University of Missouri

Motion U-Net: Multi-cue Encoder-Decoder Network for Motion Segmentation

Gani Rahmon[1], Filiz Bunyak[1], Guna Seetharaman[2], Kannappan Palaniappan[1]

[1]University of Missouri-Columbia, MO, USA

[2] U.S Naval Research Laboratory, Washington, DC.

ICPR 2020

Introduction

Detection of moving objects in a video stream

- Many computer vision applications (video surveillance and monitoring to self-driving cars)
- Provide focus of attention for later processes such as tracking, activity & event analysis
- **Performance** of the following tasks rely heavily on **moving object detection accuracy**

Challenges:

- Environmental conditions: illumination changes, shadows, glare, background clutter etc.
- Foreground complexities: occlusion, camouflage, complex motion behavior of the foreground objects;
- Imaging conditions: low resolution and/or framerate, camera jitter etc.

bad weather

night videos

Illumination variations & shadows

dynamic background

[1] Change Detection 2014 Dataset, http://www.changedetection.net/

Proposed

- **Motion U-Net (MU-Net):** A novel hybrid and multi-cue system for robust moving object detection
- Integrates motion, change, and appearance cues using a deep learning framework
- Framework: encoder-decoder deep convolutional neural network
- **Motion and change cues:** our Flux tensor and a multi-modal background subtraction modules
- **Spatio-temporal fusion + appearance cues:** deep learning

- Two versions of MU-Net are proposed:
  - MU-Net1 uses original video frame (3 channel RGB, single frame no temporal information),
  - MU-Net2 uses 3 channel input stream consisting of
    - Original video frame (RGB converted to grayscale),
    - Motion mask (flux tensor motion detection)
    - Change mask (multi-modal background subtraction)

Motion U-Net Networks: MU-Net1

Single-stream Spatial-only Detection Using Semantic Segmentation

- Input: Single RGB frame
- Output: Binary Mask
- Backbone: ResNet-18
- Network learns to segment moving objects using only appearance cue.

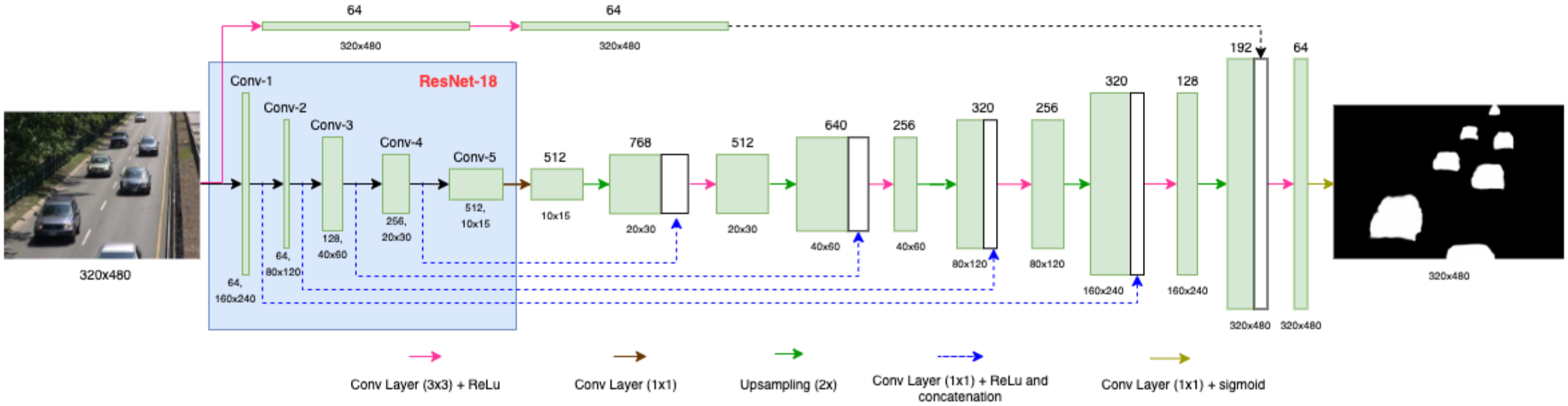


Figure 1: Single-stream Motion U-Net1 encoder-decoder network using a single RGB frame as input without temporal cues.

MU-Net1: Backbone

Single-stream Spatial-only Detection Using Semantic Segmentation

- Input: Single RGB frame
- Output: Binary Mask
- Backbone: ResNet-18
- Network learns to segment moving objects using only appearance cue.

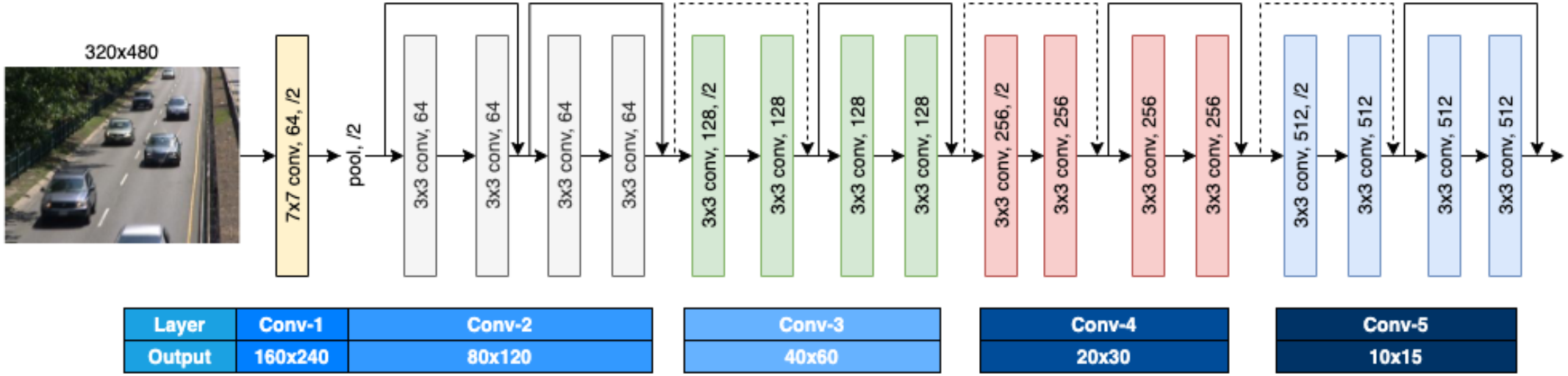


Figure 2: ResNet-18 encoder backbone used in Motion U-Net with five convolution layers.

Motion U-Net Networks: MU-Net2

Single-stream Early Fusion for Spatio-temporal Change Detection

- Input: RGB frame converted to grayscale, change (BGS) and flux motion
- Output: Binary mask
- Backbone: ResNet-18
- Network learns to segment moving objects using appearance, change and motion cues.

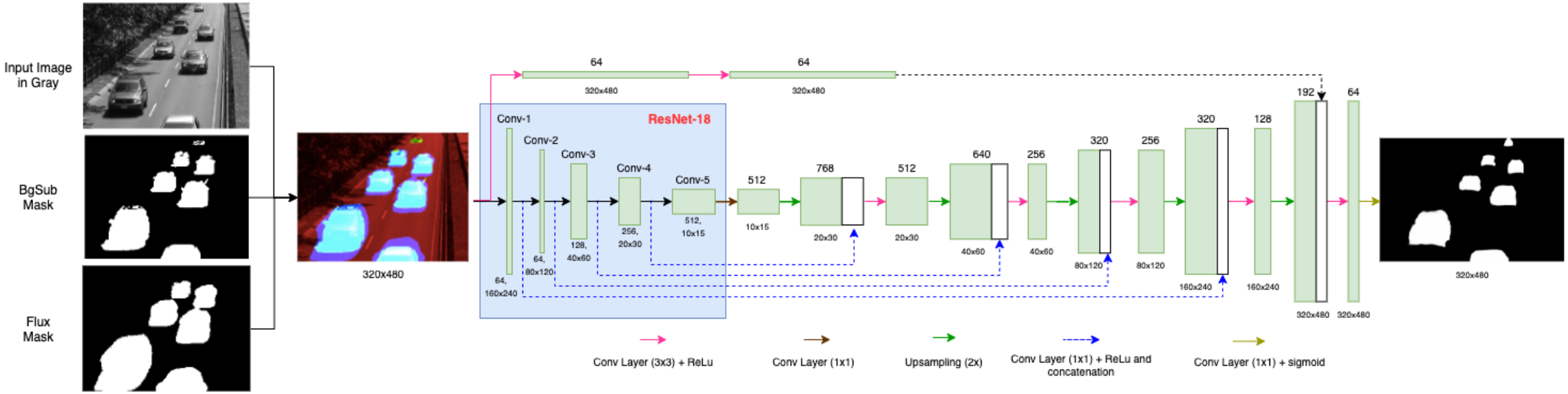


Figure 3: The single-stream 3-channel Motion U-Net2 architecture, same as Figure 2, but with a three cue input stream which includes appearance (RGB frame, time t) in the first channel, second and third channels, respectively.

Change Cue

1. Multi-modal background subtraction for change estimation

- Change: estimated using an adaptive multi-modal background subtraction approach (OpenCV: BackgroundSubtractorMOG2 [2]).

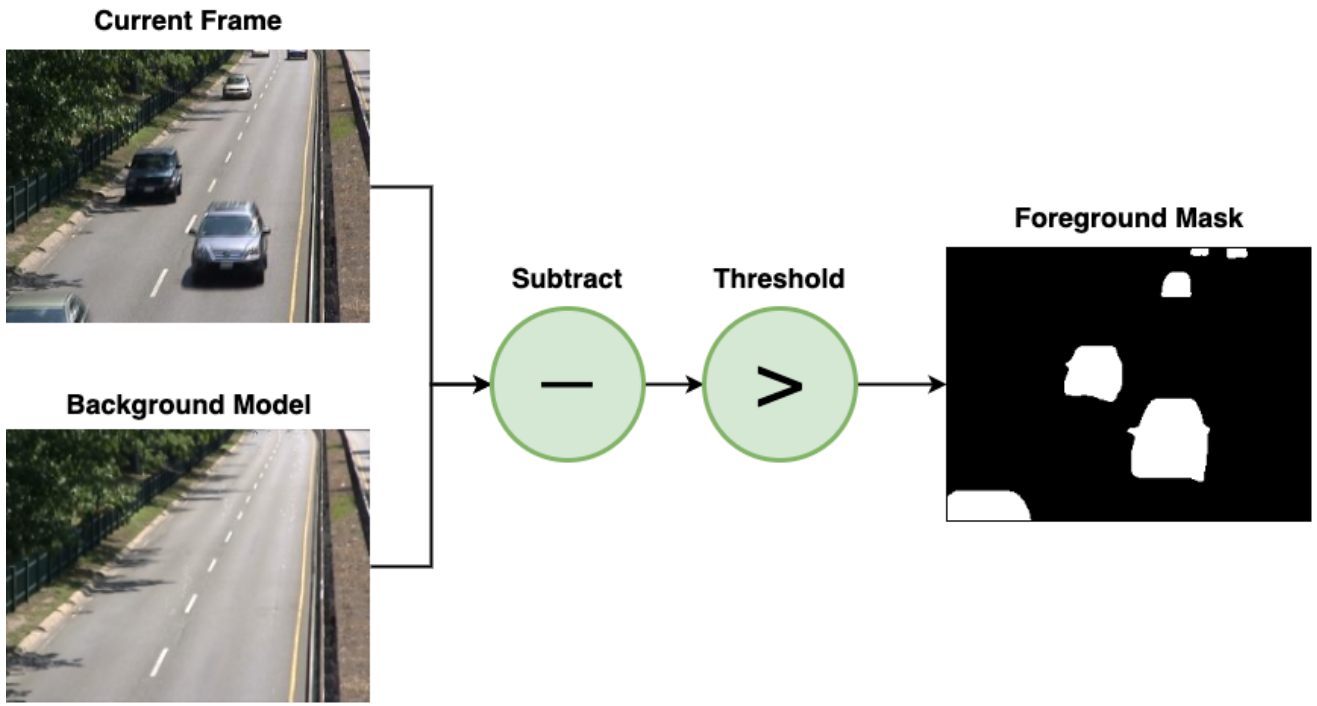


Figure 4: Foreground mask computation using background subtraction methods.

[2] *Opencv: How to use background subtraction methods*. [Online]. Available: https://docs.opencv.org/3.4/d1/dc5/ tutorial background subtraction.html.

Motion Cue

Flux tensor motion estimation

- **Flux tensor**: temporal variation of the optical flow field within local 3D spatio-temporal volume
- Minimization results in:

$$J_F(x,W) = \int_{\Omega} W(x,y) \frac{\partial}{\partial t} \nabla I(x) . \frac{\partial}{\partial t} \nabla I^T(x) dy$$

$$\frac{\partial}{\partial t}\left(\frac{dI(x)}{dt}\right) = \frac{\partial^2 I(x)}{\partial x \partial t} v_x + \frac{\partial^2 I(x)}{\partial y \partial t} v_y + \frac{\partial^2 I(x)}{\partial t^2} v_t + \frac{\partial I(x)}{\partial x} a_x + \frac{\partial I(x)}{\partial y} a_y + \frac{\partial I(x)}{\partial t} a_t$$

Advantages:

- Elements of the flux tensor incorporate information about temporal gradient changes: <u>efficient discrimination between stationary and moving image features</u>
- The trace of the flux tensor matrix can be directly used to classify moving and non-moving regions <u>without the need for expensive eigenvalue decompositions</u>.

$$J_F = \begin{bmatrix} \int_\Omega \left\{\frac{d^2 I}{dxdt}\right\}^2 dy & \int_\Omega \frac{d^2 I}{dxdt}\frac{d^2 I}{dydt} dy & \int_\Omega \frac{d^2 I}{dxdt}\frac{d^2 I}{dt^2} dy \\ \int_\Omega \frac{d^2 I}{dydt}\frac{d^2 I}{dxdt} dy & \int_\Omega \left\{\frac{d^2 I}{dydt}\right\}^2 dy & \int_\Omega \frac{d^2 I}{dydt}\frac{d^2 I}{dt^2} dy \\ \int_\Omega \frac{d^2 I}{dt^2}\frac{d^2 I}{dxdt} dy & \int_\Omega \frac{d^2 I}{dt^2}\frac{d^2 I}{dydt} dy & \int_\Omega \left\{\frac{d^2 I}{dt^2}\right\}^2 dy \end{bmatrix}$$

$$trace(J_F) = \int_\Omega || \frac{d}{dt} \nabla I ||^2 dy$$

[3] F. Bunyak, K. Palaniappan, S. Nath, and G. Seetharaman, "Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking," *Journal of multimedia*, vol. 2, no. 4, p. 20, 2007.

Evaluation

- Change Detection 2014 dataset
  - seven assessment metrics
  - Change Detection Workshop website [4]
- Seven metrics:
  - recall (Re)
  - specificity (Sp)
  - false positive rate (FPR)
  - false negative rate (FNR)
  - precision (P)
  - F-Measure (F)
  - percentage of the wrong classification (PWC)

$$Re = \frac{TP}{(TP + FN)}; Sp = \frac{TN}{(TN + FP)};$$

$$FPR = \frac{FP}{(FP + TN)}; FNR = \frac{FN}{(TP + FN)};$$

$$P = \frac{TP}{(TP + FP)}; F = \frac{2 \times P \times Re}{(P + Re)};$$

$$PWC = \frac{100 \times (FN + FP)}{(TP + TN + FP + FN)}$$

[4] Change Detection 2014 Dataset, http://www.changedetection.net/

Experimental Results

A. Motion U-Net Training Details

- Uses ResNet-18 backbone with pre-trained weights on ImageNet.
- **Training data:** 10,600 (200 x 53) CDnet-2014 (6.6%) frames for training (90/10 split train/ validation).
- **Network parameter:** input size = 320 x 480, Adam optimization, Learning Rate = 1e-4,

  40 epochs, mini-batch size = 8
- **Loss Function:**

$$Loss = w \cdot L_{BCE} + (1 - w) \cdot L_{Tversky}$$

$$L_{BCE} = -\left(y \times \log(p(y)) + (1 - y) \times \log(1 - p(y))\right)$$

$$L_{Tversky}(P, G; \alpha, \beta) = \frac{|PG|}{|PG| + \alpha|P \backslash G| + \beta|G \backslash P|}$$

Experimental Results

B. Experiment on CDnet-2014 Benchmark Videos

TABLE I: Comparison of MU-Net1 and MU-Net2 to top-performing deep learning methods on CDnet-2014

Methods	Overall				
	Rank	Re	PWC	P	F
FgSegNet_v2	1	0.9891	0.0402	0.9823	0.9847
FgSegNet_S	2	0.9896	0.0461	0.9751	0.9804
FgSegNet	3	0.9836	0.0559	0.9758	0.9770
BSPVGAN	4	0.9544	0.2272	0.9501	0.9472
BSGAN	5	0.9476	0.3281	0.9232	0.9339
Cascade CNN	6	0.9506	0.4052	0.8997	0.9209
MU-Net1		0.9277	0.2097	0.9414	0.9147
MU-Net2		0.9454	0.2347	0.9407	0.9369

TABLE II: Training duration and parameters for FgSegNet_v2 and MU-Net

Methods	# of models	GPU	Train Time
FgSegNet_v2	53	GTX 1080 Ti	29 days
MU-Net1	1	GTX 1080 Ti	4 hours
MU-Net2	1	Tesla V100	4.5 hours

Methods	Network Size (# parameters)
FgSegNet_v2	489 M (53 * 9,225,161)
MU-Net1	17.8 M (1 * 17,799,809)
MU-Net2	17.8 M (1 * 17,799,809)

Experimental Results: Generalization test using SBI-2015 Videos

Goal:

1. **Assess generalization capabilities of Motion U-Net**
2. **Assess contribution of the motion and change cues on performance**

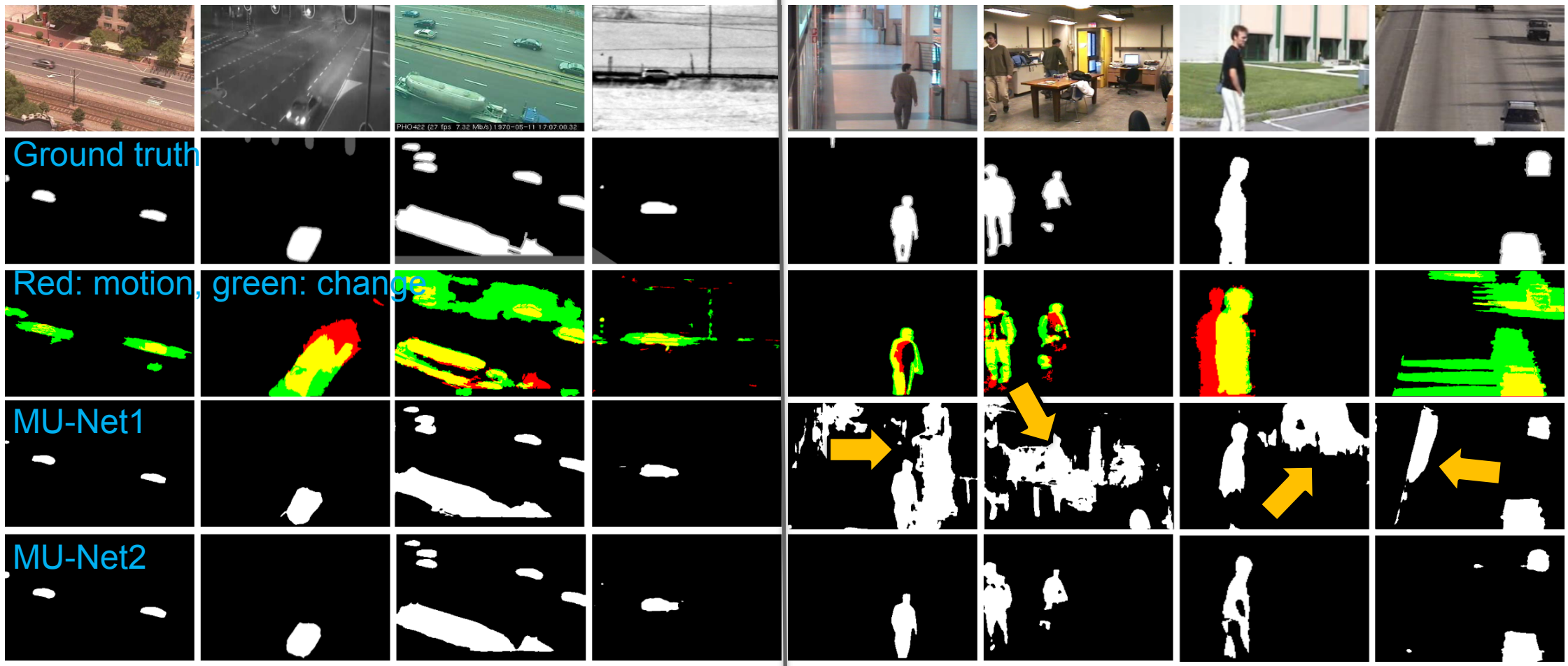
Process:

- Train MU-Net1 and MU-Net2 only on CDnet-2014
- Test CDnet-2014 trained networks on the unseen SBI-2015 videos
- FgSegNet_v2 is trained for each category of CDnet-2014 (11 models).
- Majority Voting fusion used to produce FgSegNet_v2 result

TABLE III: Architecture performance trained only on CDnet-2014 and tested on SBI-2015 (10 videos)

Methods	Overall			
	Re	PWC	P	F
MU-Net1	0.8094	15.8354	0.2881	0.3785
Motion U Change	**0.8360**	11.0881	0.4110	0.5187
MU-Net2	0.7302	**2.6826**	**0.8484**	**0.7625**
FgSegNet_v2 (50%)	0.2419	5.7716	0.8150	0.3519

Experimental Results



CDnet-2014

SBI-2015

Conclusion

- **Motion U-Net (MU-Net):** novel compact U-Net encoder-decoder network structure; integrates complementary multi-cue information (motion, change, and appearance cues) for robust moving object detection performance.
- **Motion and change:** decoupled from network; computed using unsupervised flux tensor and BG subtraction methods.
- **Decoupled, unsupervised motion and change** leads to reduced network complexity, training times, and need for training data.
- Motion U-Net able to learn fusion, object level reasoning, and semantic analysis using only 8% of the CDnet-2014 labeled video frames.
- **Unseen data performance:** great performance improvement on unseen data compared to appearance-only MU-Net1, and the top ranked FgSegNet_v2 (from 37% and 35% to 76%).