. (3) we provide ablation studies to identify the best frame differencing approach and to show that difference images can be a better source of motion information for a DCNN than optical flow or 3D convolutions.

The remainder of this paper is organized as follows: related work is discussed in Section 2. Our modification of the CDNet-2014 is presented in Section 3. The proposed approach is introduced in Section 4. Experimental results are described in Section 5. We conclude in Section 6.

2. Related Work

For a stationary camera, background modeling and subtraction is a well-established approach and is applied successfully and thus extended since many years [2, 40, 51, 53]. The main drawbacks of this method are the initialization time for modeling the background and the residual image alignment error when using a moving camera [65]. As an alternative especially for a moving camera, optical flow is incontestably a well-fitting approach for motion segmentation [11, 35, 36]. However, the quality of the optical flow vectors is crucial for the motion segmentation performance. Hence, rather large moving objects in the image can be segmented well but small objects remain an issue due to blurry edges and low resolution [44]. Frame differencing is another relevant approach [10, 42, 50]. It is highly adaptive to illumination changes [47], but objects are often just partially detected or split [57]. In order to achieve higher robustness, hybrid approaches combine background subtraction with frame differencing [61, 63] or optical flow [12, 38, 65]. Since we expect small-scale objects in the images and as we use a moving camera, frame differencing seems to be the most promising approach. Another option, however, can be short-term background modeling and subtraction with background image alignment according to the estimated camera motion [43]. The background image is calculated over few frames using either pixel-wise floating average [22] or pixel-wise median [24, 1]. Obligatory for most methods described here is that post-processing is usually performed by morphological operations to reduce noise and fill holes in the difference image followed by thresholding to get a binary motion mask [67].

The introduction of deep learning and DCNNs opened up new opportunities for improvements [6]. Many authors, however, train their DCNNs to model the background scene-specifically or scene-dependently [9, 25, 26, 60], which is a great drawback if we use a moving camera. Scene-independence is gained by focusing the DCNN on learning frame differencing rather than modeling a specific background. Two reference frames together with the current frame are passed together to the DCNN by Tezcan *et al*. [56]. Background images or single images without moving objects are potential reference frames, but they need to be picked or generated beforehand. Another promising idea inspired by frame differencing is to feed multiple consecutive frames jointly into a DCNN and use a combination of 2D and 3D convolutions to calculate spatio-temporal features and thus to extract the temporal motion information inherently contained in this input image stack [4, 21, 32, 31, 39]. Unfortunately, 3D convolutions introduce a large number of parameters (weights) and are computationally expensive. In addition, semantic segmentation can be used to introduce appearance information that can help to distinguish between relevant and irrelevant motion [8, 27, 56]. This approach, however, is highly dependent on the quality of the semantic segmentation algorithm. Furthermore, we need to know the expected object classes in advance. As we aim at being class-agnostic and as we want our approach to work with visual-optical (VIS) and thermal infrared (IR) images likewise, we avoid using semantic segmentation. In general, combining appearance and motion seems to be a good option though. Hence, leveraging one difference image for multi-modal fusion with the current image within a DCNN seems to be an intuitive and efficient approach. To the best of our knowledge, we have not seen a similar approach in the literature, yet.

3. Adaptation of the CDNet-2014 Dataset

The CDNet-2014 dataset with its 53 video sequences is the most popular public dataset for the evaluation of change detection approaches such as foreground-background segmentation or background subtraction [18]. However, the dataset is labeled for change detection only. Hence, objects that stop moving during the sequence are still labeled for the remainder of the sequence. This is not desired for motion segmentation. Before the dataset can be used for training and evaluation, it has to be adapted to meet the different requirements of motion segmentation: objects that stop moving should be considered as background and must be labeled as such in the GT. Therefore, we implemented an

[1] https://github.com/HensoldtOptronicsCV/MotionSegmentation

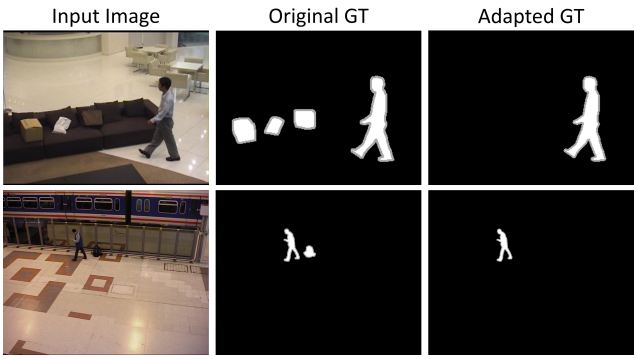

Figure 2. Examples for the CDNet-2014 dataset ground truth adapted to the task of motion segmentation.

algorithm that automatically adapts the GT. In order to recognize static objects, their movement between consecutive frames is tracked. As features we use the contours of the original GT and the surrounding bounding boxes.

The adaptation of the dataset only takes pixels annotated as *motion* into account as only these pixels are relevant for the evaluation. Pixels annotated as *unknown motion* are temporarily changed to be *background*. After the adaptation these pixels are recovered by applying a three pixel wide *unknown motion* border around each object. An object is tracked by calculating the Intersection over Union (IoU) between its bounding box in frame I_t and all bounding boxes of objects present in the next frame I_{t+1}. By setting a threshold T_{IoU} to define the highest IoU, before the object is considered static, it is possible to detect some static objects based on the change of the IoU alone. However, motion that does not change the size or position of the bounding box cannot be detected. Examples for this are the scenes *office* of the category *baseline* and the scene *library* of the category *thermal*. In both scenes a person is reading a book and occasionally turns the pages. Every motion caused by turning the pages happens within the boundary of the bounding box. This leads to the person being incorrectly removed from the GT. To reduce the number of false removals of objects that are in fact moving, the change in the position of the Center of Mass (CoM) of the GT object between consecutive frames is introduced as additional feature. By considering the change in the CoM, it is possible to detect movements that cause a change in the GT without changing the bounding box. Since the GT sometimes changes slightly between frames, even if the object does not move, a threshold value T_{CoM} is set to allow small changes in the center of mass. Considering both values, the IoU and the CoM in combination, leads to good results in the removal of static objects, as long as they do not overlap with the GT objects that are in motion. Further details are provided in the supplementary material.

In total, the 14 scenes *office*, *PETS2016*, *streetLight*, *tramstop*, *parking*, *abandonedBox*, *sofa*, *tunnelExit_0_35fps*, *copyMachine*, *diningRoom*, *corridor*, *lakeSide*, *library*, and *turbulence2* were modified to label static objects as background. This makes the dataset usable for the motion segmentation task. In Fig. 2, we can see some example images together with their related GT. The objects on the sofa and the backpack on the floor became static during the sequence. With our adaptation of the GT, they got labeled as background after they became static. We name this adaptation CDNet-2014-MotSeg.

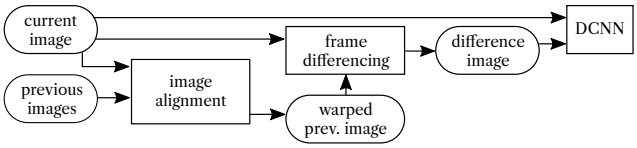

Figure 3. Overview of the proposed approach: optional image alignment followed by frame differencing and the fusion DCNN.

4. Proposed Approach

Inspired by recent methods that combine appearance and motion information using optical flow within a DCNN [49, 58], we aim at fusing appearance and motion using frame differencing. In this way, we avoid using 3D convolutions that are able to calculate spatio-temporal features [4] but significantly increase the number of weights and thus produce high computational effort during inference. However, optical flow may not be the best choice to detect slow motion of small objects in the image due to smoothing and regularization (see Section 5). An overview of the proposed approach is shown in Fig. 3. We need previous images for frame differencing that we align first if we have a moving camera. Frame differencing outputs a difference image that we feed into the DCNN together with the current image.

4.1. Frame Differencing

Frame differencing is a simple technique to subtract aligned images in order to make differences between these images apparent. If we assume that the acquisition time between the images is in a range of seconds or even milliseconds, we can expect that those differences originate from currently moving objects. Popular frame differencing approaches are two- and three-frame differencing. A good overview of frame differencing approaches in a slightly different context is provided by Sommer et al. [52]. We first perform image registration and alignment. Local image features are detected [48] and tracked across the image sequence using sparse optical flow [30]. This approach provides good accuracy without sacrificing too much speed. An affine transformation matrix (also known as homography) is estimated and used to align the previous images to the current one. This step is optional and suitable only, if the camera is moving. Then, we subtract the images. As

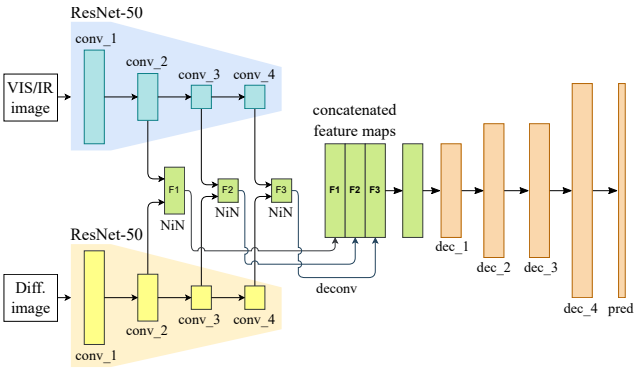

Figure 4. Proposed architecture of the multi-modal DCNN that fuses appearance and motion information using the current image and a difference image. Light blue and yellow blocks represent the layers of the ResNet-50 encoder. Green blocks represent the fusion modules and orange blocks the layers of the decoder.

we test different variants of frame differencing (18 in total), we describe the considered methods in Section 5 next to the related experiment, in which we provide an ablation study to identify the best performing frame differencing approach for our task. Finally, we apply morphological operations directly to the gray-valued difference image. We use morphological opening to remove noise followed by morphological closing to fill holes in the difference image. This size of the structuring element is discussed in Section 5.2. This image is then fed into the DCNN together with the current image.

4.2. DCNN Architecture

Figure 4 shows an overview of the proposed DCNN that is based on an encoder-decoder architecture. This DCNN fuses appearance and motion information: the appearance information provided by the RGB/IR image and the motion cues present in the related difference image. We feed the two images to the DCNN simultaneously. To fit the expected dimension of the input layer that requires three channels, the difference image is cloned, which is a common technique for gray-value infrared images, too [20]. For each modality, we use a ResNet-50 [16] backbone as encoder. We follow a *hybrid* fusion strategy: by combining information within the encoder stages and before the information is passed to the decoder, hybrid fusion strategies become more robust and achieve more accurate results than *early* and *late* fusion strategies [66]. Hence, an adaptive hybrid fusion approach [62] is chosen to fuse the appearance and motion features computed by the individual ResNet-50 encoders at different stages. Another advantage of this approach is that it is able to detect small-scale objects in the image. Rather simple fusion techniques such as concatenation, element-wise addition, multiplication, or averaging are commonly used to fuse feature maps [13]. However, when handling feature maps of different modalities, these simple techniques are unable to generate an optimal joint representation [66]. Thus, a simple yet efficient fusion module is implemented based on Network-in-Network (NiN) [28]. In the NiN block, feature maps of both modalities' individual convolutional layers (e.g., conv_2 of the VIS/IR encoder and conv_2 of the difference image encoder) are concatenated. Then, a convolutional layer with a 1×1 kernel, followed by a Rectified Linear Unit (ReLU) activation is applied to reduce the channel dimension to its original size and merge the information (rightmost green rectangle in Fig. 4). In accordance with [62], we merge intermediate feature maps at three stages after ResNet-50's conv_2, conv_3, and conv_4, followed by a final concatenation and convolutional layer. As proposed by Lui et al. [29], we deconvolve before the concatenation to have concatenated feature maps of similar size. After encoding, the resulting feature maps are passed to the decoder network.

The decoder structure is inspired by Lim et al. [25] and consists of four blocks of transposed convolutions denoted by dec_1 to dec_4 followed by a final prediction layer. The decoder network gradually restores the original spatial resolution of the input image and outputs a motion probability for each pixel. Each block except for the last one consists of a set of three transposed convolutional layers structured in a similar way as the ResNet bottleneck block without the skip connection. The first transposed convolution reduces the channels of the high dimensional feature maps to 64 dimensions by applying a 1×1 kernel with stride 1. The second layer applies a 3×3 (block 1 and 3) or 5×5 (block 2) transposed convolution without changing the size of the depth dimension. Block 2 applies this layer with a stride of 2 to upscale the features to spatial dimensions from $\frac{W}{4} \times \frac{H}{4}$ to $\frac{W}{2} \times \frac{H}{2}$ with W and H being the original image width and height, respectively. The final layer of each block increases the size of the depth dimension again. The final decoder block only consists of one transposed convolution with a kernel of size 5×5 and stride 2 to upscale the features from $\frac{W}{2} \times \frac{H}{2}$ back to the original input size. The prediction layer applies a final transposed convolution with kernel size 1×1 and reduces the depth dimension to 1. The sigmoid activation is applied in order to produce the final dense motion probability map. All other layers of the model use ReLU.

4.3. Training Strategy

As the CDNet-2014 dataset does not come with a suggested split for training and test data, we closely follow the recently published training and test strategy by Tezcan et al. [55]. The authors propose a 4-fold cross-validation strategy. All CDNet-2014 sequences are divided equally into four disjoint splits. The model is trained on three of the splits. The remaining split is used to evaluate the model's performance. Please note that this remaining split is considered as test data and not as validation data

Figure 5. Difference images variants originating from slightly different frame differencing approaches and/or parameterization.

during training. Each sequence is contained in the test data exactly once. In this way, four models are trained to obtain scene-independent results for each sequence of the dataset.

The choice of the hyper parameter setup for the training process of the model is based on the commonly used values in the literature. Adaptive Moment Estimation (Adam) [19] is used to optimize the network during the training process. The model is trained for a maximum of 30 epochs with a learning rate of $1e^{-4}$ and a batch size of 8. The motion detection task is a pixel-wise classification of *static* and *motion* pixels. Hence, binary cross-entropy is chosen as the loss function to evaluate the model's performance during training. The designated training data is divided into a training split (90 %) and validation split (10 %). At the beginning of each epoch all training samples are shuffled to avoid a repeating order of training samples, which may bias the network. Early stopping is used to prevent overfitting: if the validation loss does not improve in four consecutive epochs, the training process is terminated.

Furthermore, we use standard data augmentation techniques based on geometric transformations: *resize*, *crop*, and *flip*. Each individual transformation is applied to all inputs and their corresponding ground truth with a probability of 50 %. The resizing transformation randomly resizes the images with a scaling factor ranging from 0.5 to 1.5. Afterwards, the cropping augmentation randomly crops an area of size 320×240 pixels. Each time the training data is randomly resized and cropped, a different area of the images is used for training. This significantly increases the number of possible training samples. The final flip augmentation then randomly flips the images horizontally. We skip augmentation techniques for color space manipulations such as color jitter as they cannot be applied to a gray-scale difference image or a thermal IR image.

4.4. Frame Differencing Data Augmentation

In addition to the just mentioned standard data augmentation techniques, we also test a novel data augmentation approach specifically for difference images. Inspired by color space manipulations that aim at making the model more robust against illumination changes and eliminate color biases, the goal of our approach is to make the model invariant to the presented motion cues. To achieve this, the training strategy is modified to randomly select one of the 18 different frame differencing variants for each training sample. This essentially varies the temporal context and motion representation of each input example. The model could therefore better generalize regarding the relationship between the given input image and the presented motion cues. Figure 5 shows an example of all possible difference images resulting from the same given input image.

5. Experiments and Results

In this section, we first briefly introduce the CDNet-2014 dataset and the evaluation measures. Then, we present an ablation study to find the best frame differencing approach and parameterization for our fusion DCNN. A thorough comparison with optical flow as additional modality instead of frame differencing is provided subsequently. Then, further improvements are discussed followed by a comparison with current state-of-the-art approaches.

5.1. Dataset and Evaluation Measures

Wang *et al*. [59] introduced the CDNet-2014 benchmark, which includes over 160,000 pixel-wise annotated frames in 53 video sequences subdivided in 11 categories and two spectra: VIS and thermal IR. The 53 sequences contain a large variety of different scenes with varying image quality and resolution ranging from 320×240 to 720×480 pixels. Most scenes show an urban environment with persons or cars. The dataset contains both indoor and outdoor scenes and covers many different real world challenges such as dynamic backgrounds, shadows, and camera motion. The GT for each image is a gray-scale image that describes the 4 motion classes: static, hard shadow, unknown motion, and motion. An additional class is used to mark areas that are outside the region of interest (non-ROI). Pixels annotated as non-ROI are discarded during evaluation. Every sequence starts with a number of non-ROI frames that are intended to be used by background subtraction methods to initialize their background model. In some sequences the non-ROI annotation is also used to constrain the training and evaluation to a certain area. For our experiments, we use the CDNet-2014 dataset with our annotations according to Section 3. We use the standard evaluation measures [59] Recall (Re), Specificity (Sp), False Positive Rate (FPR), False Negative Rate (FNR), Percentage of Wrong Classification (PWC), Precision (Pr), and F_1-score. Here in the paper, we use the F_1-score only. Please refer to the supplementary material to see the other measures. Finally, we follow the evaluation protocol of Tezcan *et al*. [55] to determine different dataset splits in training and test data.

5.2. Frame Differencing Experiments

First, we conduct an extensive ablation study on different frame differencing variants and parameterizations

used as second modality for the DCNN. Difference images are created by either two-frame or three-frame differencing [42, 50]. While two-frame differencing is evaluated for temporal offsets ΔF of one, five, and ten frames between the current and the reference frame, three-frame differencing is investigated using two different fusion strategies, minimum and sum (see Eqs. 1 and 2), and a ΔF of one, two, and five frames between each frame starting from current going back. Furthermore, the frames are aligned for the categories *PTZ* and *camera jitter* using the same technique as described in Section 4.1 and the difference image is calculated with symmetric neighbourhood consideration as proposed by Saur et al. [46]. There is no binary thresholding applied to the difference images to avoid biasing the DCNN during training. To reduce noise and fill holes, we sequentially apply morphological opening and closing with a Structuring Element (SE) of size 3×3 for both operations (*small*) as well as 3×3 for opening and 15×15 for closing (*large*). The second configuration for morphological operations is chosen as large moving objects are often not entirely detected by frame differencing and the large SE for closing is expected to merge those partial detections.

$$D_{min}(x,y) = \min(|I_t(x,y) - \hat{I}_{t-1}(x,y)|, |I_t(x,y) - \hat{I}_{t-2}(x,y)|) \quad (1)$$

$$D_{sum}(x,y) = |I_t(x,y) - \hat{I}_{t-1}(x,y)| + |I_t(x,y) - \hat{I}_{t-2}(x,y)| \quad (2)$$

Equations 1 and 2 explain how frame differencing is done throughout the scope of this paper. In contrast to other three-frame differencing approaches [63] we do not use past and future frames to calculate the differences. Instead, we use only past frames as we want to enable inference of live sources (e.g., streams from surveillance cameras) without adding a constant delay. Table 1 shows the results for the above described ablation study. Referring to the F_1-score, our DCNN performs best if the difference image is calculated using three-frame differencing with ΔF of 5 frames distance between each considered frame, using the sum of differences as fusion, the small SE for morphological operations, and a confidence threshold of 0.4 to consider a pixel labeled as motion. Furthermore, there is a tendency that three-frame differencing slightly outperforms two-frame differencing. It may be worth to evaluate if this gap can by closed by explicit handling of ghosting [45]. A more detailed discussion on this ablation study is provided in the supplementary material.

5.3. Optical Flow Experiments

Optical flow is another popular approach for motion segmentation. Thus, we conduct two experiments to evaluate the usability of optical flow as motion representation. First, optical flow is investigated stand-alone and then, secondly, in combination with our best performing frame differencing approach to see if it can be beneficial to have two different representations of motion. PWC-Net [54] is utilized to get the optical flow for our experiments and by stacking together transition in x- and y-direction (ΔX and ΔY) and the transitions magnitude M to a three channel image $[\Delta X, \Delta Y, M]$ (*Mag*) we follow the suggestion of different authors [34, 58]. Similar to our experiments on frame differencing, image alignment is performed only for the categories *PTZ* and *camera jitter*. To overcome the expected issues of optical flow for small-scale objects due to smooth-

Table 1. Ablation study on frame differencing variants as input for the multi-modal DCNN.

Diff. frames	ΔF	Fusion	Morph. SE	Conf. thrs.	F_1
3	**5**	**sum**	**small**	0.4	**0.7450**
3	2	sum	small	0.3	0.7171
3	1	sum	small	0.3	0.7101
3	5	sum	large	0.4	0.7339
3	2	sum	large	0.4	0.7394
3	1	sum	large	0.5	0.6870
3	5	min	small	0.3	0.7355
3	2	min	small	0.3	0.7085
3	1	min	small	0.3	0.6581
3	5	min	large	0.3	0.7389
3	2	min	large	0.4	0.7095
3	1	min	large	0.4	0.6689
2	10		small	0.4	0.7009
2	5		small	0.3	0.7202
2	1		small	0.4	0.6810
2	10		large	0.2	0.6620
2	5		large	0.5	0.6770
2	1		large	0.5	0.6413

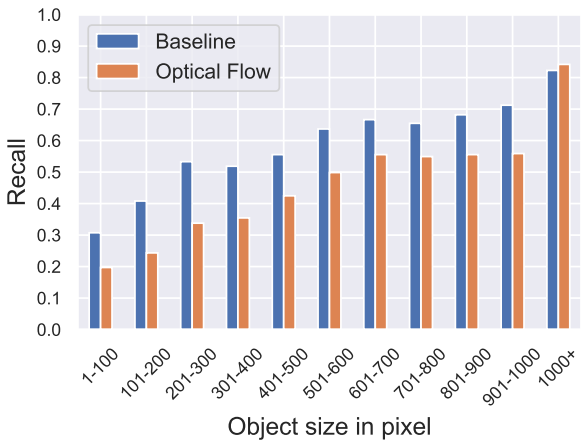

Figure 6. Size dependent Recall for our baseline and the best optical flow based approach to represent motion in the fusion DCNN.

Table 2. F_1-score for our experiments using optical flow as input.

ΔF	Third Channel	Conf. Thrs.	F_1
1	Mag	0.3	0.604
1	Diff	0.3	0.743
5	Mag	0.5	0.595
5	Diff	0.3	0.712
current baseline			**0.745**

Table 3. F_1-score for our experiments on further improvements.

Approach	Conf. Thrs.	F_1
Frame Diff. Data Augmentation	0.4	0.687
Multi-Scale Backbone	0.4	0.734
current baseline		**0.745**

ing and regularization, we replace the magnitude with our best performing frame difference image in order to provide an input beneficial for small-scale objects (*Diff*).

Figure 6 shows a bar diagram that plots the Recall against different object sizes. The expectation is that the mentioned issues of optical flow can be made visible in false negative detections and hence in the Recall. We compare our current baseline and optical flow as input. We can observe a significant gap in performance preferring our baseline. This gap gets larger for smaller objects confirming our expectation. Table 2 shows our results for the both experiments described above. For a fair comparison, we also vary the frame gap ΔF for the optical flow. According to the F_1-score our frame differencing baseline approach outperforms all approaches that utilize optical flow including the combination of optical flow and frame differencing.

5.4. Potential Further Improvements

Considering further improvements for our baseline approach, data augmentation and a multi-scale backbone [14] are investigated as further optimizations. Data augmentation is done by randomly selecting one of the 18 difference image variants. As described in Section 4.4, the goal of this augmentation is to make the model invariant to the presented motion cues. Furthermore, as seen in Figure 6, small-scale objects are most challenging. To overcome this issue the multi-scale backbone Res2Net [14] is used instead of ResNet-50. Due to its multi-scale architecture it is expected to provide stronger features for small-scale objects.

Table 3 shows that none of our improvement attempts seems to be promising. Even the new backbone did not improve the performance. The limited amount of training data available in CDNet-2014 could be the reason for this lack of improvement, as the multi-scale backbone introduces additional weights that need more data to be trained effectively.

Table 4. Quantitative evaluation for comparing our approach with the state-of-the-art. The F_1-score can be lower compared to Pr and Re due to the evaluation protocol's averaging strategy [55].

Approach	Conf. Thrs.	Pr	Re	F_1
Ours	0.40	**0.774**	**0.751**	**0.745**
Bosch [4]	0.40	0.626	0.673	0.553
Xiao *et al.* [63]	0.05	0.462	0.513	0.420
STBGS	0.10	0.406	0.549	0.401
Frame Diff	0.15	0.375	0.580	0.389

5.5. Comparison with the State-of-the-Art

It is rather difficult to identify methods for comparing our proposed approach with the state-of-the-art. Most existing methods in the literature are either scene-specific or they need too much time (e.g., 100 frames) to train a background model or they cannot compensate for camera motion. Furthermore, most authors do not provide code, which is an issue as we changed the GT annotations. However, we can provide a quantitative and qualitative evaluation by comparing our approach with three different algorithms for motion and segmentation: Bosch's [4] DCNN based approach that utilizes 2D and 3D convolutions, a part of Xiao's [63] approach with three-frame differencing and pixel-wise minimum calculation, Short-Term Background Subtraction (STBGS) by aligning and averaging ten subsequent images as for example proposed by several authors [22, 52], and three-frame differencing with pixel-wise summation. Table 4 shows the quantitative results regarding Precision, Recall, and F_1-score. Our approach outperforms the other methods by a large margin.

The qualitative evaluation is visualized in Fig. 7. The first row shows the GT followed by the results of the considered approaches. Columns one and two each show a frame taken from sequences with turbulence and small-scale objects. Our approach has by far the smallest number of false positives (red color), but is still able to recognize both objects. Column three and four show scenes with irrelevant foreground motion and dynamic background, respectively. In column three the fountains produce irrelevant motion causing false positives for all approaches. However, our approach again has the lowest number of false positives while detecting the car entirely. Column four shows another challenging scene as the moving water is highly dynamic and introduces variations between frames due to glint, waves, and reflections. Both frame differencing and STBGS produce false positives on the water's surface. Bosch's approach produces false positives and false negatives on the canoe. Our approach segments the moving region best.

In summary, we state that we found a simple, effective, and efficient approach for motion segmentation in challenging environmental conditions. Our approach achieves the highest F_1-score on the CDNet-2014-MotSeg benchmark.

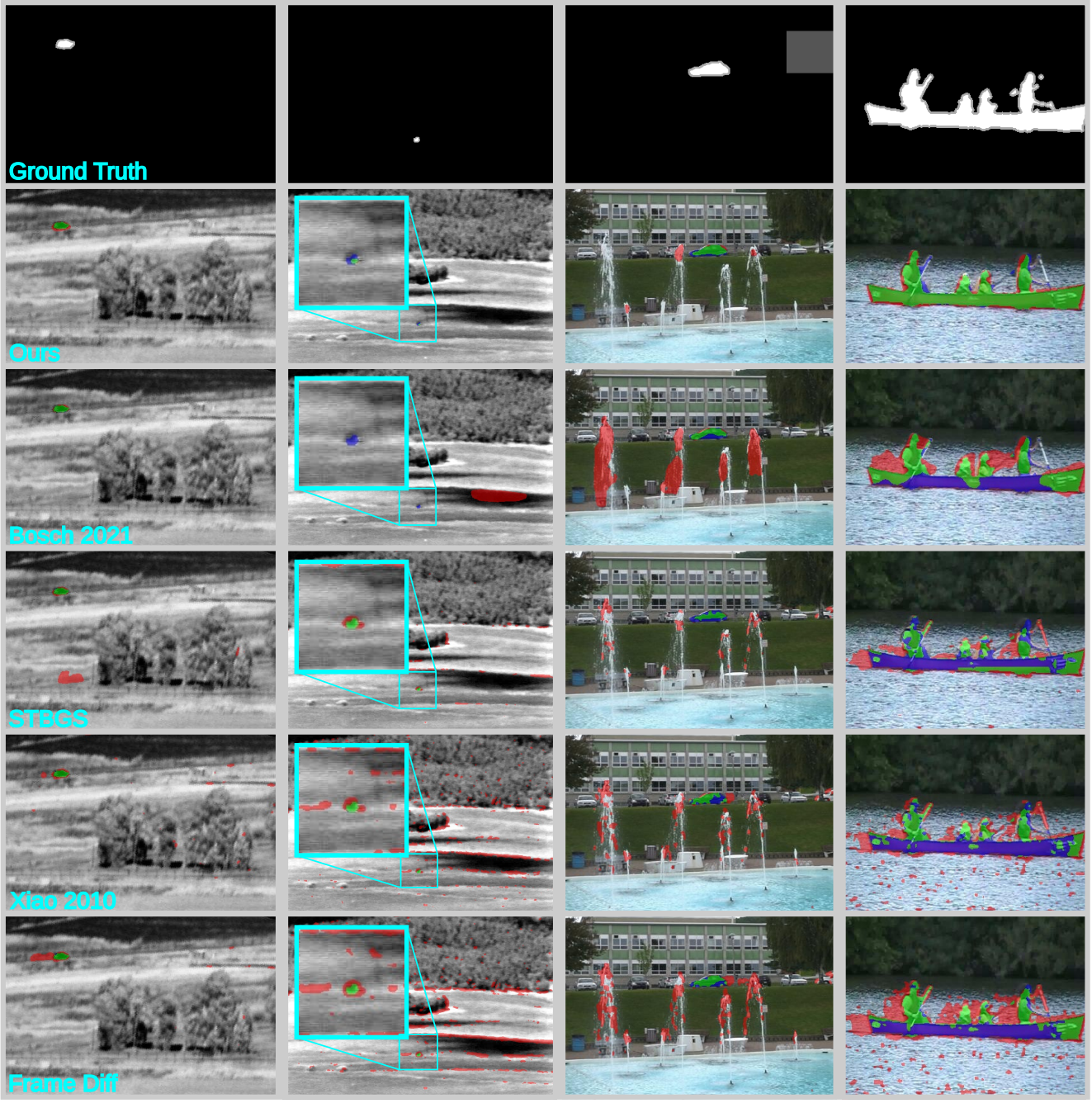

Figure 7. Qualitative evaluation of our approach against other approaches for motion segmentation. True positive detection are visualized in green color, false positives in red color, and false negatives in blue color. STBGS denotes Short Term Background Subtraction.

6. Conclusion

We presented a novel DCNN architecture for motion segmentation under difficult environmental conditions. The core idea is to fuse appearance and motion information. Besides the current image that represents the appearance information, we add the related difference image as a second modality and feed it into the DCNN together with the current image. The proposed approach outperformed other state-of-the-art methods based on optical flow or 3D convolutions on the challenging CDNet-2014 dataset. As the CDNet-2014 dataset is labeled for change detection only, we re-labeled the dataset for motion segmentation and call the result CDNet-2014-MotSeg. The new labels are publicly available. We achieve an overall F_1-score of 0.745, so we still see space for improvement.

References

[1] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll. A Deep Convolutional Neural Network for Background Subtraction. *Pattern Recognition*, 76:635–649, 2017. 2

[2] Olivier Barnich and Marc Van Droogenbroeck. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, 2011. 2

[3] Pia Bideau and Erik Learned-Miller. A Detailed Rubric for Motion Segmentation. *arXiv preprint arXiv:1610.10033*, 2016. 1

[4] Markus Bosch. Deep learning for robust motion segmentation with non-static cameras. Master's thesis, Ulm University of Applied Sciences, Ulm, Germany, 2021. 2, 3, 7

[5] Thierry Bouwmans. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*, 11–12:31–66, 2014. 1

[6] Thierry Bouwmans, Sajid Javed, Maryam Sultana, and Soon Ki Jung. Deep neural network concepts for background subtraction:A systematic review and comparative evaluation. *Neural Networks*, 117:8–66, 2018. 2

[7] Thierry Bouwmans, Fatih Porikli, Benjamin Höferlin, and Antoine Vacavant (Eds.). *Background Modeling and Foreground Detection for Video Surveillance*. CRC Press, 2014. 1

[8] Marc Braham, Sebastien Pierard, and Marc Van Droogenbroeck. Semantic background subtraction. In *IEEE ICIP*, 2017. 2

[9] Marc Braham and Marc Van Droogenbroeck. Deep background subtraction with scene-specific convolutional neural networks. In *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2016. 2

[10] Caiyuan Chen and Xiaoning Zhang. Moving Vehicle Detection Based on Union of Three-Frame Difference. In *International Conference on Electronic Engineering, Communication and Management (EECM 2011)*, 2012. 2

[11] Tao Chen and Shijian Lu. Object-Level Motion Detection from Moving Cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2333–2343, 2017. 2

[12] Ali Elqursh and Ahmed Elgammal. Online Moving Camera Background Subtraction. In *ECCV*, 2012. 2

[13] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2021. 4

[14] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2019. 7

[15] Belmar Garcia-Garcia, Thierry Bouwmans, and Alberto Jorge Rosales Silva. Background subtraction in real applications: Challenges, current models and future directions. *Computer Science Review*, 35, 2020. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016. 4

[17] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *IEEE CVPR*, 2017. 1

[18] Rudrika Kalsotra and Sakshi Arora. A Comprehensive Survey of Video Datasets for Background Subtraction. *IEEE Access*, 7:59143–59171, 2019. 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2014. 5

[20] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully Convolutional Region Proposal Networks for Multispectral Person Detection. In *IEEE CVPR*, 2017. 4

[21] Rodney LaLonde, Dong Zhang, and Mubarak Shah. ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information. In *IEEE CVPR*, 2017. 2

[22] Jian Li, Zhong-Ming Pan, Zhuo-Hang Zhang, and Heng Zhang. Dynamic ARMA-Based Background Subtraction for Moving Objects Detection. *IEEE Access*, 7:128659–128668, 2019. 2, 7

[23] Sheng Li, Kang Li, and Yun Fu. Temporal Subspace Clustering for Human Motion Segmentation. In *IEEE ICCV*, 2015. 1

[24] Xinxin Li, Michael K. Ng, and Xiaoming Yuan. Median filtering-based methods for static background extraction from surveillance video. *Numerical Linear Algebra with Applications*, 22(5):845–865, 2015. 2

[25] Long Ang Lim and Hacer Yalim Keles. Foreground segmentation using convolutional neural networks for multiscale feature encoding. *Pattern Recognition Letters*, 112:256–262, 2018. 2, 4

[26] Long Ang Lim and Hacer Yalim Keles. Learning Multi-scale Features for Foreground Segmentation. *Pattern Analysis and Applications*, 23:1369–1380, 2019. 2

[27] Chuming Lin, Bo Yan, and Weimin Tan. Foreground Detection in Surveillance Video with Fully Convolutional Semantic Network. In *IEEE ICIP*, 2018. 2

[28] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *ICLR*, 2014. 4

[29] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. Center and Scale Prediction: A Box-free Approach for Pedestrian and Face Detection. *arXiv preprint arXiv:1904.02948*, 2019. 4

[30] Bruce Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *DARPA Image Understanding Workshop*, 1981. 3

[31] Sabarinath Mahadevan, Ali Athar, Aljosa Osep, Sebastian Hennen, Laura Leal-Taixe, and Bastian Leibe. Making a Case for 3D Convolutions for Object Segmentation in Videos. In *BMVC*, 2020. 2

[32] Murari Mandal, Vansh Dhar, Abhishek Mishra, and Santosh Kumar Vipparthi. 3DFR: A Swift 3D Feature Reductionist

Framework for Scene Independent Change Detection. *IEEE Signal Processing Letters*, 26(12):1882–1886, 2019. 2

[33] Murari Mandal and Santosh Kumar Vipparthi. Scene Independency Matters: An Empirical Study of Scene Dependent and Scene Independent Evaluation for CNN-Based Change Detection. *IEEE Transactions on Intelligent Transportation Systems*, Early Access:1–14, 2020. 2

[34] Eslam Mohamed, Mahmoud Ewaisha, Mennatullah Siam, Hazem Rashed, Senthil Yogamani, Waleed Hamdy, Muhammad Helmi, and Ahmad El-Sallab. Monocular Instance Motion Segmentation for Autonomous Driving: KITTI InstanceMotSeg Dataset and Multi-task Baseline. *arXiv preprint arXiv:2101.09585*, 2021. 6

[35] Manjunath Narayana, Allen Hanson, and Erik Learned-Miller. Coherent Motion Segmentation in Moving Camera Videos using Optical Flow Orientations. In *IEEE ICCV*, 2013. 1, 2

[36] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, 2014. 2

[37] Omar Oreifej, Xin Li, and Mubarak Shah. Simultaneous Video Stabilization and Moving Object Detection in Turbulence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):450–462, 2013. 1

[38] Anestis Papazoglou and Vittorio Ferrari. Fast Object Segmentation in Unconstrained Video. In *IEEE ICCV*, 2013. 2

[39] Prashant W. Patil, Akshay Dudhane, and Subrahmanyam Murala. Multi-frame Recurrent Adversarial Network for Moving Object Segmentation. In *IEEE WACV*, 2021. 2

[40] Massimo Piccardi. Background subtraction techniques: a review. In *International Conference on Systems, Man and Cybernetics*, 2004. 2

[41] Richard Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005. 1

[42] Syaimaa Solehah Mohd Radzi, Shahrul Nizam Yaakob, Zulaikha Kadim, and Hon Hock Woon. Extraction of Moving Objects Using Frame Differencing, Ghost and Shadow Removal. In *International Conference on Intelligent Systems, Modelling and Simulation*, 2014. 2, 6

[43] Vladimir Reilly, Haroon Idrees, and Mubarak Shah. Detection and tracking of large number of targets in wide area surveillance. In *ECCV*, 2010. 2

[44] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Detecting Flying Objects Using a Single Moving Camera. *IEEE Transactions on Image Processing*, 39(5):879–892, 2016. 2

[45] Imran Saleemi and Mubarak Shah. Multiframe Many-Many Point Correspondence for Tracking of Swarms of Vehicles in Wide Area Aerial Videos. *International Journal of Computer Vision (IJCV)*, 104(2), 2013. 6

[46] Günter Saur, Wolfgang Krüger, and Arne Schumann. Extended image differencing for change detection in UAV video mosaics. In *Proc. SPIE Vol. 9026*, 2014. 1, 6

[47] Kamal Sehairi, Fatima Chouireb, and Jean Meunier. Comparative study of motion detection methods for video surveillance systems. *SPIE Journal of Electronic Imaging*, 26(2), 2017. 1, 2

[48] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE CVPR*, 1994. 3

[49] Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand, and Ahmad El-Sallab. MODNet: Motion and Appearance based Moving Object Detection Network for Autonomous Driving. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2018. 1, 3

[50] Nishu Singla. Motion Detection Based on Frame Difference Method. *International Journal of Information & Computation Technology*, 4(15):1559–1565, 2014. 2, 6

[51] Andrews Sobral and Thierry Bouwmans. BGS Library: A Library Framework for Algorithm's Evaluation in Foreground/Background Segmentation. In *Background Modeling and Foreground Detection for Video Surveillance*. CRC Press, 2014. 2

[52] Lars Sommer, Michael Teutsch, Tobias Schuchert, and Jürgen Beyerer. A survey on moving object detection for wide area motion imagery. In *IEEE WACV*, 2016. 3, 7

[53] Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE CVPR*, 1999. 2

[54] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *IEEE CVPR*, 2018. 6

[55] M. Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. BSUV-Net 2.0: Spatio-Temporal Data Augmentations for Video-Agnostic Supervised Background Subtraction. *arXiv preprint arXiv:2101.09585*, 2021. 4, 5, 7

[56] M. Ozan Tezcan, Janusz Konrad, and Prakash Ishwar. BSUV-Net: A Fully-Convolutional Neural Network for Background Subtraction of Unseen Video. In *IEEE WACV*, 2020. 2

[57] Chun-Ming Tsai and Zong-Mu Yeh. Intelligent Moving Objects Detection via Adaptive Frame Differencing Method. In *Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, 2013. 2

[58] Johan Vertens, Abhinav Valada, and Wolfram Burgard. SMSnet: Semantic motion segmentation using deep convolutional neural networks. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2017. 3, 6

[59] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. CDnet 2014: An Expanded Change Detection Benchmark Dataset. In *IEEE CVPR Workshops*, 2014. 1, 5

[60] Yi Wang, Zhiming Luo, and Pierre-Marc Jodoin. Interactive deep learning method for segmenting moving objects. *Pattern Recognition Letters*, 96:66–75, 2017. 2

[61] Zhihu Wang, Jiulong Xiong, and Qi Zhang. Motion saliency detection based on temporal difference. *SPIE Journal of Electronic Imaging*, 24(3), 2015. 2

[62] Alexander Wolpert, Michael Teutsch, M. Saquib Sarfraz, and Rainer Stiefelhagen. Anchor-free Small-scale Multispectral Pedestrian Detection. In *BMVC*, 2020. 4

[63] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, and Feng Han. Vehicle detection and tracking in wide field-of-view aerial video. In *IEEE CVPR*, 2010. 2, 6, 7

[64] Xun Xu, Loong-Fah Cheong, and Zhuwen Li. 3D Rigid Motion Segmentation with Mixed and Unknown Number of Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):1–16, 2021. 1

[65] Yan Zhang, Stephen J. Kiselewich, William A. Bauson, and Riad Hammoud. Robust Moving Object Detection at Distance in the Visible Spectrum and Beyond Using A Moving Camera. In *IEEE CVPR Workshops*, 2006. 2

[66] Yifei Zhang, Desire Sidibe, Olivier Morel, and Fabrice Meriaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:1–16, 2021. 4

[67] Yanzhu Zhang, Xiaoyan Wang, and Biao Qu. Three-frame difference algorithm research based on mathematical morphology. *Procedia Engineering*, 29:2705–2709, 2012. 2