Elsevier Editorial System(tm) for Computer Vision and Image Understanding
Manuscript Draft

Manuscript Number: CVIU-19-772

Title: ChangeNet-v2: Semantic change detection with Convolutional Neural Networks

Article Type: Research paper

Keywords: Change Detection; Image Segmentation; Convolutional Neural Networks; Deep Learning; Surveillance; Computational Intelligence

Corresponding Author: Dr. Jayavardhana Gubbi,

Corresponding Author's Institution: University of Melbourne

First Author: K Ram Prabhakar

Order of Authors: K Ram Prabhakar; Akshaya Ramaswamy; Suvaansh Bhambri; Jayavardhana Gubbi; Venkatesh Babu; Balamuralidhar Purushothaman

Abstract: Identifying areas of change between two images of the same scene captured at different times is a fundamental processing step in many image processing applications. Common challenges involved are illumination variation, camera jitter, shadows, and many more. The task becomes even more challenging when the test and reference images are captured at a different time of the day, season, and viewpoint. In different applications, change can manifest in
three types: insertion of an object, removal of an object, and change in its state (including movement, shape, pose or form). The human mind interprets the change by comparing the current status with historical data at the
intelligence level rather than using only visual information. In this paper, we present a deep architecture called ChangeNet-v2 for detecting changes between pairs of images and express the same semantically (label the change).
The main objective is to detect relevant changes rather than detecting background changes, which are irrelevant to the application. A parallel deep Convolutional Neural Network (CNN) architecture for localizing and identifying the changes between image pairs has been proposed. We start with computing the convolutional feature pyramid for both images. Later, the changes are segmented by finding a correlation between them. We benchmark our method on three different datasets: VL-CMU-CD, GSV, and TSUNAMI. Compared with several traditional and other deep learning-based change detection methods, our proposed method achieves higher overall accuracy and f-score in our experiments on all three datasets.

*1. Cover Letter

December 03, 2019

To
Nikos Paragios,
Editor-in-Chief,
Computer Vision and Image Understanding (CVIU)

Dear Dr. Nikos,

We would like to submit a manuscript titled '*ChangeNet-v2: Semantic change detection with Convolutional Neural Networks*' to be considered for publication in the Elsevier Computer Vision and Image Understanding (CVIU).

In this paper, we present a deep architecture called ChangeNet-v2 for detecting changes between pairs of images and express the same semantically (label the change). Common errors in change detection arise due to illumination variations, camera jitter or shadows. The task becomes even more challenging when the test and reference images are captured at a different time of the day, season, and viewpoint. Our main objective in this work is to detect relevant changes rather than detecting background changes, which are irrelevant to the application. We propose a parallel deep Convolutional Neural Network (CNN) architecture, that first computes the convolutional feature pyramid for both images, and then finds correlation between the features to segment the changes. We benchmark our method on three different datasets: VL-CMU-CD [1], GSV [2], and TSUNAMI [2]. Compared with several traditional and other deep learning-based change detection methods, our proposed method achieves higher overall accuracy and f-score in our experiments on all three datasets. This work is an original contribution, and an initial part of this has been published [3] in the proceedings of the UAVision Workshop held in conjunction with the European Conference on Computer Vision (ECCV) 2018.

This work has not been submitted (or under consideration) to any other publication. We believe this work will be of much interest to the readers of CVIU. We have no conflicts of interest to disclose. Please do not hesitate to contact me if you need further information on this submission.

Thank you very much for your consideration.

Sincerely,
Jayavardhana Gubbi,
E-mail: j.gubbi@gmail.com

References

1. Alcantarilla *et al.* "Street-view change detection with deconvolutional networks." Autonomous Robots 42.7 (2018): 1301-1322.

2. Ken Sakurada, and Takayuki Okatani. "Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation." BMVC 2015.

3. Ashley Varghese *et al.* "ChangeNet: a deep learning architecture for visual change detection." ECCV 2018.

ChangeNet-v2: Semantic change detection with Convolutional Neural Networks

K. Ram Prabhakar[a], Akshaya Ramasamy[b], Suvaansh Bhambri[a], Jayavardhana Gubbi[b,*], R. Venkatesh Babu[a], Balamuralidhar Purushothaman[b]

[a]*Department of Computational and Data Sciences, Indian Institute of Science, Bangalore 560012, India*
[b]*Embedded Systems and Robotics, TCS Research and Innovation, Bangalore - 560066, India*

Abstract

Identifying areas of change between two images of the same scene captured at different times is a fundamental processing step in many image processing applications. Common challenges involved are illumination variation, camera jitter, shadows, and many more. The task becomes even more challenging when the test and reference images are captured at a different time of the day, season, and viewpoint. In different applications, change can manifest in three types: insertion of an object, removal of an object, and change in its state (including movement, shape, pose or form). The human mind interprets the change by comparing the current status with historical data at the intelligence level rather than using only visual information. In this paper, we present a deep architecture called ChangeNet-v2 for detecting changes between pairs of images and express the same semantically (label the change). The main objective is to detect relevant changes rather than detecting back-

*Corresponding author:
Email address: j.gubbi@tcs.com (Jayavardhana Gubbi)

ground changes, which are irrelevant to the application. A parallel deep Convolutional Neural Network (CNN) architecture for localizing and identifying the changes between image pairs has been proposed. We start with computing the convolutional feature pyramid for both images. Later, the changes are segmented by finding a correlation between them. We benchmark our method on three different datasets: VL-CMU-CD, GSV, and TSUNAMI. Compared with several traditional and other deep learning-based change detection methods, our proposed method achieves higher overall accuracy and f-score in our experiments on all three datasets.

1. Introduction

Change detection is often used as a stepping stone for high-level scene understanding. In its conventional form, the methods are used for identifying changes in the background by comparing any two consecutive frames or limited to short term temporal analysis spanning a few seconds ([1]). In remote sensing literature, change detection is referred to as surface component alteration that is very useful in automatic land use analysis ([2]). The fact that the satellite images are registered helps in pixel-level change detection tasks that have been successfully extended to object-level change analysis. Some of the critical challenges for visual change detection between any two images include variations in lighting or illumination, contrast, quality, resolution, noise, scale, pose, and occlusion. The first five attributes are experienced in any change detection scenario, but the last three attributes are either not experienced in short term temporal analysis, or can be easily handled by dropping frames.

Most of the methods in the literature that model background pixels to detect change address the first five attributes. In the case of remote sensing, where change detection is widely used, change in scale, pose, and occlusion is rarely seen, and the above methods can be easily deployed with suitable pre-processing. Although these approaches are a part of decision making, it involves low-level image analytics such as foreground-background segmentation. In more complex inferencing using visual input, particularly in pattern recognition and category formation, higher-level cognition is essential. For instance, when two images are being compared that have variations in pose, illumination, color information, and occlusion, the methods in literature often fail due to unregistered images, pose and scale variations as well as occlusions.

The traditional approach to this problem is to detect moving objects from the difference between the current frame and reference frame, often called *background image*, or *background model*. Several methods in the literature follow a two-stage process. The first stage involves developing a model for static or background pixels. In the second stage, the developed background model is used to detect pixels (*foreground* pixels) that deviate from the estimate. The success of such methods relies on the accuracy of the estimated background model. They need to be updated for new and challenging scenes.

Recently, Convolutional Neural Networks (CNN) are used to learn problem-dependent features that outperform traditional methods in foreground extraction. [3] proposed a method to intelligently fuse multiscale CNN features with feature pooling, to learn class-specific foreground extractor. [4] proposed a new CNN based change detection method called CDnet. For training CDnet, the authors curated a new urban change detection dataset called as

VL-CMU-CD. VL-CMU-CD dataset consists of 1362 registered image pairs captured at different time instances over a year. It is larger compared to GSV and TSUNAMI. It contains challenging changes like structural, construction, and natural seasonal changes.

[5] use a combination of superpixel segmentation and pre-trained deep neural network weights to detect changes. Also, they have created a new dataset called TSUNAMI and Google Street View (GSV) for benchmarking change detection algorithms. Both TSUNAMI and GSV contain 100 panoramic image pairs each. These datasets cover changes on surfaces of objects (like changes in billboard signs), structural changes (such as changes in a building structure).

More recently, [6] proposed ChangeNet, a deep neural network-based approach that uses pre-trained CNN features to detect changes. Currently, ChangeNet is the current state-of-the-art in the VL-CMU-CD dataset. However, they have shown to perform lower than [4] in GSV and TSUNAMI dataset. As a whole, there is no single method that outperforms in all the three datasets. Such a method would be able to detect both pixel level and semantic level changes irrespective of the challenges posed in real-life conditions.

Figure 1 shows an example from the VL-CMU-CD change detection dataset, where higher-level inferencing is required to detect the rubbish dumping on the pavement, and the appearance changes are spread throughout the images. Motivated by the above problem, we develop a semantic change detection approach that can be used in various environmental conditions. In this paper, a novel deep learning architecture, called ChangeNet-v2, is pro-

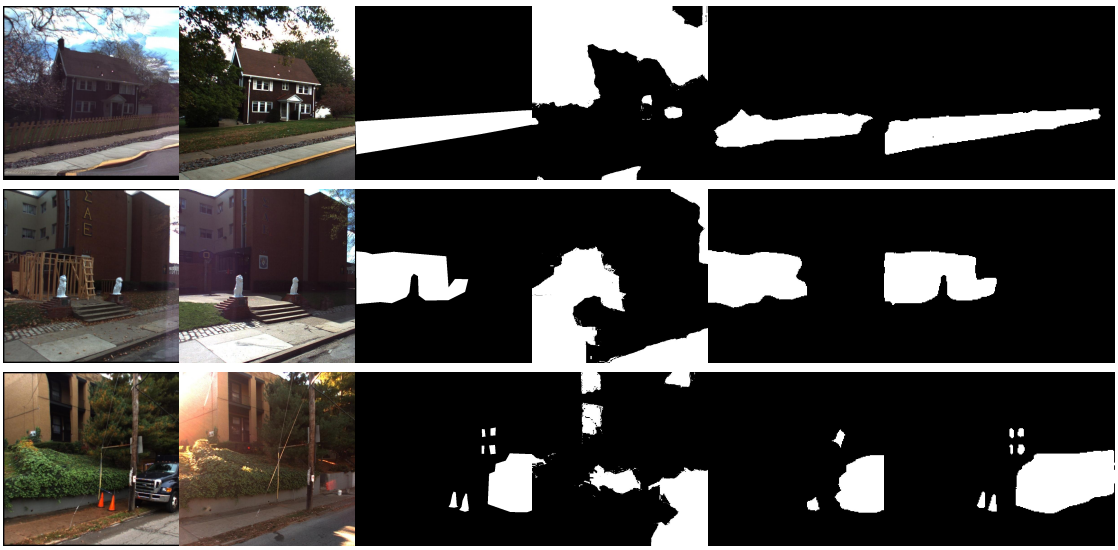
posed for change detection that targets higher-level inferencing. The new network architecture involves extracting features using CNN and combining filter outputs at different levels to localize the change. The detected changes are identified using the same network, and output is an object-level change detection with the label of a changed object.

The proposed ChangeNet-v2 architecture is different from the ChangeNet approach. Similar to ChangeNet model, we use parallel weight tied networks for feature extraction. It ensures both the network learn the same features from the two images. Therefore, the features from both the images can be compared easily. However, compared to ChangeNet, instead of concatenating the image features, we compute the correlation at different levels of convolution layers. Additionally, ChangeNet uses bilinear interpolation to upsample the features; however, in ChangeNet-v2, the upsampling is performed with the help of a learnable deconv layer. Another feature of ChangeNet-v2 is that it combines predictions from different levels of the convolution layer. Such an approach helps the model to capture both coarse and fine details of the object and also makes the approach scene agnostic.

In summary, our contributions are:

- We propose a CNN-based change detection method that is robust to camera motion and various environmental challenges.
- Through experimental evaluation, we show the efficacy of the proposed method in VL-CMU-CD, GSV, and TSUNAMI datasets.

The rest of the paper is organized as follows. In Section 2, we discuss the related works. The proposed method is described in Section 3. The implementation details, evaluation protocols, and results are discussed in



(a) Test (b) Reference (c) Ground truth (d) SuperPixel (e) ChangeNet (f) ChangeNet-v2

Figure 1: Our novel neural network architecture, *ChangeNet-v2*, can accurately segment the semantic changes between test and reference image. It can not only detect the changes with better boundary precision, but can also label them. Qualitative comparison for binary segmentation with SuperPixel ([7]), ChangeNet ([6]) and proposed method for images from VL-CMU-CD dataset.

Section 4. Finally, we conclude the paper in Section 5.

2. Related Works

Change detection has a rich history of related works. It will be exhaustive to describe them all in this paper. Thus, we only focus on popular and recent best-performing change detection algorithms. To model the background, statistical (or parametric) methods using Gaussians are proposed [[8], [9], [10]] to model each pixel as a background or foreground pixel. However, parametric methods are computationally inefficient; to alleviate this problem, various

non-parametric methods [[11], [12], [1]] are proposed. "Vibe" proposed by [11] makes use of randomly selected pixels over time to compute the background model. Additionally, Vibe updates neighboring pixel backgrounds as well whenever a pixel is updated. This iterative process makes Vibe robust to motion artifacts. [1] proposed an improved version of Vibe called "SubSCENE". They use color and local binary pattern information to increase spatial coherency.

One exciting development in semantic change detection was reported by [13] on satellite image processing. Although the work is on simulated data, for the first time, they have reported the perspective of change detection at different inference levels such as object, theme, and database akin to our work. As far as we are aware, this is the first work alluding to different levels in change detection. In a similar work, [14] talk about semantic change detection by adding semantic meaning to the changing area. First, they find changed area using hyper maps, and then add semantic meaning to that changed area. Since the last few years, after deep learning has become the main approach in computer vision, there have been some efforts in the creation of the dataset as well as in building change detection procedures.

[5] was the first such attempt, and they built two data sets with 100 image pairs known as the TSUNAMI dataset and Google Street View (GSV) dataset. These are panoramic images created using street view separated temporally by several days or months. In addition to the creation of a dataset, they proposed a super pixel-based approach that uses the Convolutional Neural Network (CNN) for feature extraction. The low-resolution feature map generated from the CNN network is combined with superpixel segmentation

to get precise segmentation boundaries of the changes. Although deep learning is used in the pipeline, there are many other hyperparameters in the procedure that needs finetuning for different scenarios.

Going one step further, [4] proposes a network called CDnet for finding structural changes in street view video. They create a new dataset of 152 categories with 11 unique object classes called the VL-CMU-CD dataset. To create nearly registered images, they use visual simultaneous localization and mapping (SLAM) to get the 3D point cloud and then project the points onto a 2D reference plane after determining the reference pose. It is a pixel-level change detection approach and uses contraction and expansion layers for pixel-level classification. The contraction block creates data representation. In this process, it stores max-pooling output for later use in the expansion network. The expansion block has been used for improving change localization.

Apart from deep learning approaches, a multiscale superpixel approach for drone image analysis has been proposed by [7] with limited success on VL-CMU-CD dataset.

In the recent past, there has been quite a good amount of success in pixel-level image analysis using deep architecture. [15] proposed a new architecture for predicting surface normal that is useful in 2D-3D alignment. They use the pre-trained VGG-16 network for feature extraction followed by three layers of fully connected layers for predicting surface normal for every pixel. [16] generalized their earlier work and created PixelNet architecture and demonstrated semantic segmentation and edge detection in addition to surface normal estimation using an extended VGG-16 network. Such work

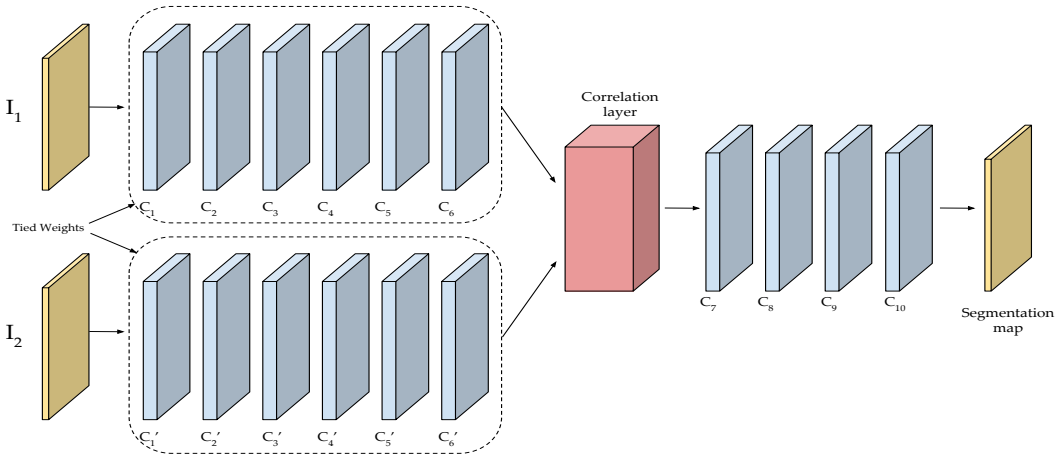


Figure 2: Overview of proposed Correlation model. Test image (I_1) and reference image (I_2) are passed as input to two parallel CNN architecture with with tied weights. Correlation between C_6 and C_6' estimates the likelihood of finding similar feature in a fixed neighbourhood. Convolutional layers C_7 to C_{10} convert the correlation map into binary segmentation output.

has demonstrated that CNN can learn pixel-level information in addition to its success in image categorization. Similar work has been extended for regions of interest where they propose a new network for simultaneously predicting human eye fixations and segmenting salient objects. In addition to single image pixel analysis, there has been some recent work, including the similarity between two images or signal pairs. [17] propose a Siamese CNN network for checking whether two handwritten texts are written by the same person or not. Both the inputs are encoded with the same network, and then concatenated output is fed to a two-class classifier to determine whether handwriting is the same or not.

With the recent success of deep learning models in pixel-level image analysis, we are motivated to solve the hard problem of change detection using a

deep network. VL-CMU-CD dataset is our target as the scene pairs are complex and taken at different view angles, illumination, and seasons as well. It has 11 different classes of structural changes like construction-maintenance, bin on the pavement, new signboards, traffic cone on the road, vehicles, etc., including the background. To the best of our knowledge, it is a novel architecture for visual change detection, mainly resulting in scene labels that can be viewed as semantic change detection. We further train the network to determine the category of change in addition to the changing area. Both the tasks happen within the network and involve single training. Most of the background information is irrelevant in our case since those changes could be due to season, illumination, or viewpoint variation. It mainly looks for changes at the object level as compared to [4]. The model input is the test and reference images. The output is detection, localization, and categorization of the changed region. It mainly answers the following three questions in the presence of seven variations, which have been discussed earlier: is there any change? If yes, what is the change? Furthermore, where is the change in the image?

3. ChangeNet-v2

In this section, we introduce our proposed CNN based foreground extraction method and its variation. The input to the model are two images: dynamic *test* frame and a static *reference* frame. Both the frames are passed onto two parallel CNN architectures with tied weights to extract similar features from both of them. The pixel-level similarity between features of test and reference images is computed using a correlation layer. The similar-

ity features are passed onto a further set of convolution layers to provide a probability score of belonging to background or foreground.

3.1. Correlation model

In Figure 2, we show our simple correlation model (named *Corr-model* further). Let I_1 be test frame with moving foreground object, and I_2 be reference frame static background. The objective is to segment the moving foreground object from the test frame. I_1 and I_2 are passed as input to a Siamese network, which consists of six convolutional layers with shared/tied weights. These features contain semantic information about both the test and the reference frame. One simplistic way to segment out the moving object is to subtract both features. However, that holds only for the case where both test and reference frames are registered. To solve this problem, we make use of the correlation layer to compute pixel similarity.

A correlation layer computes patch comparison between features maps f_1 and f_2. The correlation between two patches p_1 (from f_1) and p_2 (from f_2) centered at a (x, y) is defined as follows:

$$C(p_1, p_2) = \sum_{o \in [-s,s] \times [-s,s]} \langle f_1(p_1 + o), f_2(p_2 + o) \rangle \tag{1}$$

where s is the size of the support window sampled around the pixel. Bigger the value of s, higher the robustness towards false matching. However, that also leads to more computation steps. To find the relative displacement of p_1, the correlation operation is applied to all pixels in a search area $T \times T$ of f_2 centered at (x, y). This results in an output of T^2 correlation values for every pixel in f_1. The computed correlation map is passed onto set of convolutional layers to obtain binary segmentation mask.

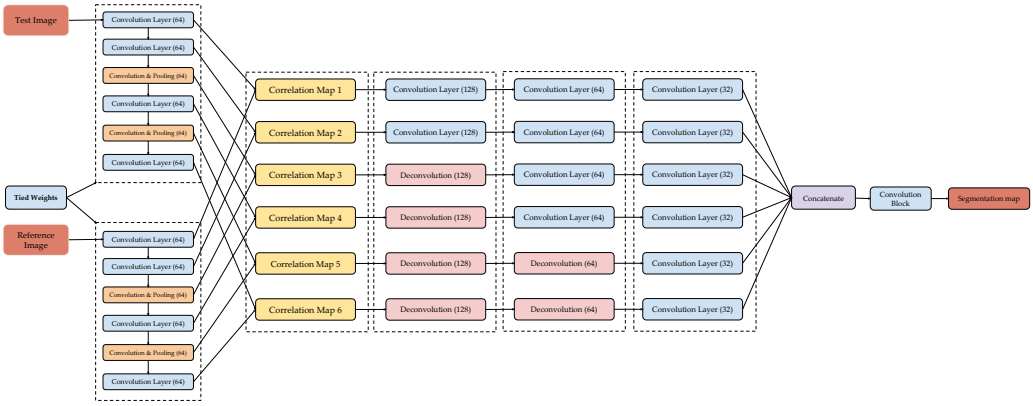


Figure 3: Overview of proposed ChangeNet-v2 model. In contrast to the model in Fig. 2, we compute the correlation between all feature maps in the tied-weight encoder. Also, as the tied-weight encoder architecture has max-pooling in later layers, their feature maps are deconvolved to obtain feature map with the same size.

To find the right value for s, we did ablation with two settings: pixel correlation (termed as $corr - model(pixel)$ in Table 1) and patch correlation (termed as $corr - model(patch)$ in Table 1) with $s = 5$. From Table 1, we find that patch correlation does not offer enough improvement over pixel correlation for the extra computation.

Further, we also did ablation on the suitable feature maps to apply correlation, as shown in Table 1. Along with correlation computed at the sixth convolutional layer, we append correlation computed at the first and second convolutional layer as well. However, appending such correlation maps offer minimal improvement, suggesting the fact that the sixth convolutional layer has enough semantic information for foreground extraction.

3.2. ChangeNet-v2 model

Motivated by the success of pyramidal feature pooling in CNNs, we modify previous architecture to extract pyramidal convolutional features (Figure 3). The Siamese network from the previous model is modified to include two max-pooling layers to condense maximum spatial features. Also, we compute correlation for all feature maps in the Siamese architecture. However, as the features of later layers in tied weights (Siamese) network have different resolutions, we upsample them to the same resolution with the help of transposed convolution layers.

4. Experiments

4.1. Network Implementation

The Siamese network consists of six convolutional Layers and two Max-pool layers (with stride=2 for correlation-deconv model). Each convolutional layer consist of a convolution with kernel size = 3 with Batch Normalization and Leaky ReLU activation (with *alpha* = 0.01). For corr-deconv model, we use transpose-convolution layer (with kernel size=3, stride =2) to get the feature maps of a size equivalent to the largest correlation map. The concatenated correlation maps are passed through four convolutional layers with kernel size = 5 and Leaky ReLU. Finally, the output segmentation map is obtained with the Sigmoid activation function. The network is trained in an end-to-end fashion with binary cross-entropy loss between predicted and ground truth segmentation map. We use Adam optimizer with learning rate of 0.0001 and $\beta_1 = 0.9$ and $\beta_1 = 0.999$ to optimize the loss function.

4.2. Datasets and protocols

We train and evaluate our method in three publicly available datasets: VL-CMU-CD ([4]), GSV, and TSUNAMI ([5]). VL-CMU-CD dataset consists of 1187 image pairs in total. We follow the dataset split used in ChangeNet ([6]) for our experiments as well. The dataset is split into a ratio of 70:15:15% for training, validation, and testing. Similarly, GSV and TSUNAMI datasets consist of 100 image pairs each, out of which 70 image pairs were used for training and 15 image pairs each for validation and testing. We perform a quantitative evaluation on all three datasets using five-fold cross-validation. For training the proposed model, we computed cross-entropy loss between ground truth and predicted output with class rebalancing weights. We implemented our model in Tensorflow ([18]) installed on a workstation with Intel Xeon at 3.50 GHz CPU with 32GB RAM and an NVIDIA Titan X GPU card.

The three datasets and related methods in the literature focus on binary classification. That is, the final output is to detect change or no-change. We refer to this scenario as binary for the rest of the paper. We are also interested in labeling the object after the change is detected. We call this scenario multiclass. Currently, the system is built for multiclass classification of 10 commonly appearing objects in VL-CMU-CD dataset: barrier, bin, construction, person/bicycle, rubbish bin, sign board, traffic cone, and vehicle.

Table 1: Comparison with State-of-art methods in VL-CMU-CD dataset.

Model	f-score
SuperPixel	0.1567
CDNet	0.5802
ChangeNet	0.8005
Corr-model (Pixel)	0.9222
Corr-model (Patch)	0.9270
Corr-model (2^{nd} and 6^{th} layer)	0.9277
Corr-model (1^{st} and 6^{th} layer)	0.9041
ChangeNet-v2 Model	**0.9385**

4.3. Results

4.3.1. Evaluation metrics

We evaluate the performance of the proposed method with state-of-the-art methods using standard F1 evaluation metrics. Also, we evaluated proposed method on three datasets in Table 7 using following standard metrics: Accuracy, Precision, Recall, F1 score, mean Intersection over Union (mIoU), Matthew's correlation coefficient (MCC), Sensitivity, Percentage of Wrong Classification (PWC), Specificity, False Positive Rate (FPR) and False Negative Rate (FNR).

4.3.2. Quantitative comparison

We compare our proposed ChangeNet-v2 method with three state-of-the-art methods in VL-CMU-CD dataset: SuperPixel ([7]), CDNet ([4]) and ChangeNet ([6]). The results are shown in Table 1. Compared with

ChangeNet, the proposed method offers over 14% improvement in f-score. The improvement in accuracy is attributed to the fact that the proposed model is robust enough for image misalignment. Also, it is observed that corr-deconv model performs better than corr-model. Thus, for further evaluation in other datasets, we report only the performance of the corr-deconv model.

Table 2: Analysis of ChangeNet-v2 results at class level on VL-CMU-CD data set. Miscellaneous class has been excluded from the table as all the values were 0.

Classification	Class(→) Metric(↓)	Barrier	Bin	Construction	Other objects	Person/ Bicycle	Rubbish bin	Sign board	Traffic cone	Vehicle
Pixel-based	Precision	0.74	0.76	0.90	0.67	0.84	0.56	0.78	0.67	0.92
	Recall	0.70	0.72	0.85	0.65	0.79	0.50	0.69	0.60	0.88
	f-score	0.72	0.74	0.87	0.66	0.81	0.53	0.73	0.63	0.90
Object-based	Precision	1.00	0.97	0.88	1.00	1.00	0.96	1.00	1.00	1.00
	Recall	0.78	1.00	1.00	0.63	1.00	1.00	0.87	0.58	0.97
	f-score	0.87	0.98	0.94	0.78	1.00	0.97	0.93	0.73	0.98

The ChangeNet-v2 architecture was specifically designed keeping VL-CMU-CD dataset in mind due to its complexity. In order to validate the architecture, the five-fold cross-validation was conducted. The results are as

Table 3: Average results of 5-fold cross validation for binary and multi-class categories in VL-CMU-CD dataset.

	Accuracy	Precision	Recall	f-score
Binary	99.2	93.7	93.9	93.8
Multi-class	78.5	76.0	71.3	73.4

Table 4: Performance metrics for ChangeNet-v2.

Metrics(→) Method(↓)	Pixel accuracy	Mean pixel accuracy	Mean IOU	Frequency weighted IOU
ChangeNet-v2	99.2	84.77	88.3	97.8

Table 5: The quantitative comparison of our method with other approaches for FPR = 0.1 and FPR = 0.01. The best scores are highlighed in **bold** and the second best in blue color.

	FPR = 0.1			FPR = 0.01		
Metrics (→) Methods (↓)	Precision	Recall	f-score	Precision	Recall	f-score
Super-pixel	0.17	0.35	0.23	0.23	0.12	0.15
CDnet	0.40	0.85	0.55	0.79	0.46	0.58
ChangeNet	0.79	0.80	0.79	0.80	0.79	0.79
ChangeNet-v2	**0.93**	**0.93**	**0.93**	**0.90**	**0.94**	**0.93**

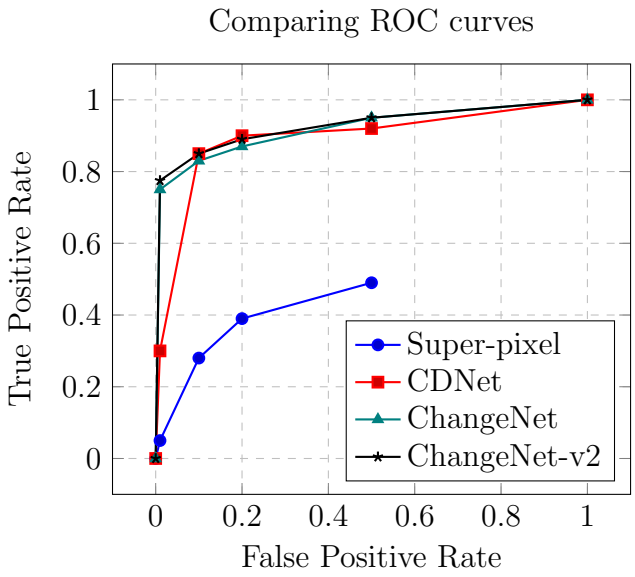


Figure 4: ROC and TPR-FPR curve for binary class.

Table 6: Comparison with state-of-the-art methods in VL-CMU-CD, GSV and TSUNAMI dataset with f-score.

Datasets (→) Methods (↓)	VL-CMU-CD	GSV	TSUNAMI
Super-pixel	0.15	0.26	0.38
CDNet	0.58	0.61	0.77
ChangeNet	0.80	0.45	0.73
ChangeNet-v2	**0.93**	**0.68**	**0.86**

Table 7: Performance metrics of ChangeNet-v2 for binary classification (change or no-change) on different datasets.

Metric (→) / Dataset (↓)	Accuracy	Precision	Recall	F1	mIoU	MCC	Sensitivity (TPR)	PWC	Specificity (TNR)	FPR	FNR
TSUNAMI	0.921	0.845	0.881	0.863	0.827	0.808	0.881	0.078	0.938	0.062	0.119
GSV	0.840	0.629	0.759	0.688	0.665	0.587	0.759	0.159	0.865	0.135	0.241
VL-CMU-CD	0.992	0.937	0.939	0.938	0.883	0.934	0.939	0.765	0.995	0.004	0.060

shown in Table 3, and a healthy average f-score of 88.8% was obtained for binary classification and 73.4% for multiclass classification. To further confirm its performance, multi-scale SuperPixel [7], CDnet [4] and ChangeNet [6] were compared to the proposed architecture. In order to make a fair comparison, the results of binary classification (change or no change) of all the methods are compared by converting our class-based output into binary form. The predicted change map of baseline approaches and our method are shown in Figure 1. Each sample exhibits different lighting and seasonal condition. The first column is the test image, which is compared against the reference image in the second column. The ground truth change map is shown in the third column. The fourth and fifth columns are the change detection results of SuperPixel and ChangeNet methods. The results of ChangeNet-v2 is shown in the last column.

As shown in Figure 5, ChangeNet-v2 performs better than other approaches both in terms of the output as well as change class labeling. It gives a better performance in terms of accuracy and precision. Compared to other approaches, it gives additional information like what is the structural changes in the scene. In other words, our approach can locate the change

in the scene as well as what the change is. ChangeNet-v2 performs well even though the background between image pair is different due to seasonal alterations and lighting conditions. For example, the image pair in row 2 are taken at different lighting conditions, and ChangeNet-v2 was able to detect the changed area. An example of multiple changes in the same scene is depicted in row 4, where a vehicle and a signboard are depicted as change. ChangeNet-v2 can identify both of them accurately. Results in row 5 show the performance of ChangeNet-v2 when images are captured at a different seasonal condition. Reference image is taken in, and it has snow in the background. The same case applies to row 6 as well. The model performed well in this case, and it could detect and locate the rubbish bin. Since we approached the change detection problem at the semantic level, we could mitigate irrelevant background information and reduce false alarms, if any.

The quantitative performance of our method is evaluated in two aspects. The first aspect is how accurately it localized the change. Once it localized the change, what is the pixel labeling accuracy? Mainly, Intersection over Union (IoU) and pixel accuracy metrics are used for evaluating the performance. We considered 11 classes, including background for this performance measurement. The model is evaluated with 177 image pairs, and the results are generated. The performance metric for ChangeNet is given in Table 4. We achieved 98.4% pixel-level accuracy, and 83.93% mean pixel accuracy. In other words, 98.4% pixels are classified as change correctly. In that, 83.93% pixels are classfied correctly per class basis. Also, we achieved 76.35% IoU. It compares the ground truth and predicted changed area on a per-class basis. IoU is changed to 96.3% once we assigned the weights to class IoU based on

their appearance frequency.

Table 2 shows the results of ChangeNet-v2 in identification of class-based change. Other than traffic cone and other objects, all other classes resulted in a f-score of over 0.8 for object-level change detection. At the pixel level, small objects, including traffic cone, rubbish bin, and signboard, resulted in lower f-scores. Table 5 shows quantitative comparison of ChangeNet-v2 with ChangeNet, CDnet and Super-pixel methods for two different false-positive rates of 0.1 and 0.01. As can be seen, ChangeNet-v2 outperforms both the methods with an impressive f-scores.

Figure 4 shows the Receiver Operator Characteristic (ROC) curve for binary classification. All the classes except the background are considered as a logical one. ChangeNet-v2 resulted in a steep ROC curve with a maximum true positive rate and minimum false positive rate. The area under the ROC curve, i.e., AUC is 99.4%. Finally, the performance of the three different methods on three different datasets is presented in Table 6. The proposed Changenet-v2 model performs better than compared methods in all three datasets in terms of f-score.

Detailed results of ChangeNet-v2 on the three datasets with eleven metrics are presented in Table 7. As can be seen, the results on VL-CMU-CD dataset are very high with good performance on the TSUNAMI dataset. There is a drop in GSV performance. The drop in performance on the GSV dataset is attributed to the way the ground truth is created in these datasets. ChangeNet-v2 focuses on structural changes, but GSV ground truth represents cars on the road as change. Hence, the overall performance seems to dip. The results of ChangeNet-v2 on TSUNAMI and GSV dataset are shown

in Figures 6 and 7.

4.3.3. Qualitative comparison

We show the results generated by SuperPixel, ChangeNet and our method for images from GSV, TSUNAMI and VL-CMU-CD datasets in Fig. 1, 5, 6, and 7.

5. Conclusion

In this paper, we have presented ChangeNet-v2, a novel CNN-based method for detecting changes from a pair of images. We handle the changes at various semantic levels, from simple structural change to illumination or seasonal changes, by using low-level to high-level convolutional features extracted at different depths in our encoder. Additionally, we make use of the correlation layer to handle the misregistration between input pairs. Hence, the burden of having a perfect pixel-to-pixel alignment is alleviated. Through extensive evaluation on three different datasets, we have shown that our proposed ChangeNet-v2 method offers better accuracy over existing state-of-the-art methods. In particular, ChangeNet-v2 offers 13% boost in f-score at VL-CMU-CD dataset.

References

[1] P.-L. St-Charles, G.-A. Bilodeau, R. Bergevin, Subsense: A universal change detection method with local adaptive sensitivity, IEEE Transactions on Image Processing 24 (1) (2015) 359–373.

[2] M. Hussain, D. Chen, A. Cheng, H. Wei, D. Stanley, Change detection from remotely sensed images: From pixel-based to object-based approaches, ISPRS Journal of photogrammetry and remote sensing 80 (2013) 91–106.

[3] L. A. Lim, H. Y. Keles, Learning multi-scale features for foreground segmentation, arXiv preprint arXiv:1808.01477 (2018).

[4] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, R. Gherardi, Street-view change detection with deconvolutional networks, Autonomous Robots 42 (7) (2018) 1301–1322.

[5] K. Sakurada, T. Okatani, Change detection from a street image pair using cnn features and superpixel segmentation., in: BMVC, 2015, pp. 61–1.

[6] A. Varghese, G. Jayavardhana, R. Akshaya, P. Balamuralidhar, Changenet: A deep learning architecture for visual change detection, in: European Conference on Computer Vision Workshops (ECCVW), IEEE, 2018.

[7] J. Gubbi, A. Ramaswamy, N. Sandeep, A. Varghese, P. Balamuralidhar, Visual change detection using multiscale super pixel, in: Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on, IEEE, 2017, pp. 1–6.

[8] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, Pfinder: Real-time tracking of the human body, IEEE Transactions on Pattern Analysis & Machine Intelligence (7) (1997) 780–785.

[9] C. Stauffer, W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in: Computer Vision and Pattern Recognition, IEEE, 1999, p. 2246.

[10] M. S. Allili, N. Bouguila, D. Ziou, A robust video foreground segmentation by using generalized gaussian mixture modeling, in: Fourth Canadian Conference on Computer and Robot Vision, IEEE, 2007, pp. 503–509.

[11] O. Barnich, M. Van Droogenbroeck, Vibe: A universal background subtraction algorithm for video sequences, IEEE Transactions on Image processing 20 (6) (2011) 1709–1724.

[12] M. Hofmann, P. Tiefenbacher, G. Rigoll, Background segmentation with feedback: The pixel-based adaptive segmenter, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, IEEE, 2012, pp. 38–43.

[13] A. Gressin, N. Vincent, C. Mallet, N. Paparoditis, Semantic approach in image change detection, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, 2013, pp. 450–459.

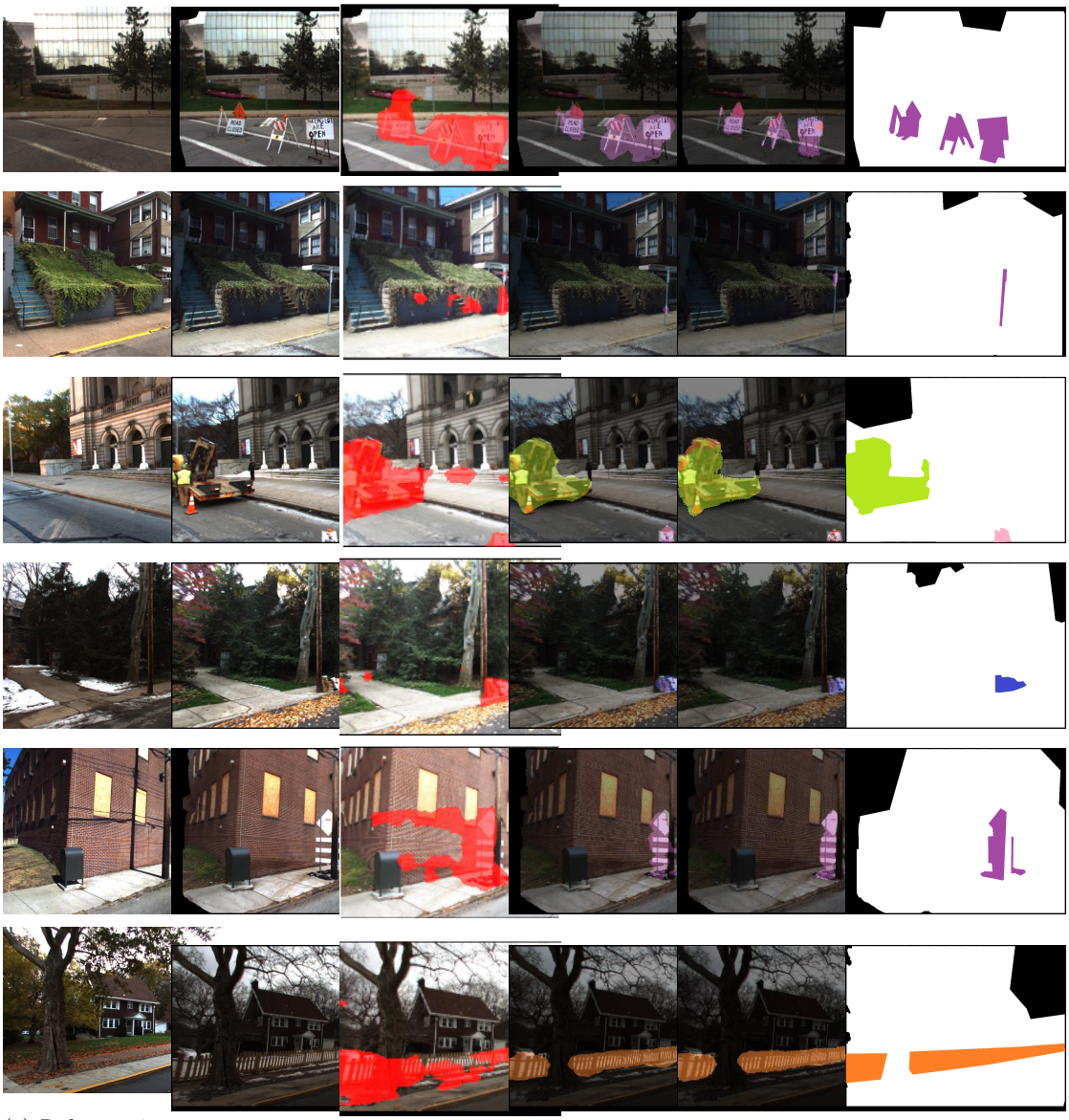
[14] H. Kataoka, S. Shirakabe, Y. Miyashita, A. Nakamura, K. Iwata, Y. Satoh, Semantic change detection with hypermaps, arXiv preprint arXiv:1604.07513 2 (4) (2016).

[15] A. Bansal, B. Russell, A. Gupta, Marr revisited: 2d-3d alignment via surface normal prediction, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5965–5974.

[16] A. Bansal, X. Chen, B. Russell, A. Gupta, D. Ramanan, Pixelnet: Representation of the pixels, by the pixels, and for the pixels, arXiv preprint arXiv:1702.06506 (2017).

[17] W. Du, M. Fang, M. Shen, Siamese convolutional neural networks for authorship verification, 2017.

[18] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning., in: OSDI, Vol. 16, 2016, pp. 265–283.



(a) Reference image (b) Test image (c) CDnet (d) ChangeNet (e) ChangeNet-v2 (f) Ground truth

Figure 5: Qualitative comparison for multi-class segmentation with CDnet, ChangeNet and proposed method for images from VL-CMU-CD dataset.

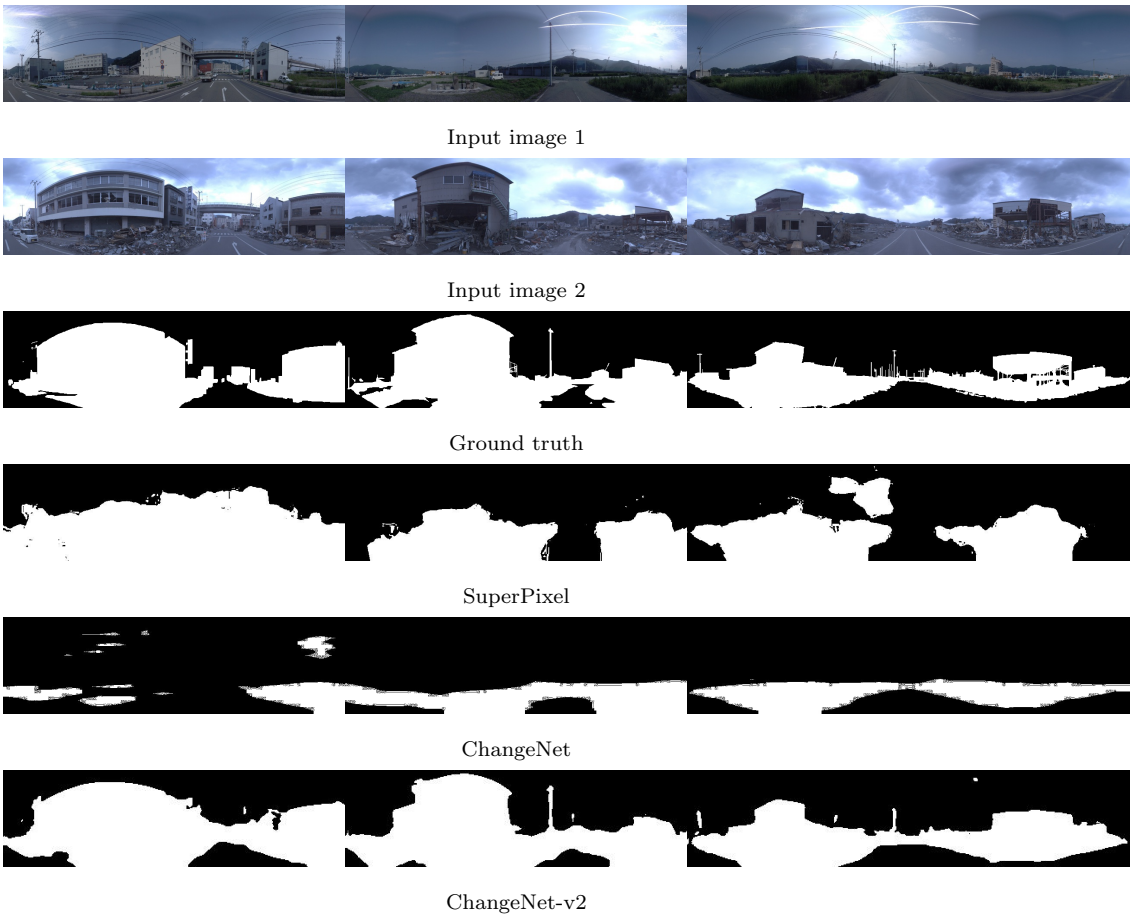


Figure 6: Qualitative comparison with SuperPixel ([7]), ChangeNet ([6]) and proposed method for images from TSUNAMI dataset.

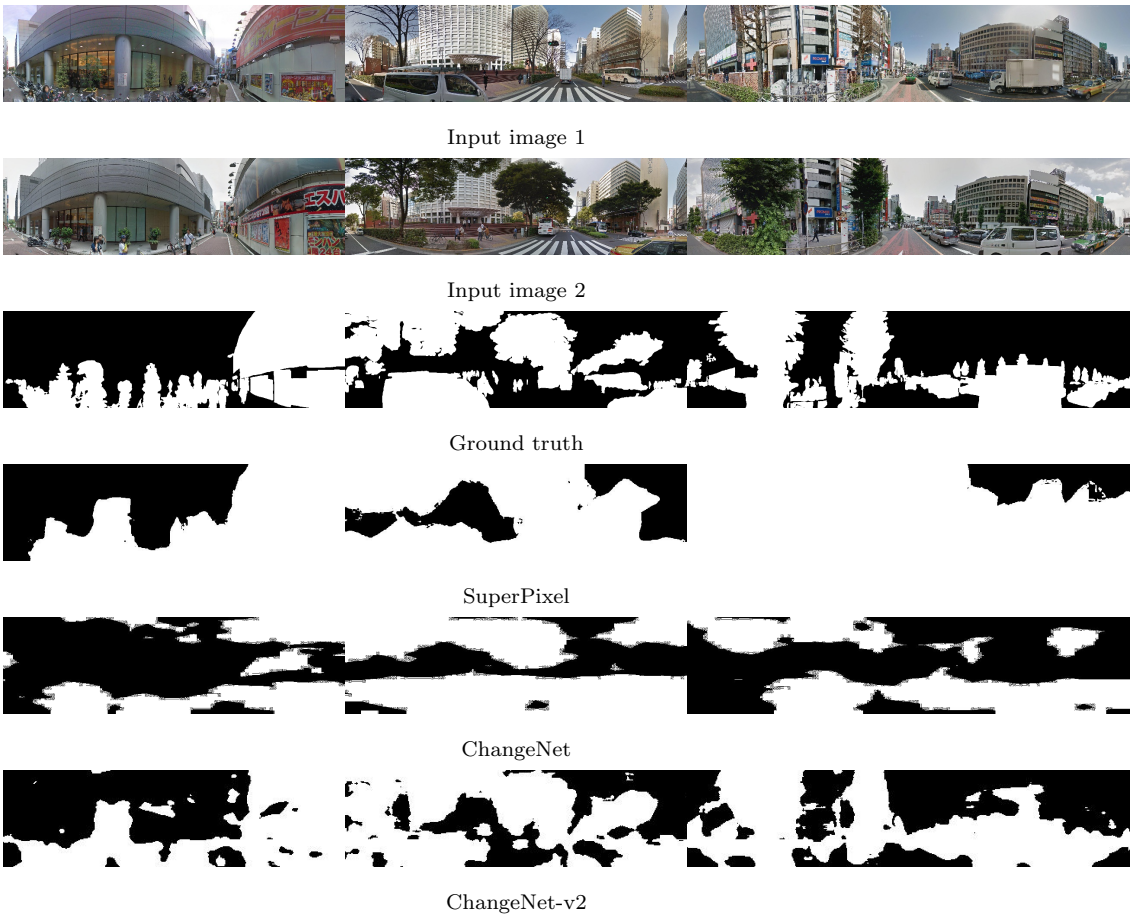


Figure 7: Qualitative comparison with SuperPixel ([7]), ChangeNet ([6]) and proposed method for images from GSV dataset.

Click here to download high resolution image

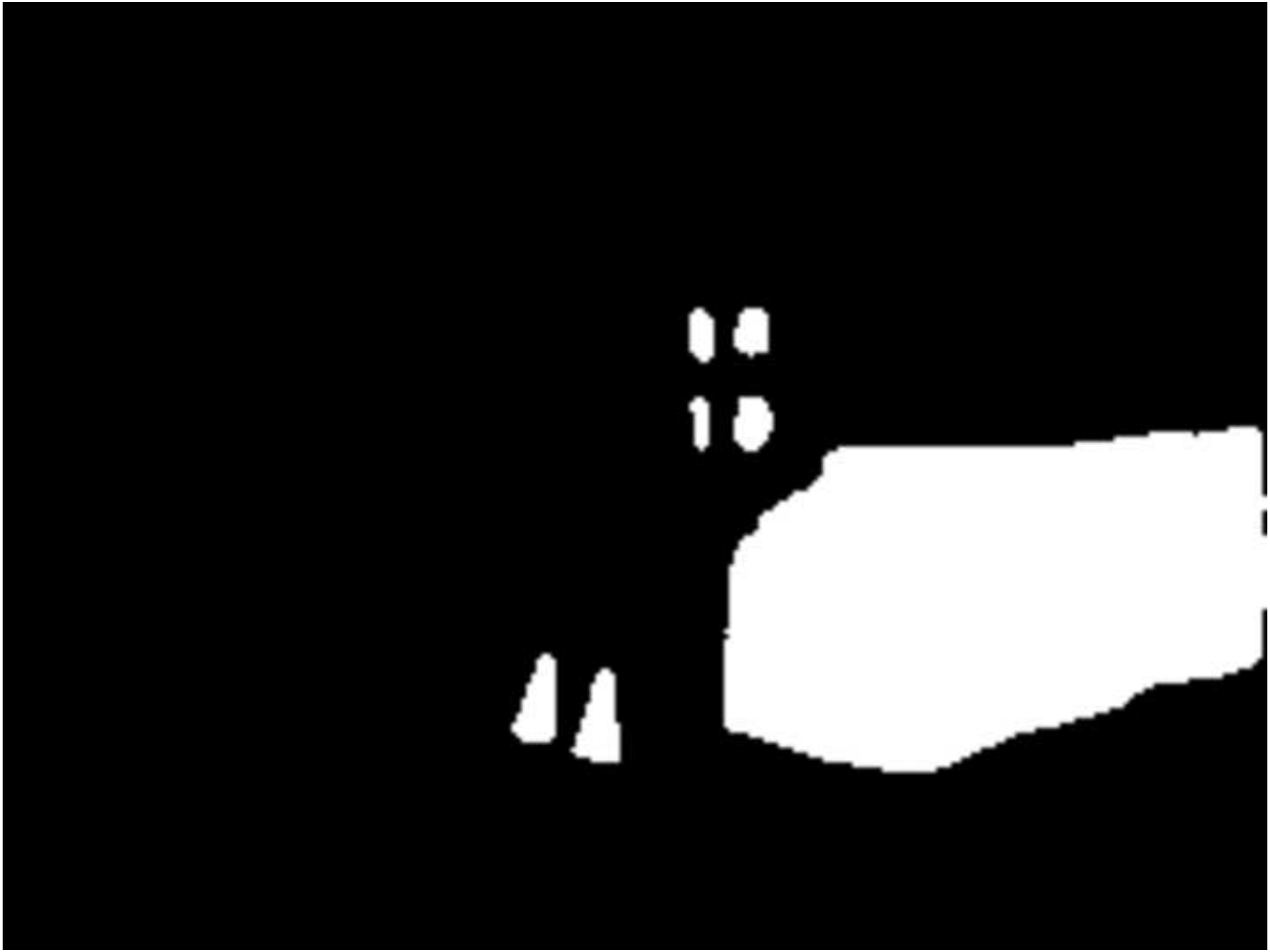
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

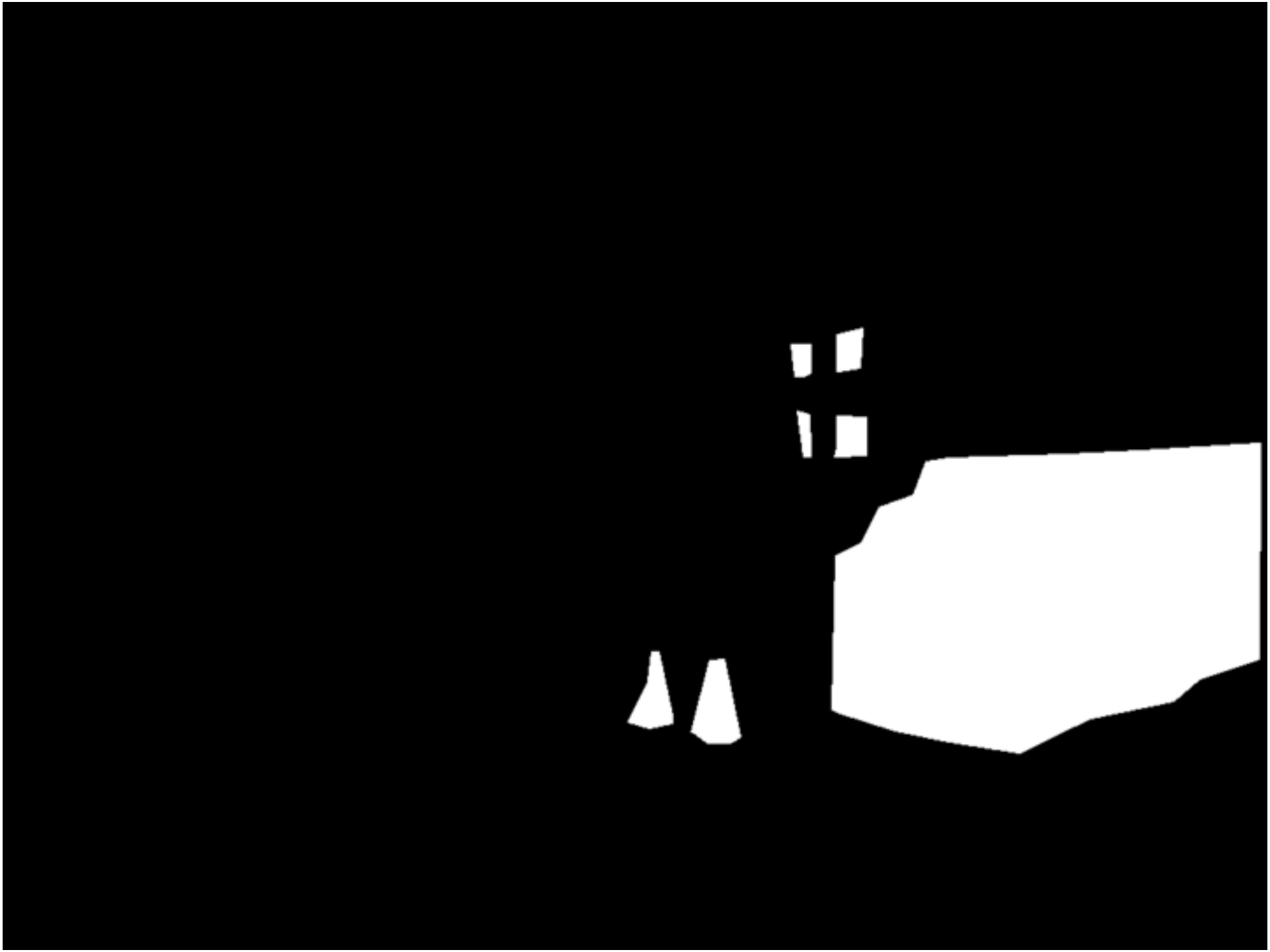


Click here to download high resolution image

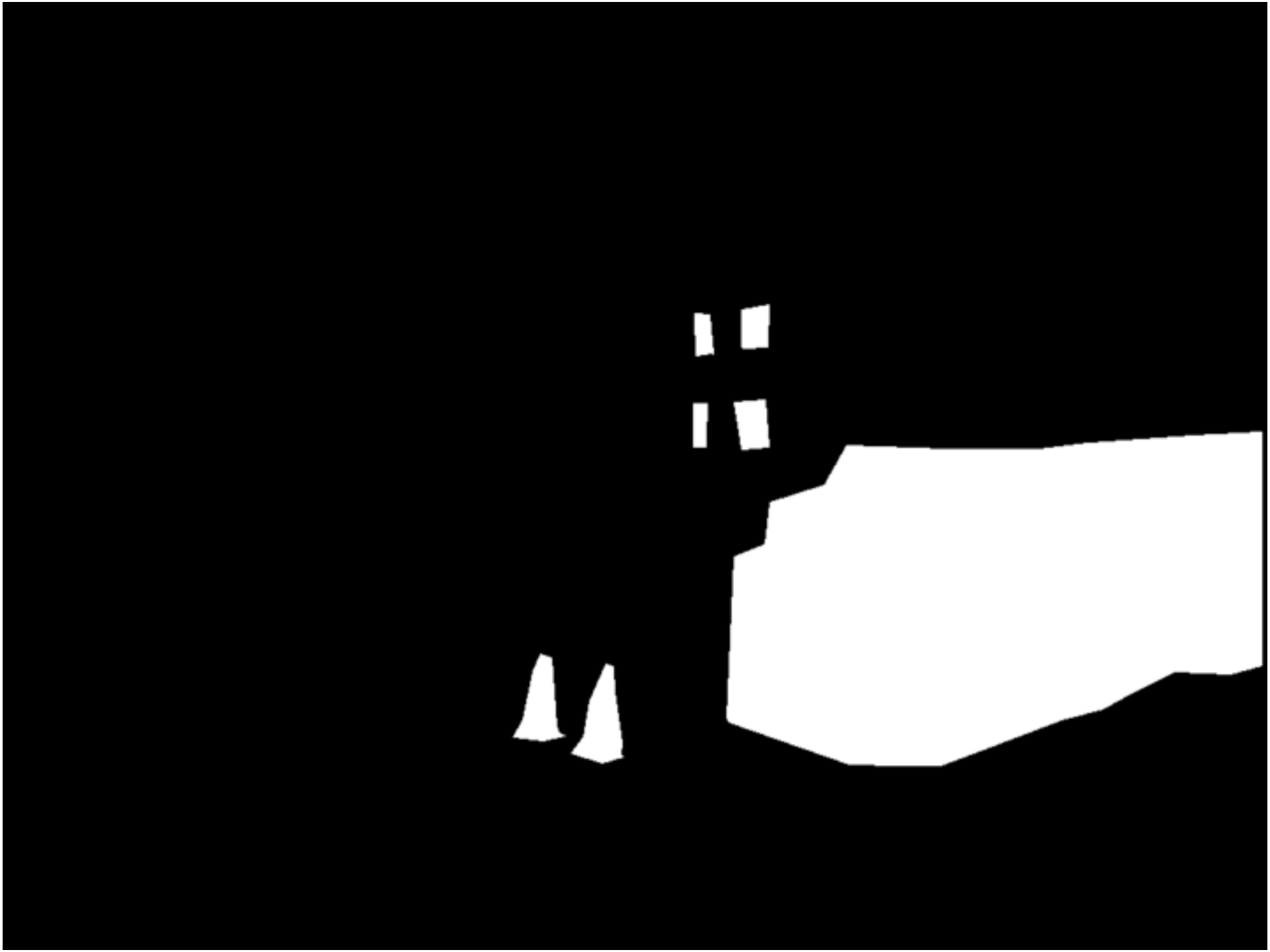
Click here to download high resolution image

Click here to download high resolution image

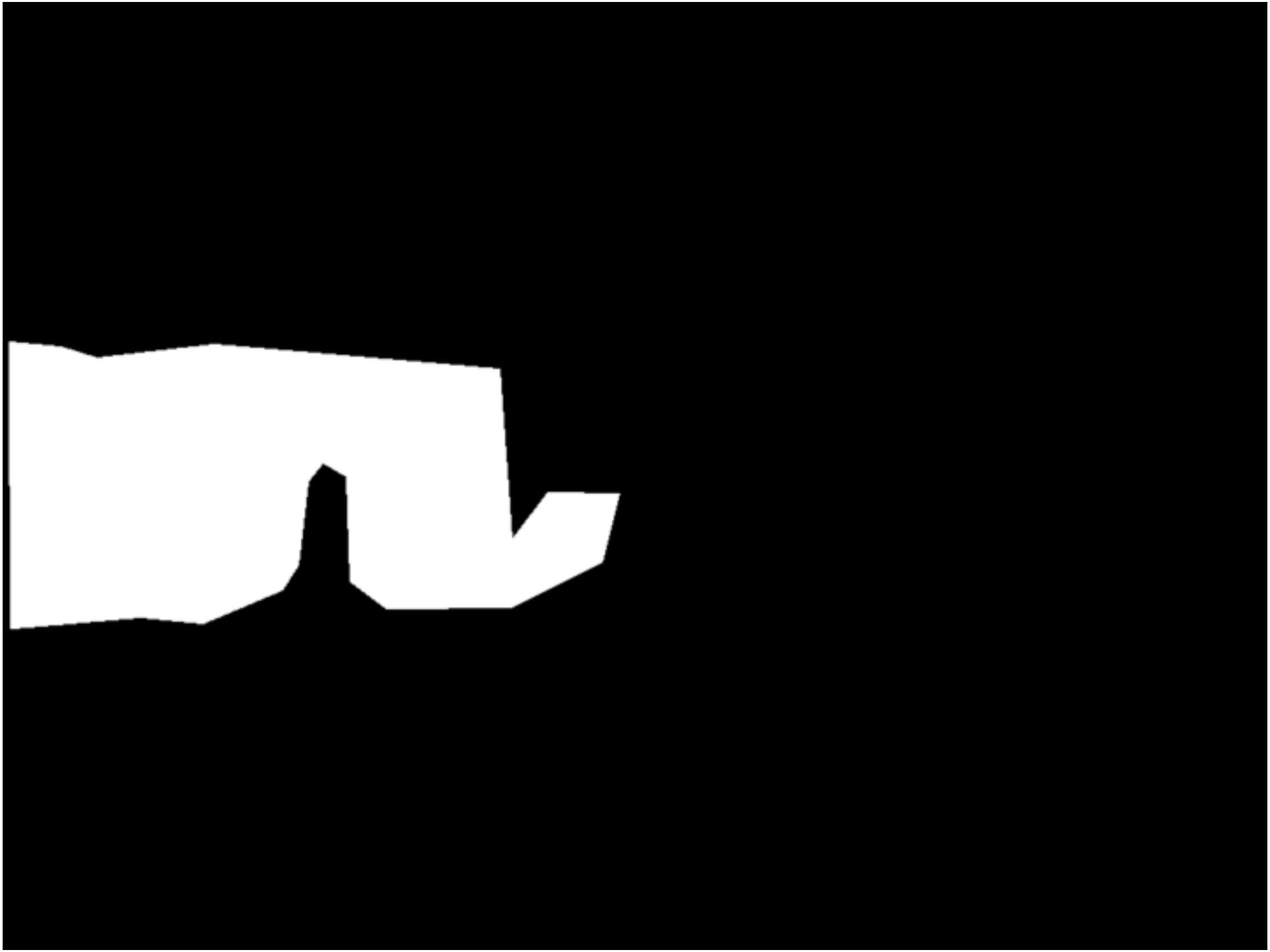
Click here to download high resolution image



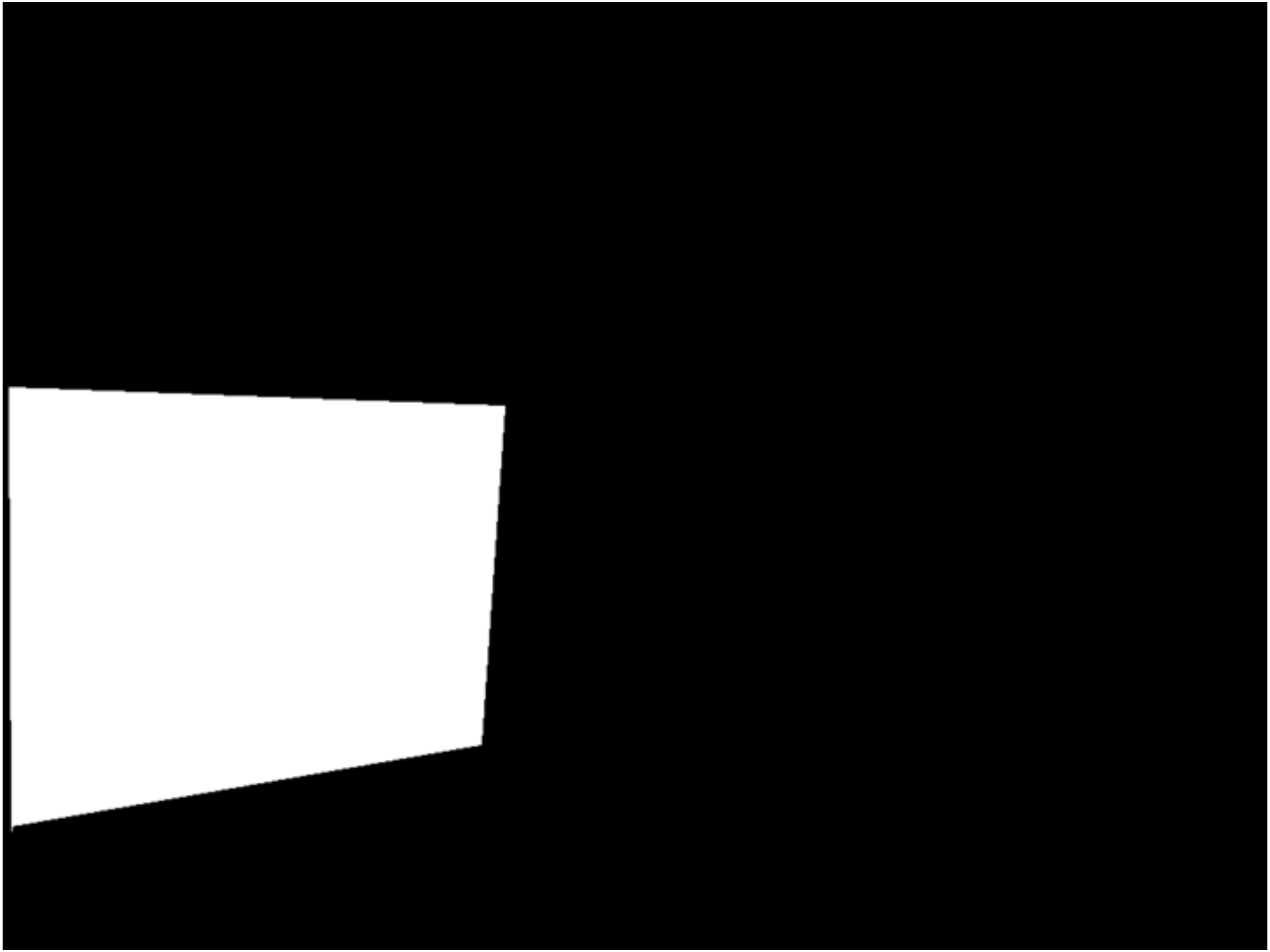
Click here to download high resolution image



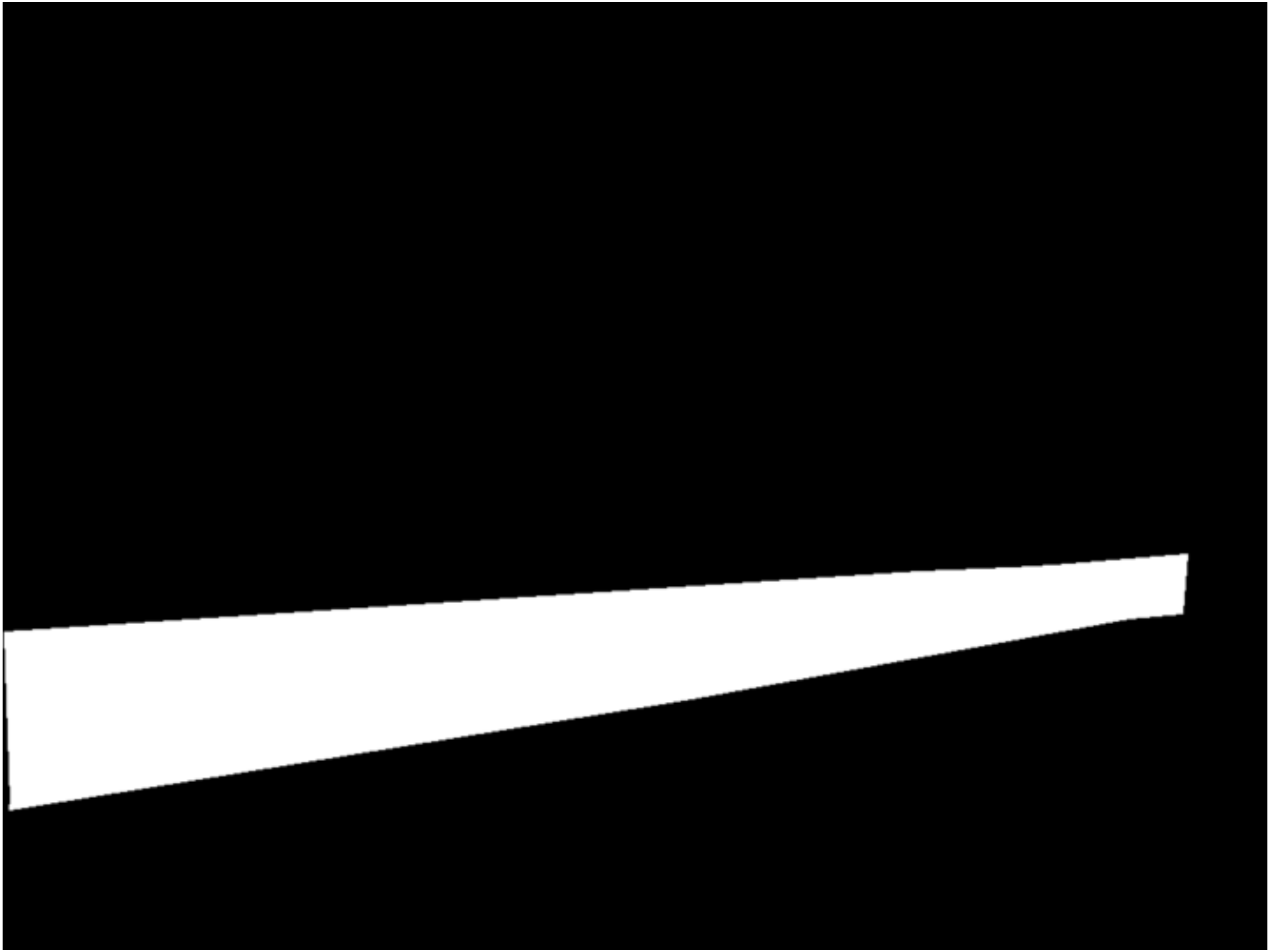
Click here to download high resolution image



Click here to download high resolution image



Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

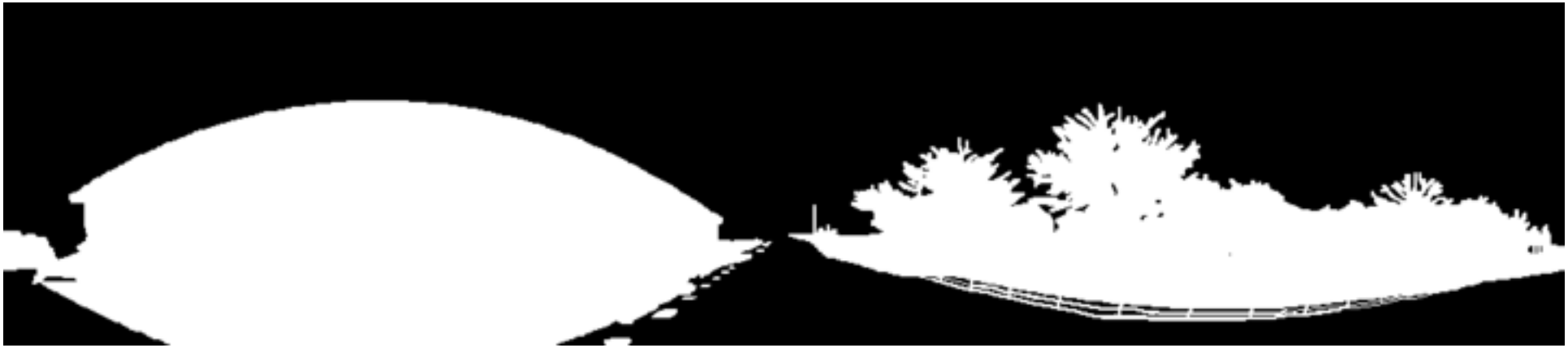
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

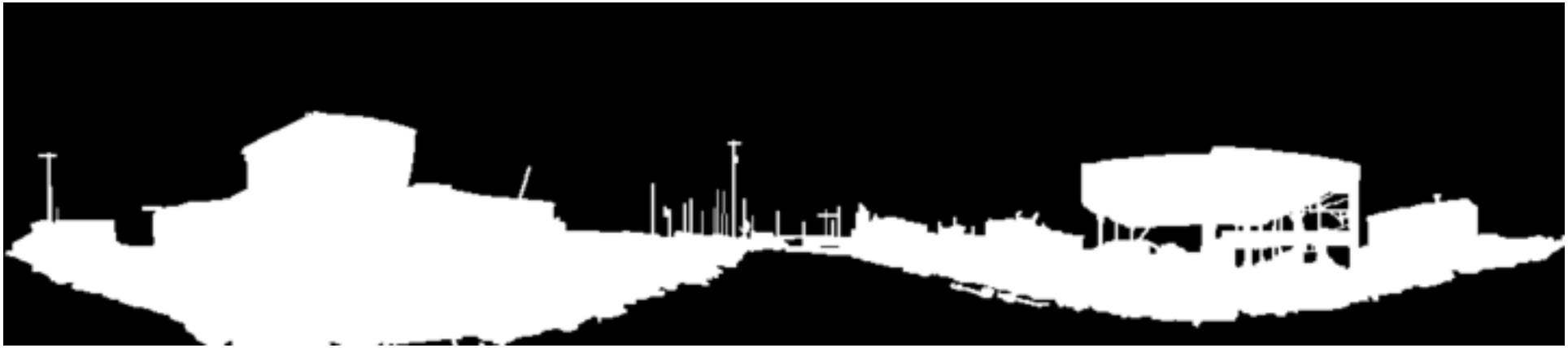
Click here to download high resolution image

Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

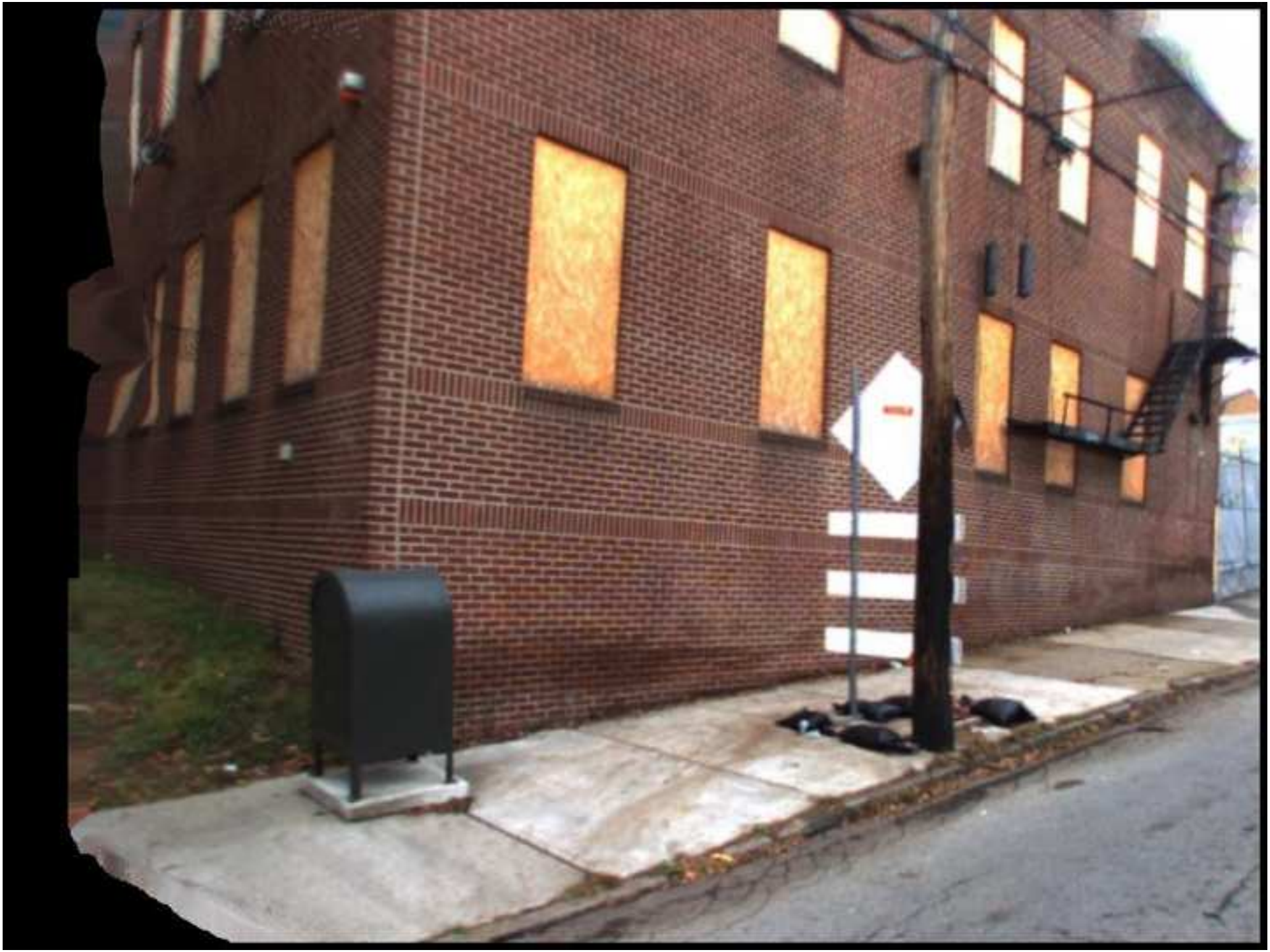
Click here to download high resolution image

Click here to download high resolution image

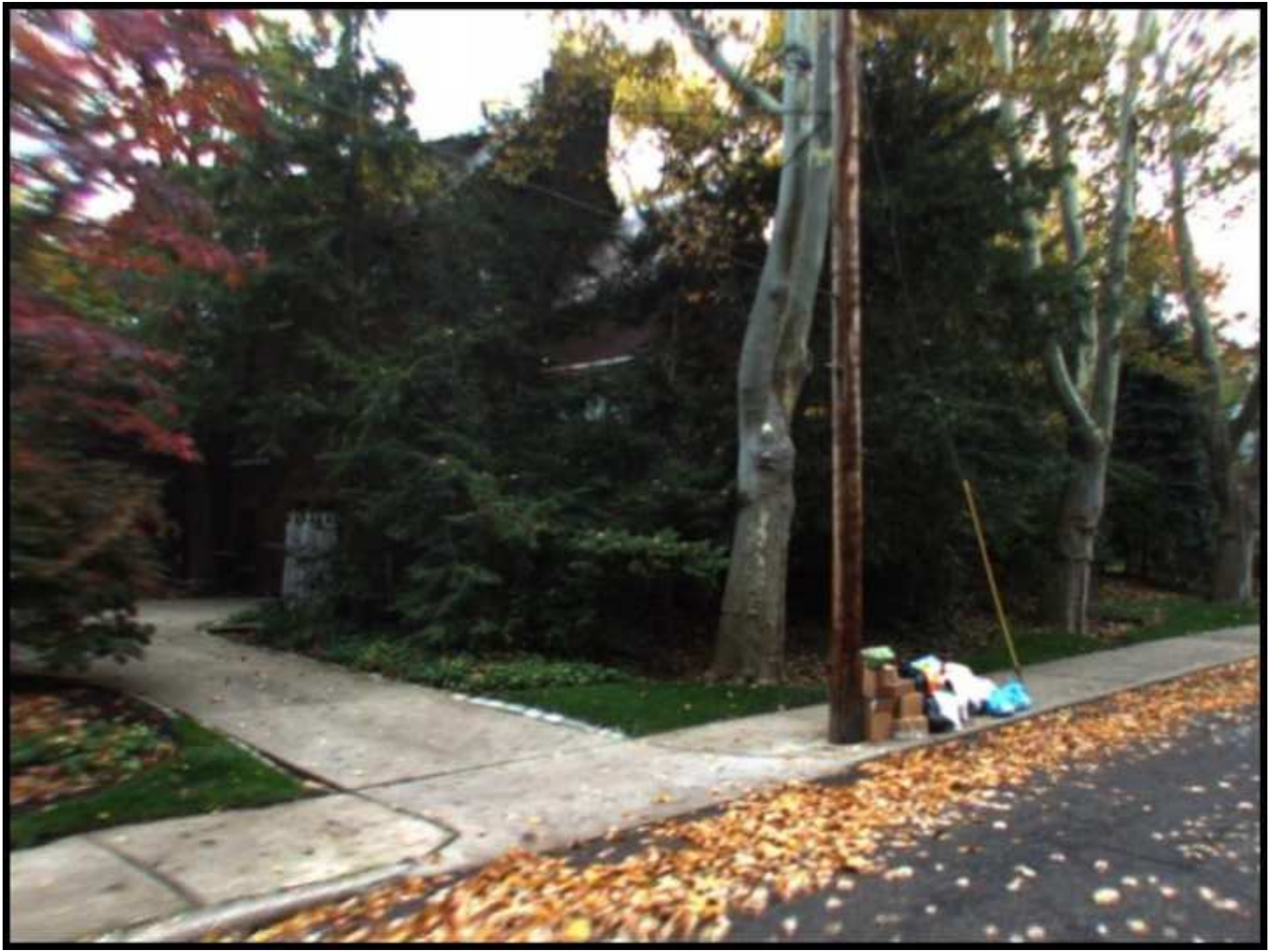
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image



Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

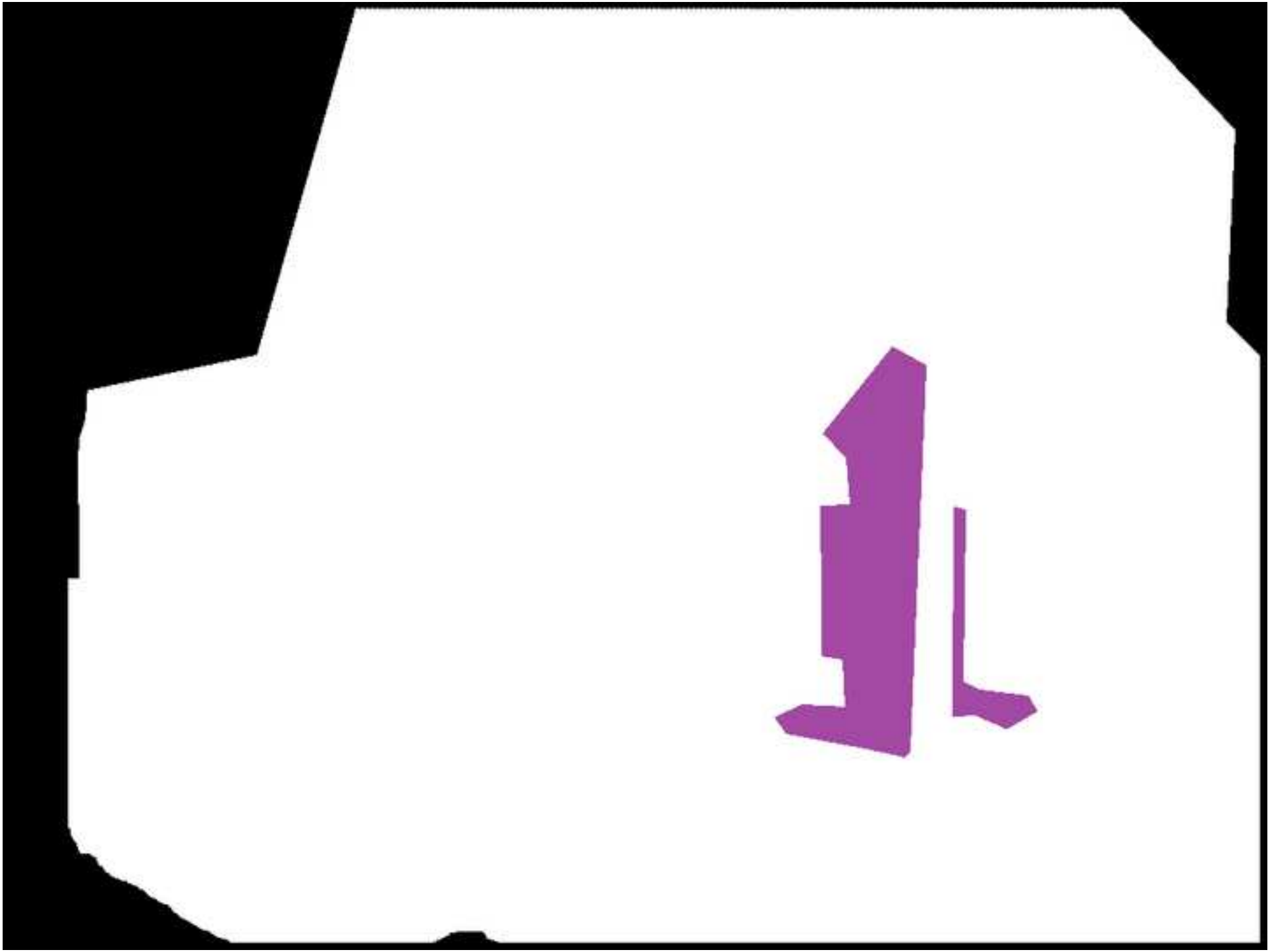
Click here to download high resolution image

Click here to download high resolution image

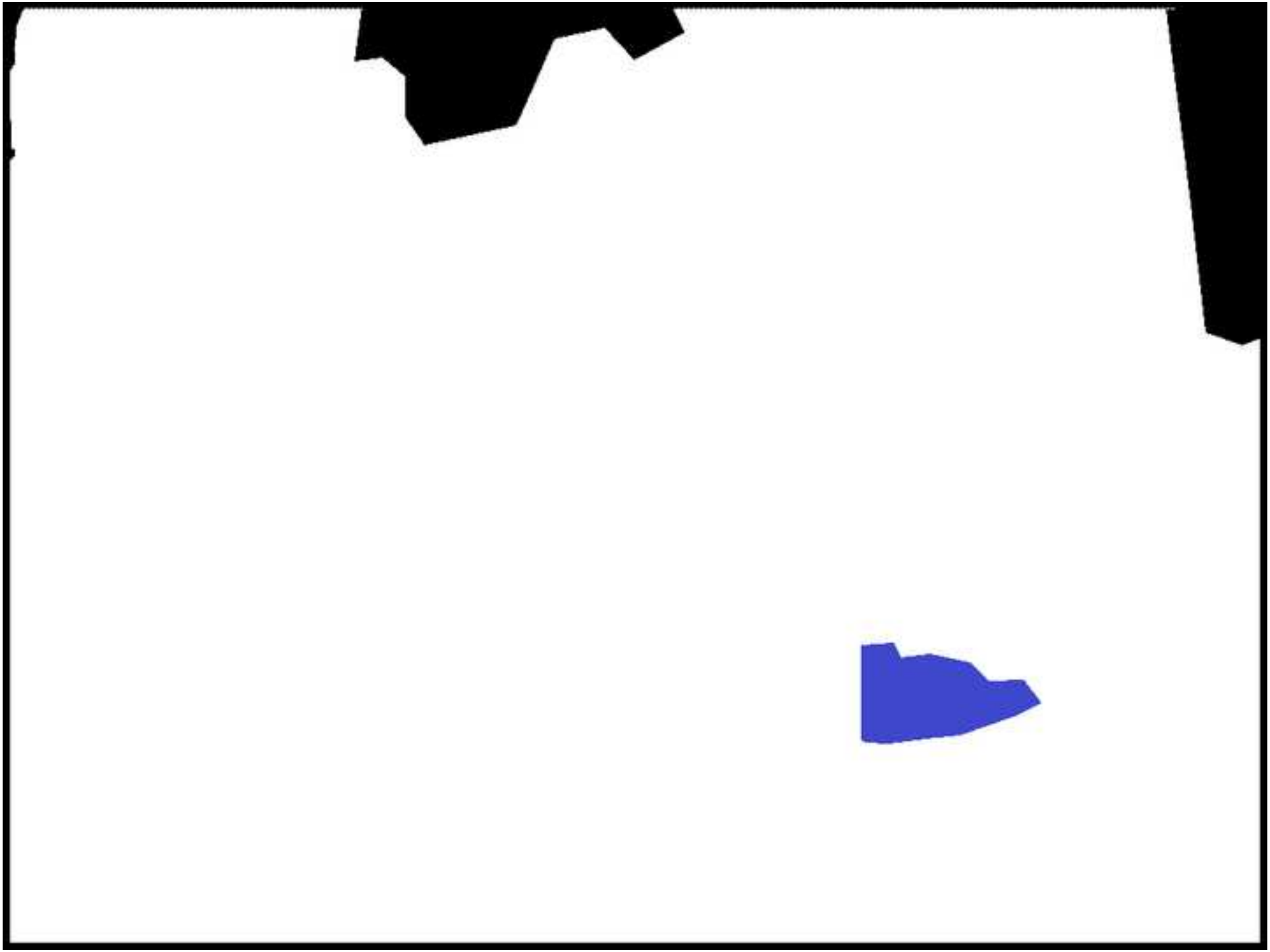
Click here to download high resolution image

Click here to download high resolution image

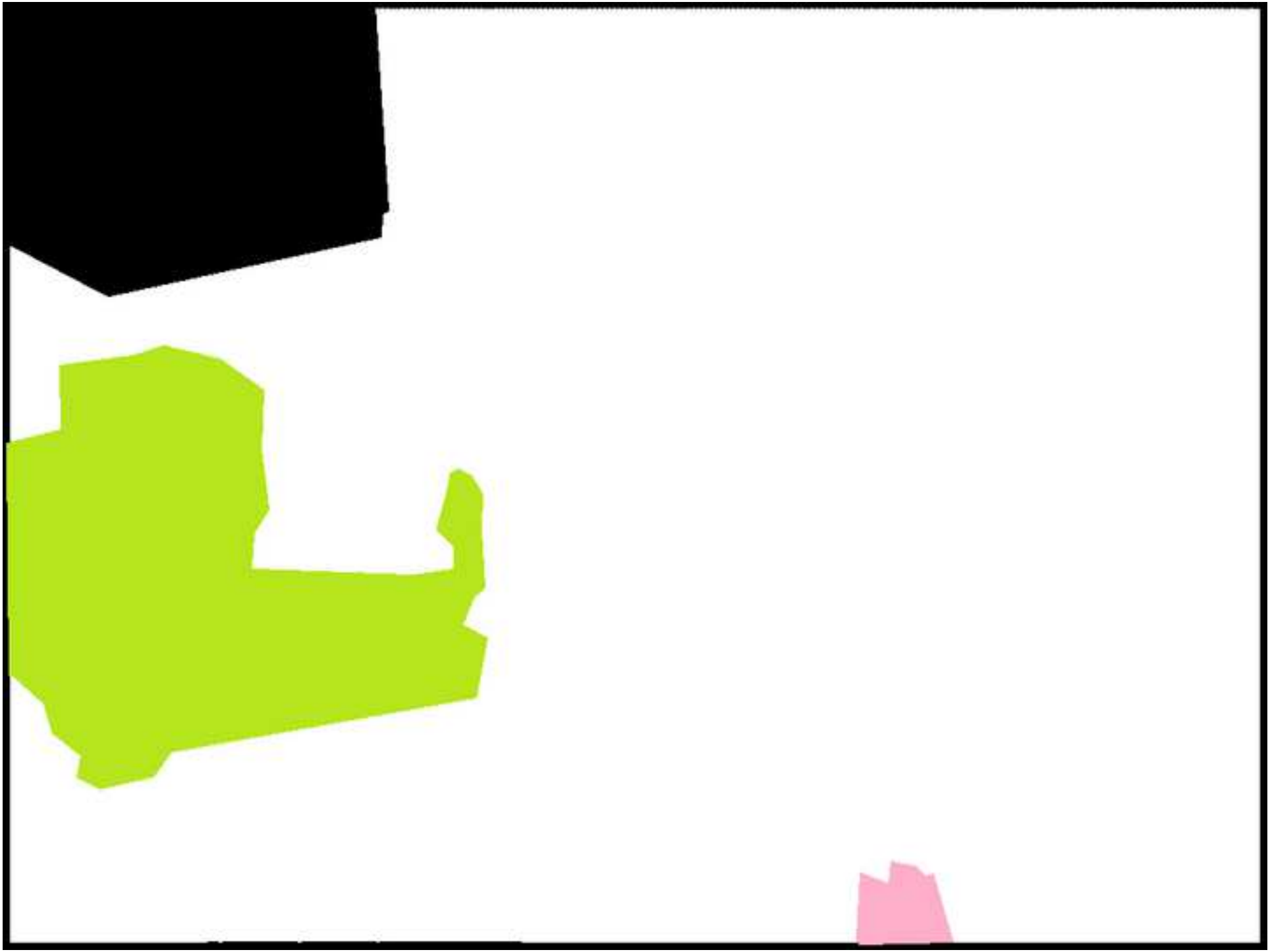
Click here to download high resolution image



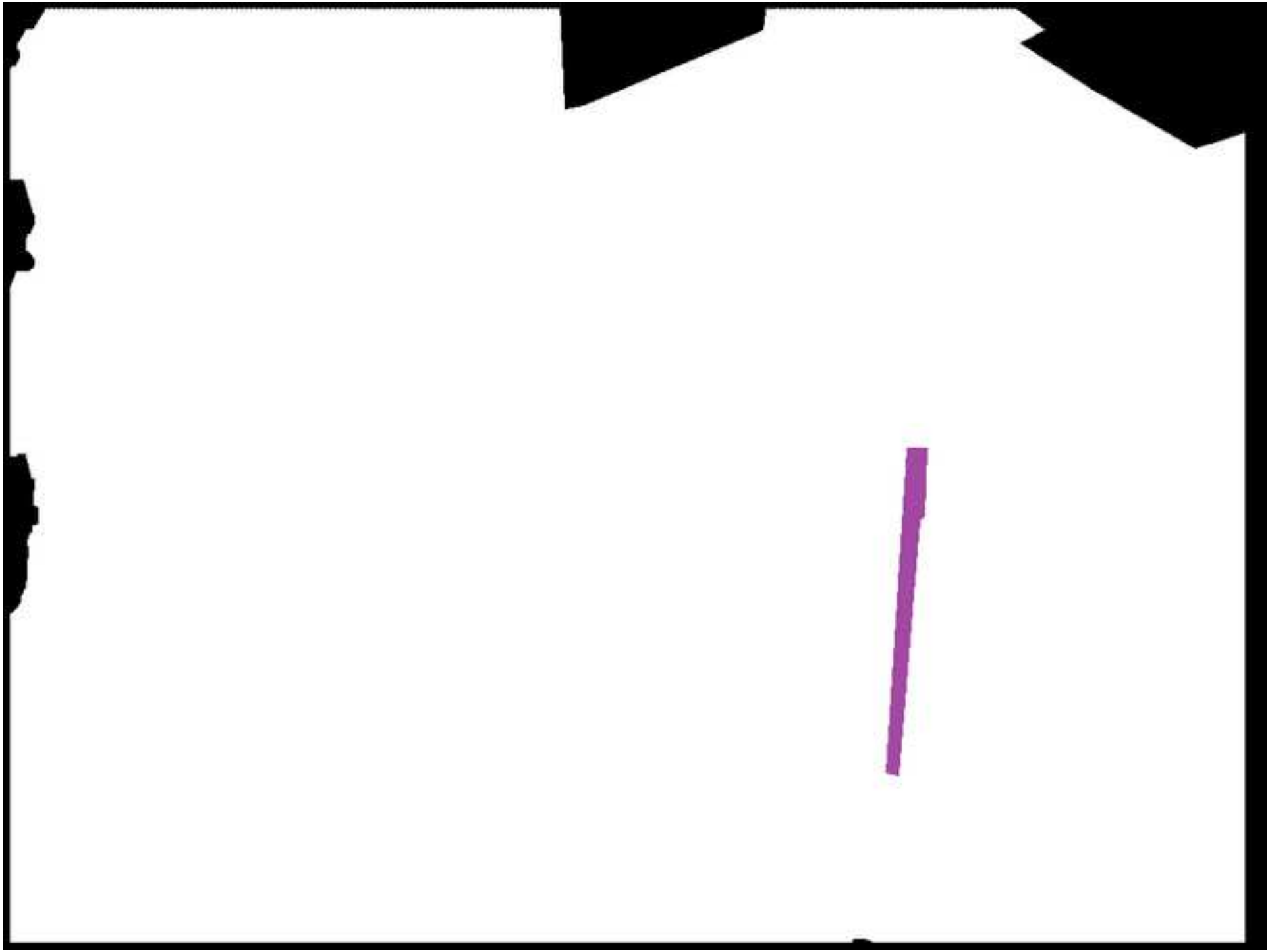
Click here to download high resolution image



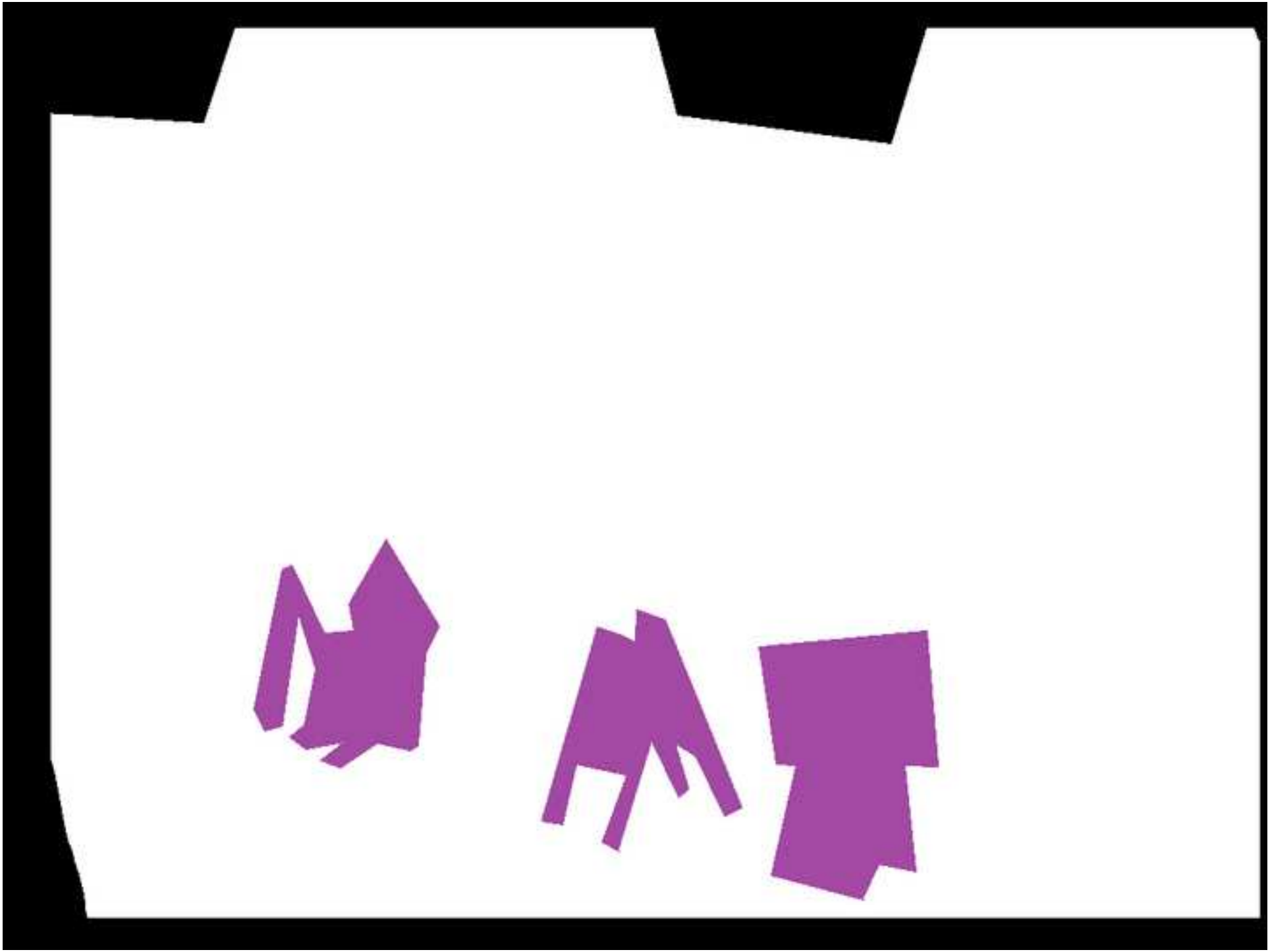
Click here to download high resolution image



Click here to download high resolution image



Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

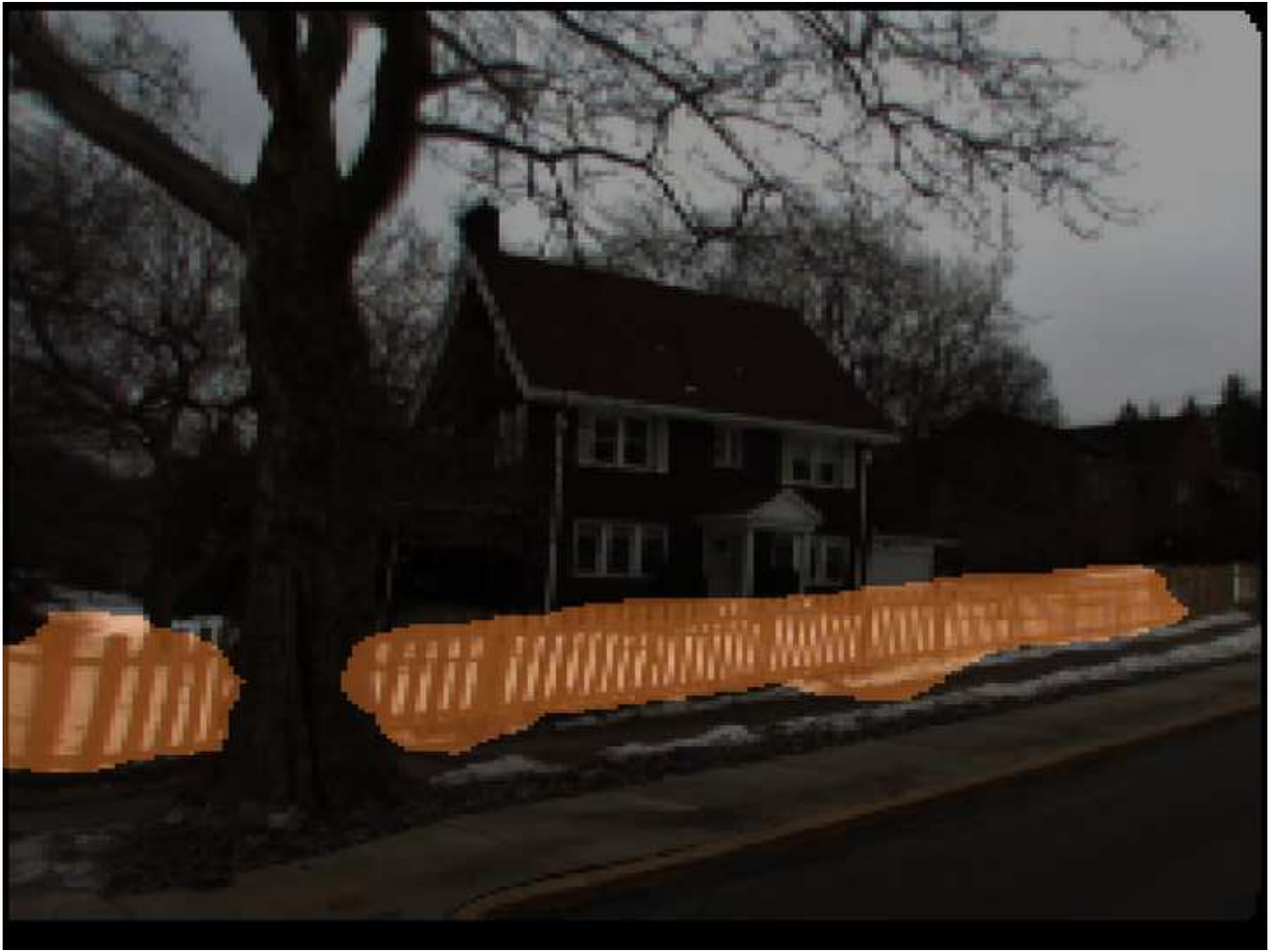
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

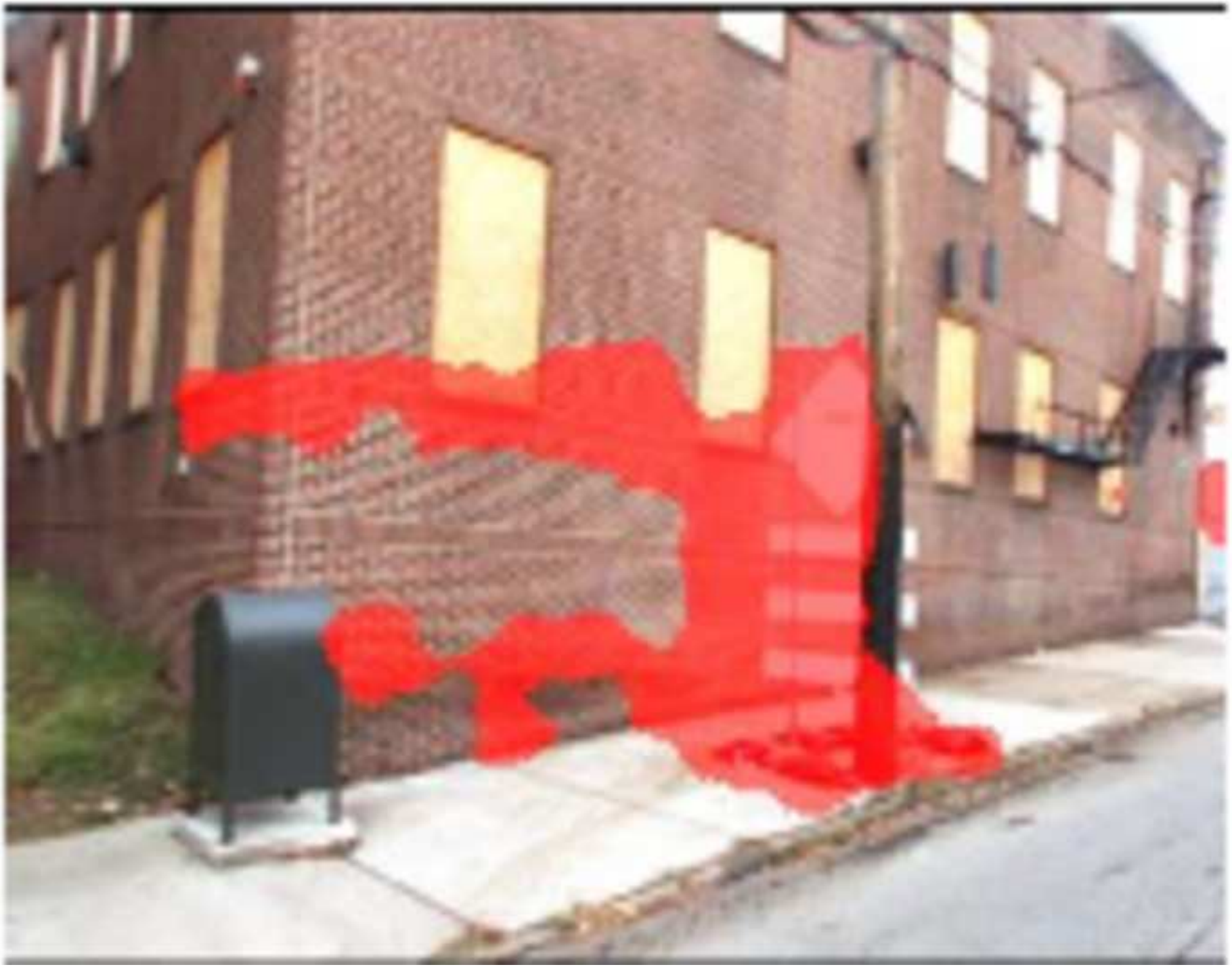
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

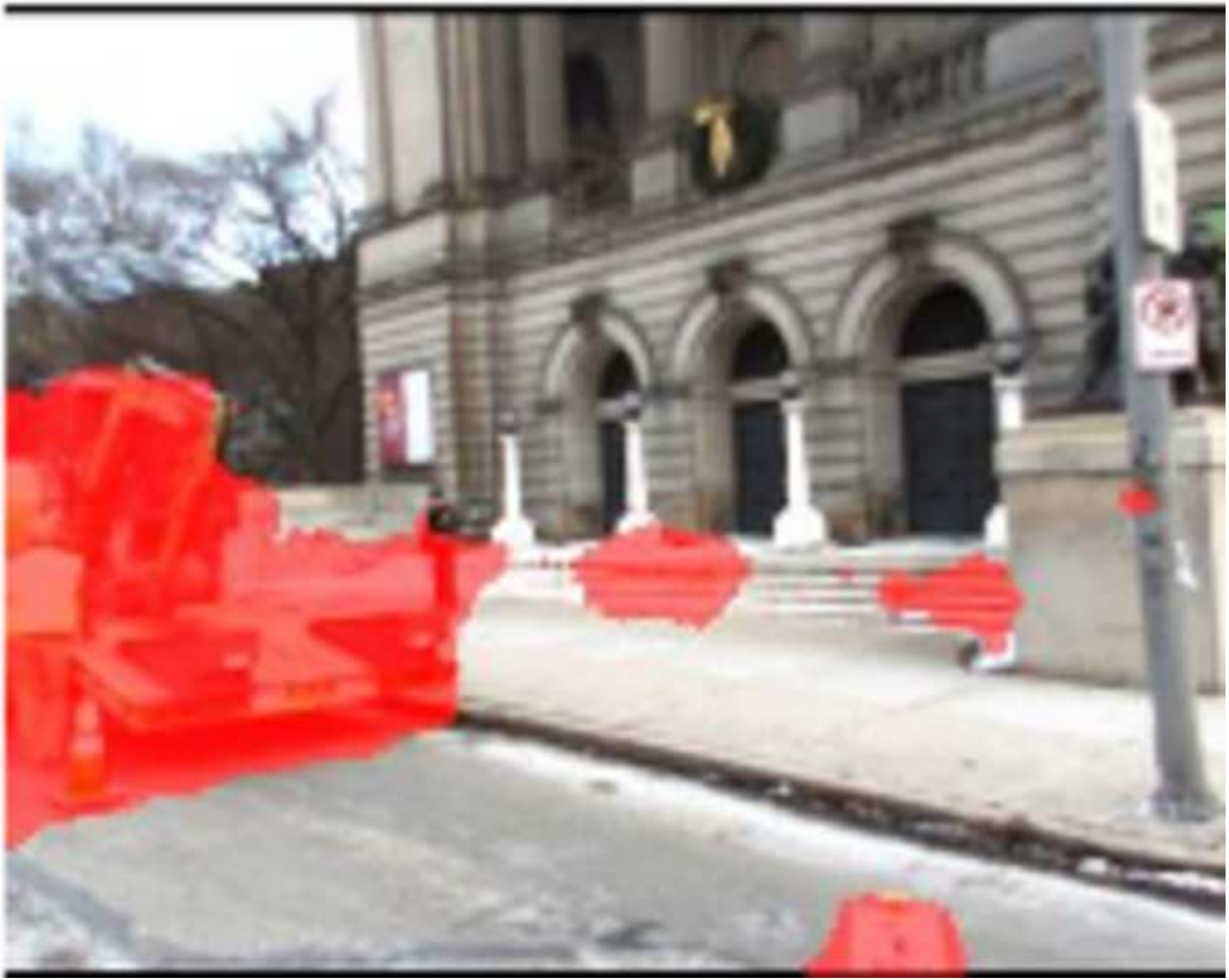
Click here to download high resolution image

Click here to download high resolution image



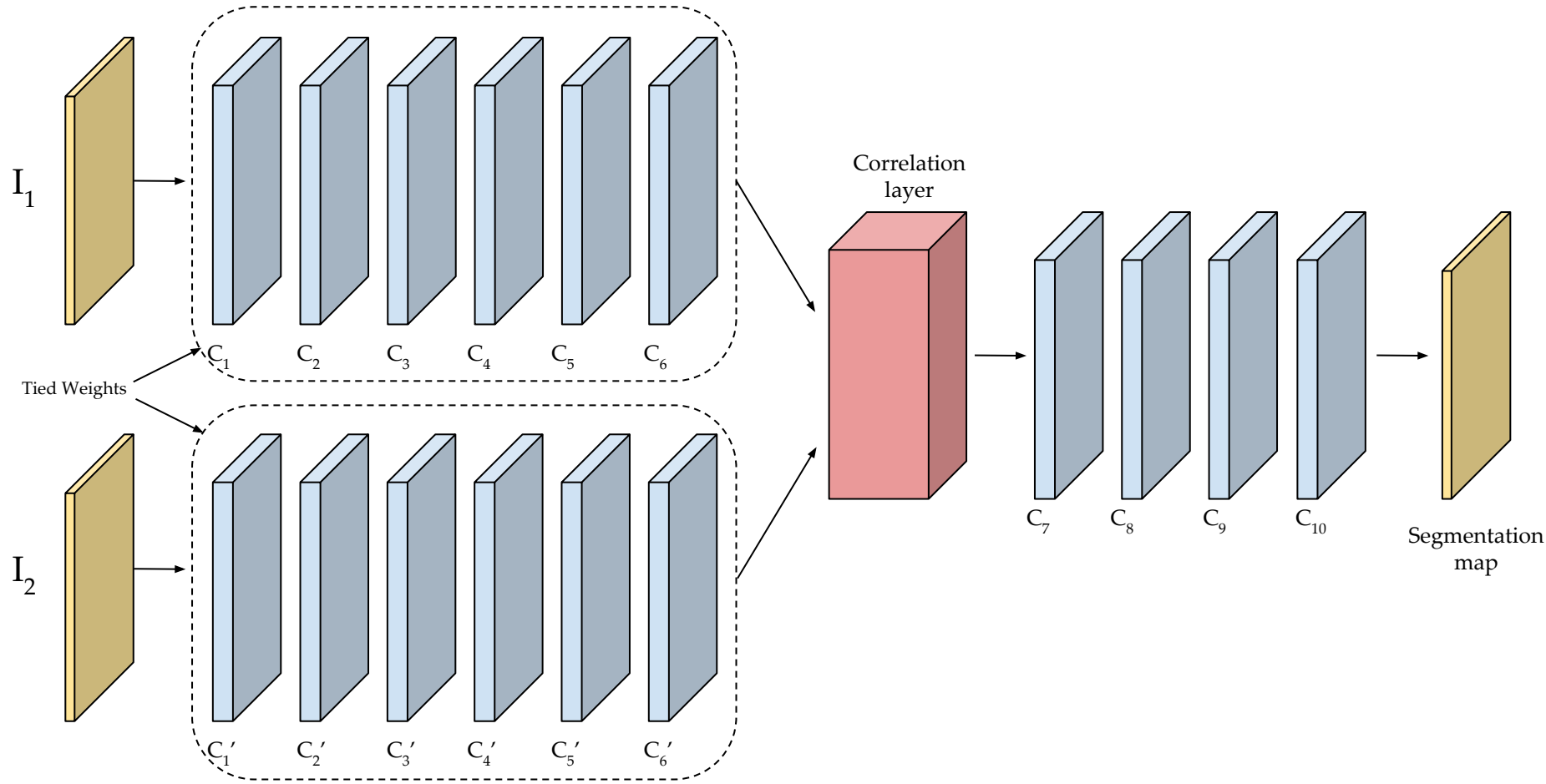
Click here to download high resolution image

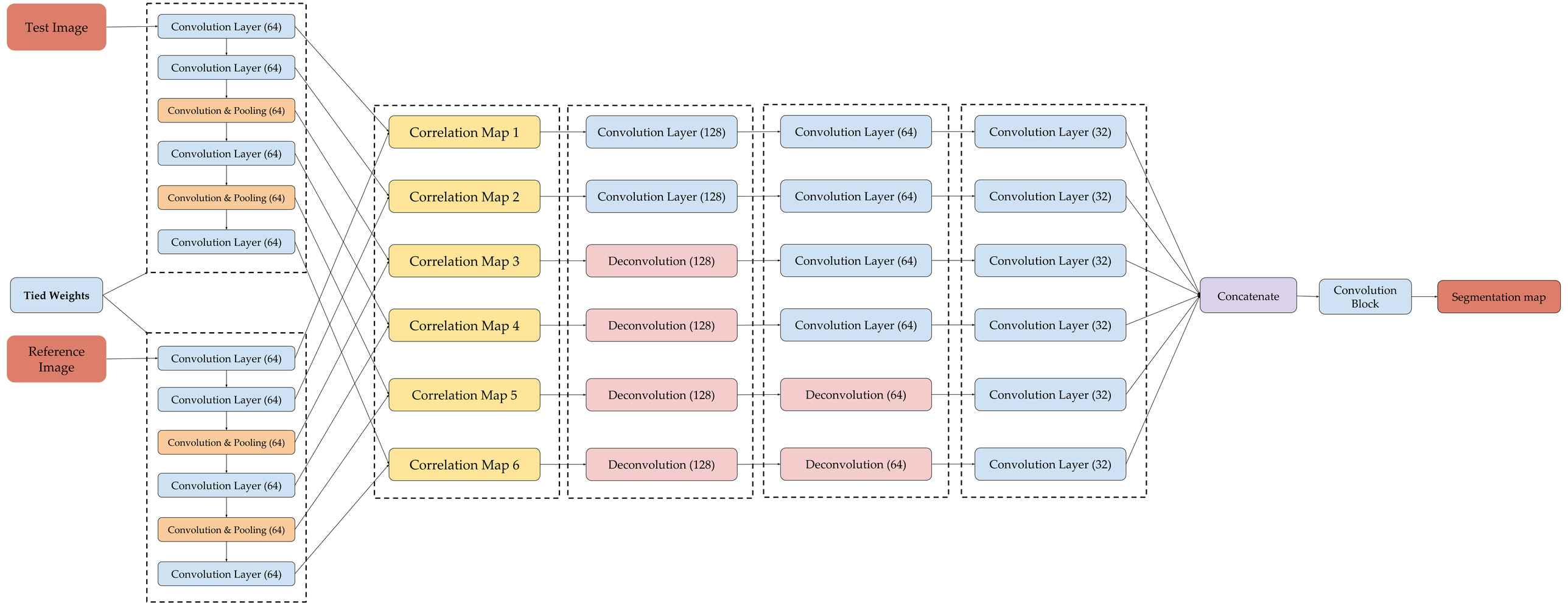
Click here to download high resolution image



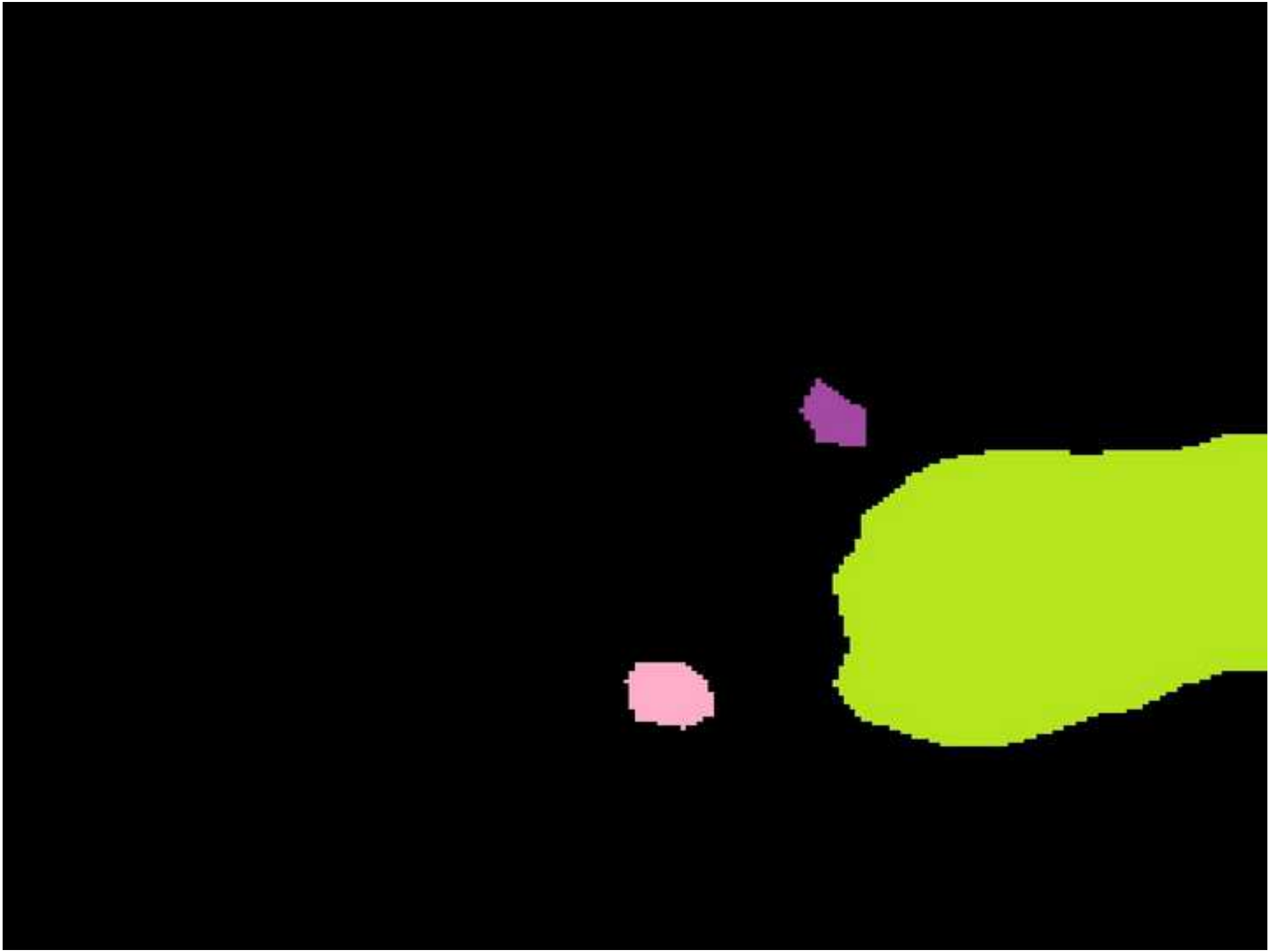
Click here to download high resolution image

Click here to download high resolution image

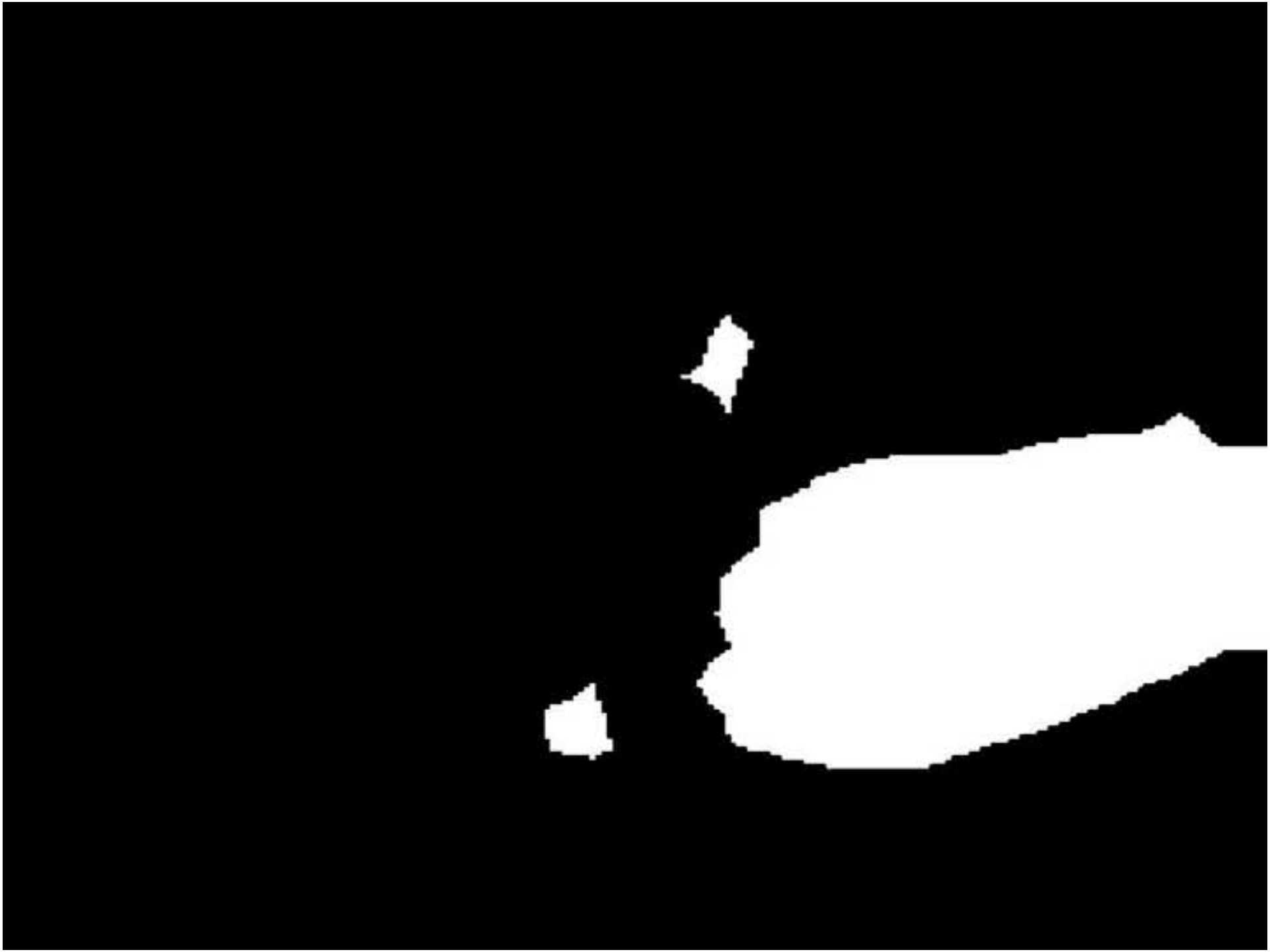

I_1
C_1
C_2
C_3
C_4
C_5
C_6
Tied Weights
I_2
C_1'
C_2'
C_3'
C_4'
C_5'
C_6'
Correlation layer
C_7
C_8
C_9
C_{10}
Segmentation map


Test Image
Convolution Layer (64)
Convolution Layer (64)
Convolution & Pooling (64)
Convolution Layer (64)
Convolution & Pooling (64)
Convolution Layer (64)
Tied Weights
Reference Image
Convolution Layer (64)
Convolution Layer (64)
Convolution & Pooling (64)
Convolution Layer (64)
Convolution & Pooling (64)
Convolution Layer (64)
Correlation Map 1
Correlation Map 2
Correlation Map 3
Correlation Map 4
Correlation Map 5
Correlation Map 6
Convolution Layer (128)
Convolution Layer (128)
Deconvolution (128)
Deconvolution (128)
Deconvolution (128)
Deconvolution (128)
Convolution Layer (64)
Convolution Layer (64)
Convolution Layer (64)
Convolution Layer (64)
Deconvolution (64)
Deconvolution (64)
Convolution Layer (32)
Convolution Layer (32)
Convolution Layer (32)
Convolution Layer (32)
Convolution Layer (32)
Convolution Layer (32)
Concatenate
Convolution Block
Segmentation map

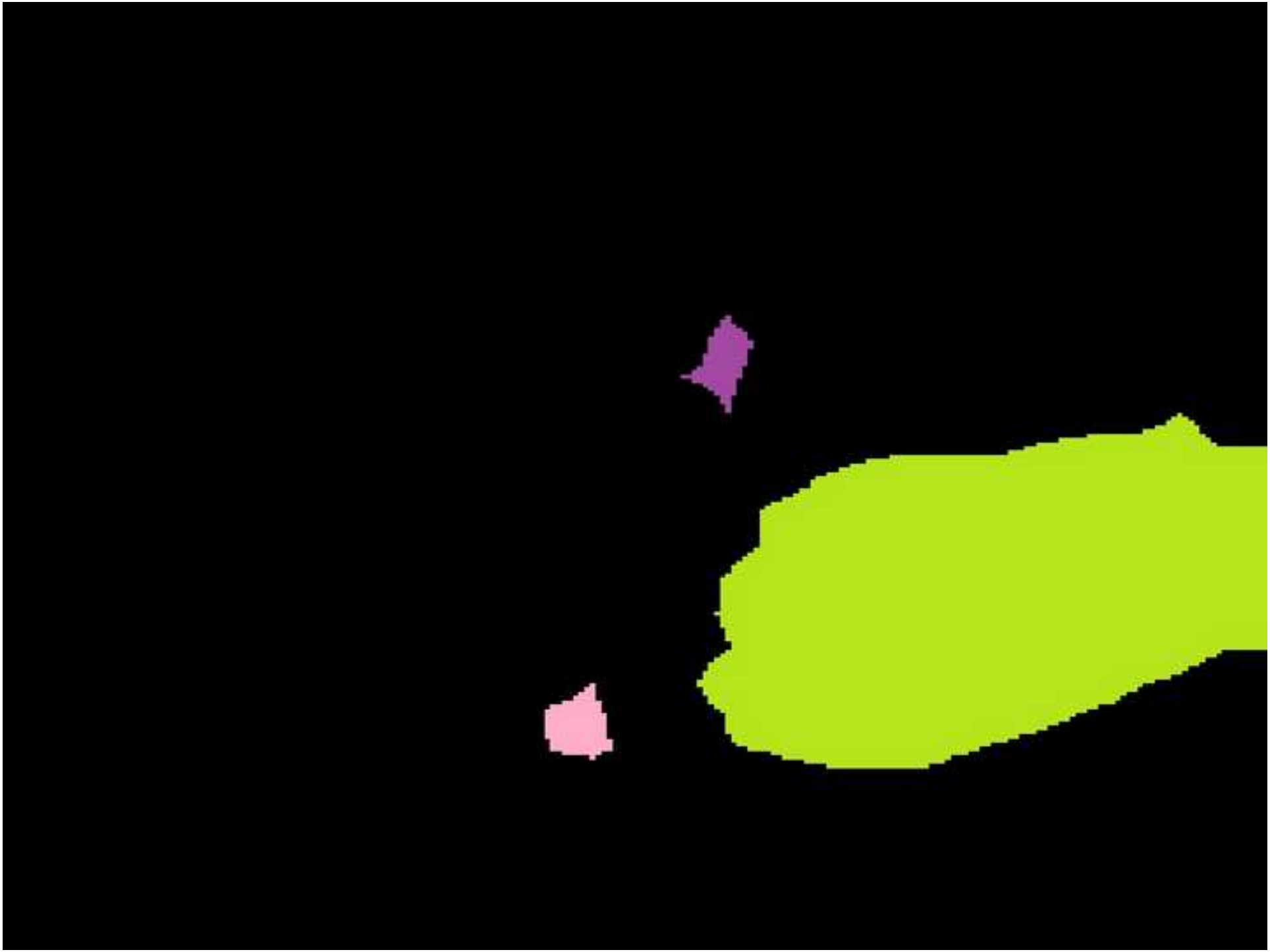
Click here to download high resolution image



Click here to download high resolution image

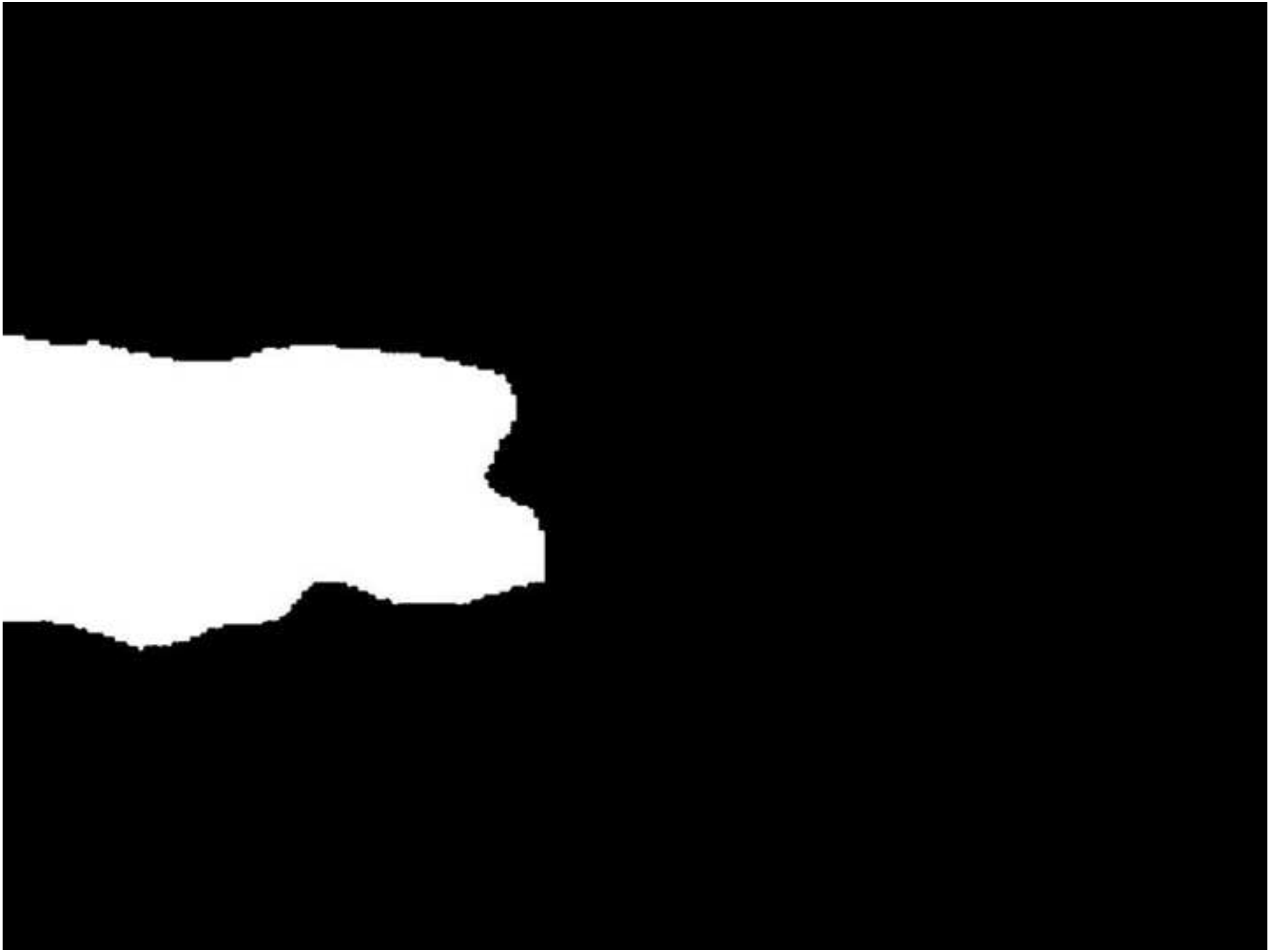


Click here to download high resolution image

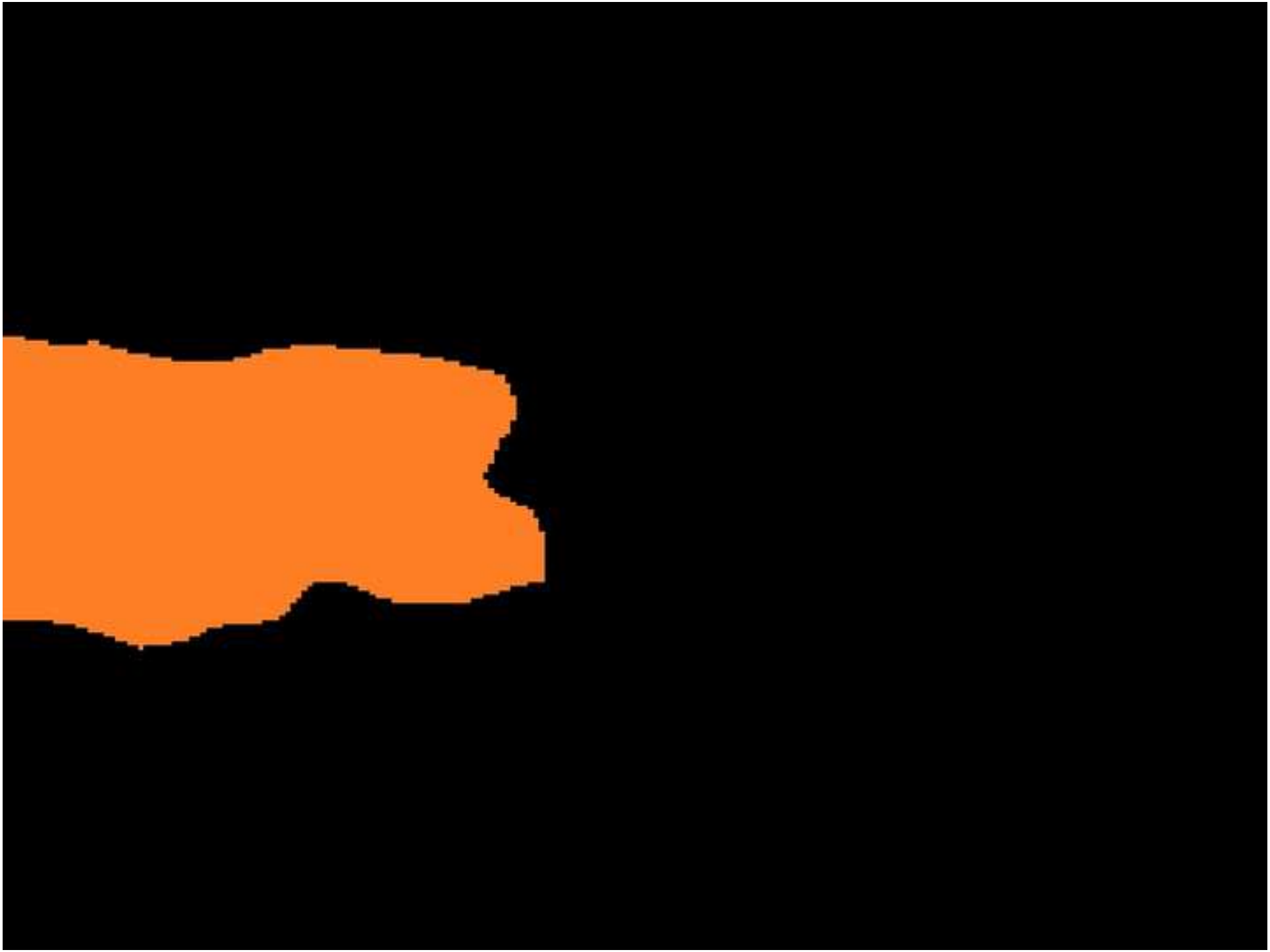


Click here to download high resolution image

Click here to download high resolution image



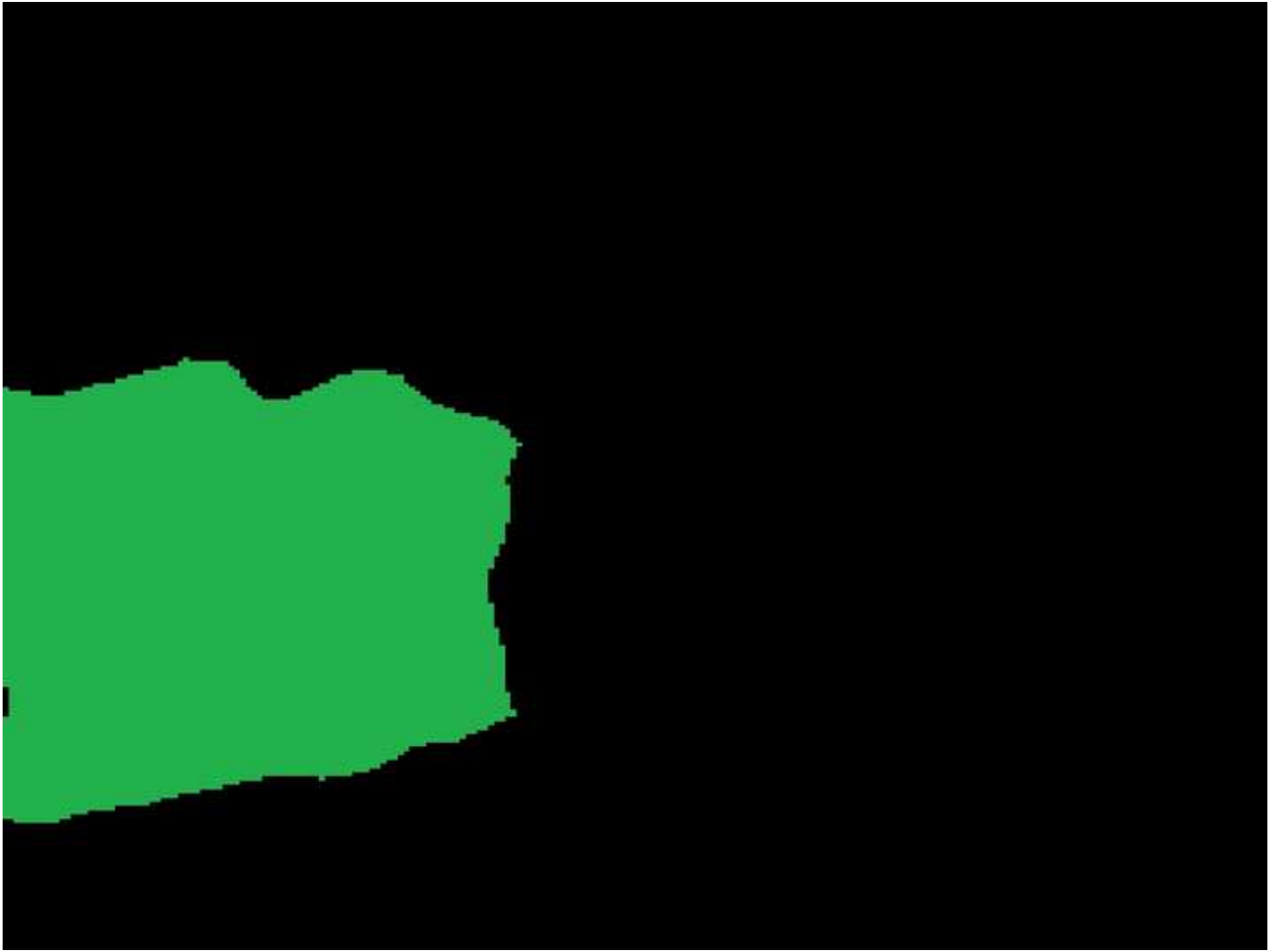
Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

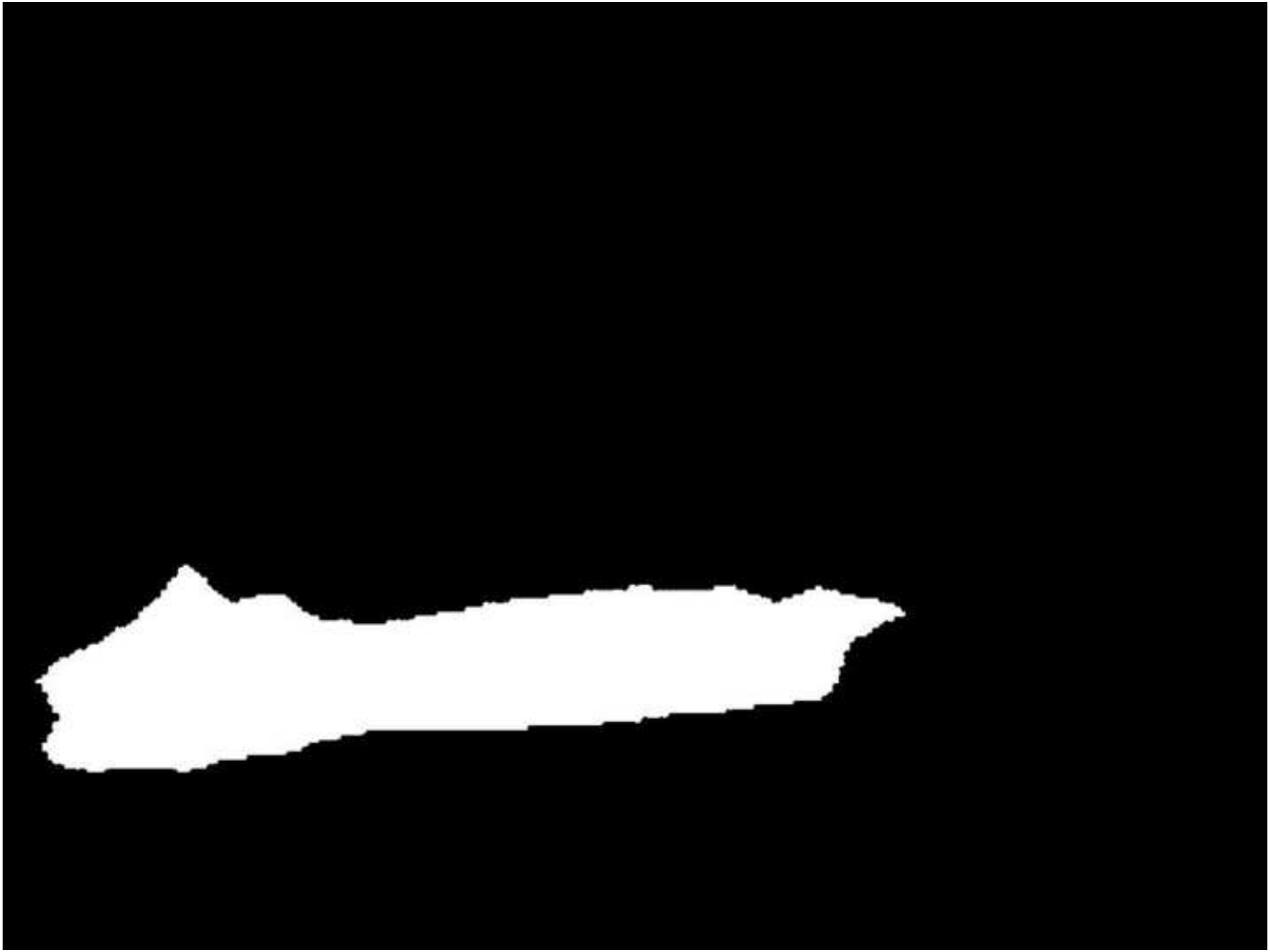


Click here to download high resolution image

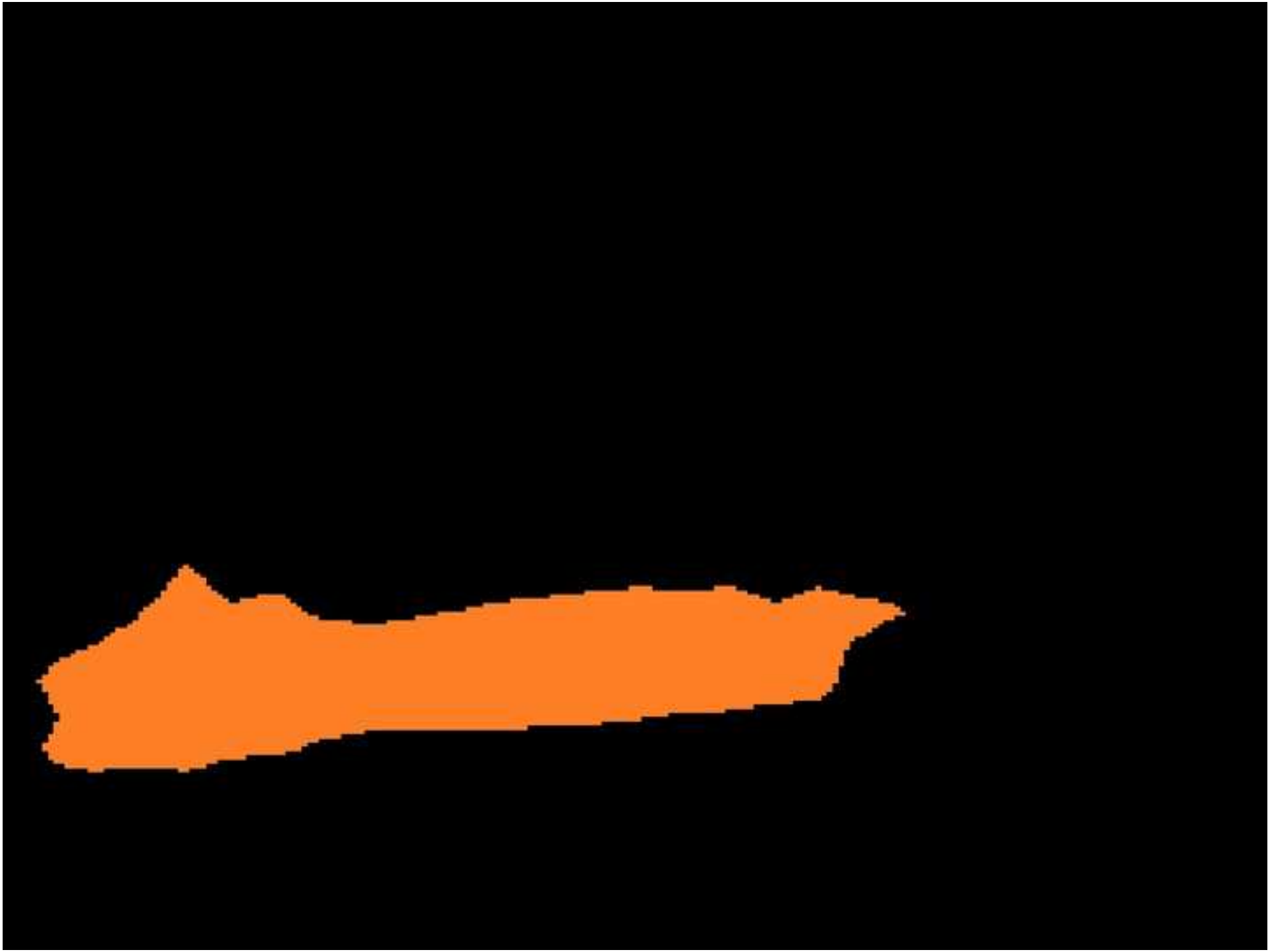
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image



Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

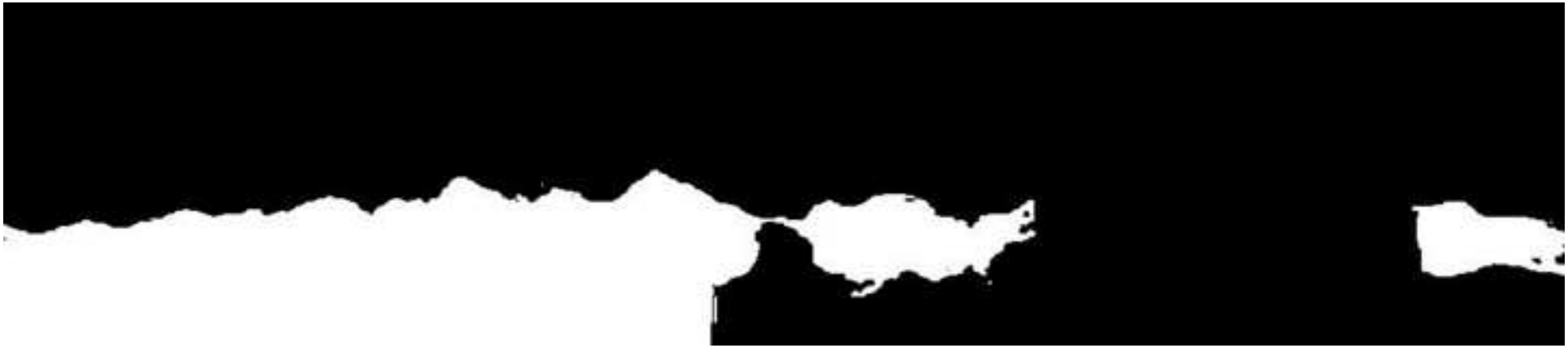
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

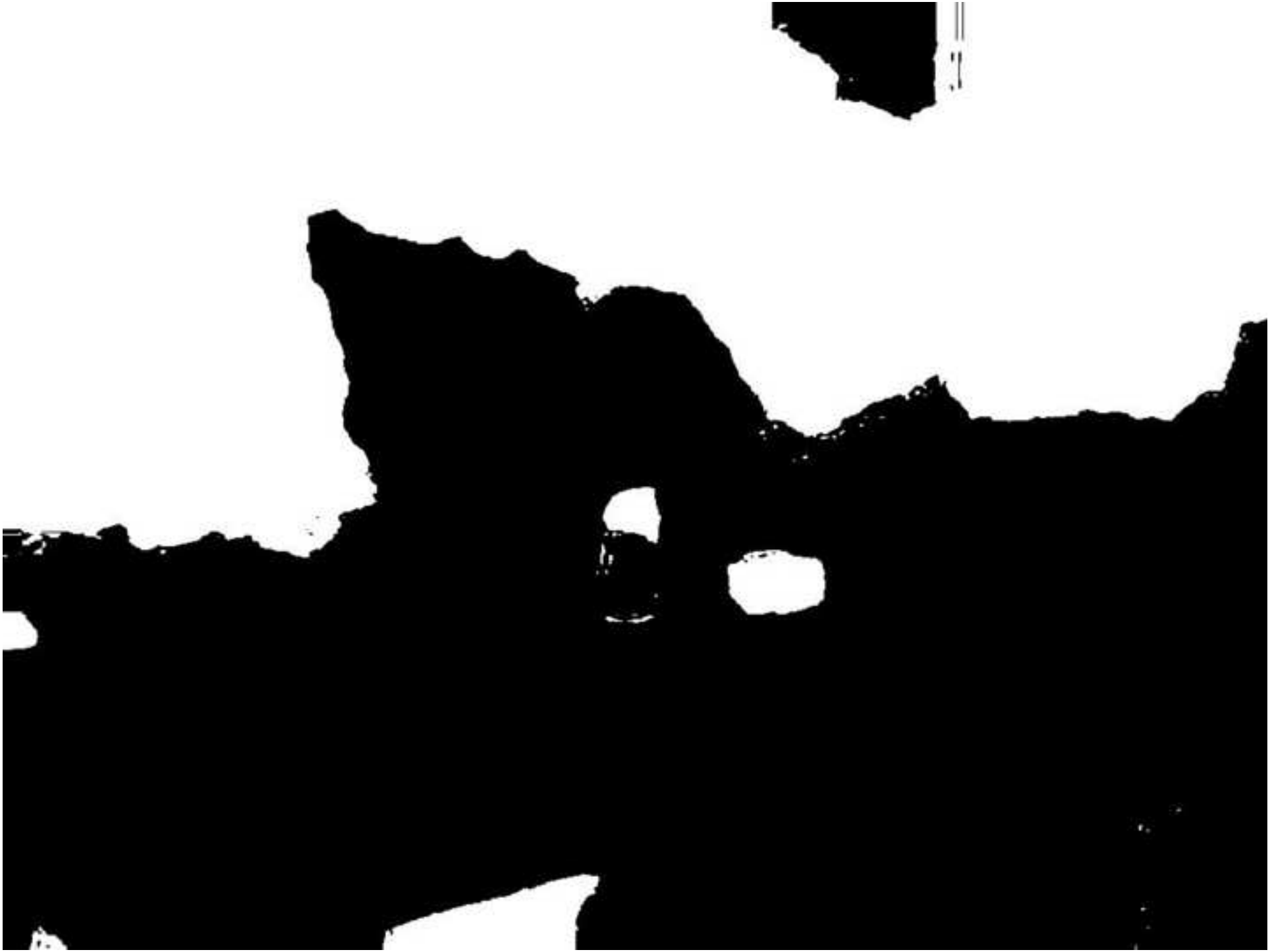
Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image



Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download high resolution image

Click here to download LaTeX Souce Files: refs.bib

Click here to download LaTeX Souce Files: main.tex

Click here to download LaTeX Souce Files: main.bbl

Click here to download LaTeX Souce Files: elsevier-logo.pdf

Click here to download LaTeX Souce Files: elsevier-logo.eps

Click here to download LaTeX Souce Files: ycviu-authorship.pdf

Click here to download LaTeX Souce Files: top-elslogo-fm1.pdf

*Declaration of Interest Statement

Declaration of interests

|√| The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: