Crowd Attention Estimation Automatisation Based on Semi-Automatic Image Semantic Segmentation by Using UNet and CRF Networks

Stanislav Sholtanyuk, Yakov Malionkin
The Department of Computer Applications and Systems
Belarusian State University
Minsk, Belarus
ssholtanyuk@bsu.by, malionkinjr@gmail.com

Bin Lei
Nanjing Research Institute of Electronics Engineering
Nanjing City, China
40220720@qq.com

Alexander Nedzved
The Department of Information Management Systems
Belarusian State University
Minsk, Belarus
Anedzved@bsu.by

***Abstract*—Semantic segmentation of crowd images plays a pivotal role in various applications such as crowd management, surveillance, and urban planning. In this paper, we propose an approach for dense and sparse crowd image semantic segmentation based on semi-automatic labeling by employing a combination of UNet and Conditional Random Field (CRF).**

We introduce a technique for generating segmentation maps for crowd images. We utilize UNet for initial rough segmentation followed by refinement using CRF. Experimental results demonstrate the model performs better in binary segmentation (crowd and on-crowded regions) rather than ternary segmentation (dense crowds, sparse crowds, and non-crowded areas). However the latter shows better results in terms of crowd detection (regardless of its type). Besides, we show the CRF refinement is significant in ternary segmentation.

Also, we highlight some crowd behavior patterns based on the proposed segmentation model. They differ in people's attention types, connections within and between crowds, and possibilities of emergencies.

***Keywords*—artificial neural networks, computer vision, crowd images, crowd detection, crowd behavior, image analysis, machine learning, semantic segmentation**

I. Introduction

Segmentation is the process of breaking the image into distinct segments or regions that represent objects of interest or their structure. Image segmentation is one of the pivotal stages in computer vision and image analysis. The main purpose of segmentation is to highlight key objects and their features for a more detailed understanding of the content represented in the image. More rigorously, during image segmentation, a label representing a certain class is assigned to each image pixel so pixels in the same class stand for a joint object and demonstrate some shared characteristics, and pixels with different labels somehow differ from each other.

Image segmentation can be compared with a classification task with some initial classes given. However, the image classification task implies a label for the whole image as the result, whereas in semantic segmentation, a label is assigned to each image pixel. Thus, unlike classification tasks, not only does segmentation have a purpose to determine the main object of interest but also to examine its optical and morphological characteristics like its edges, position on the image, and its position relative to other objects (if any).

In computer vision, there are two main types of image segmentation: semantic and instance segmentation. In semantic segmentation, each pixel must be associated with one of the predefined classes (e. g. background, person, vehicle, building, etc.), and such classes are represented by their colors. In instance segmentation, each pixel is also classified based on some given classes, but distinct objects within a class are highlighted by different colors.

Semantic segmentation of crowd images is an effective tool for analyzing and understanding crowded scenes, crowd structure, and behavior. Crowd semantic segmentation can be used in the following applications:

1) Counting and analyzing the crowd. Segmentation can facilitate such tasks by decomposing the image into clusters, and some basic techniques could be applied to them to analyze the crowd structure within them. It is useful for monitoring crowded places like stadiums, markets, malls, fairs, social events, etc. [1]–[4].
2) Security and surveillance. Some segmentation techniques allow highlighting single persons which is important for detecting abnormal, troublesome,

The research was supported by the project "Technology Development Agreement of developing of algorithms of remote sensing image processing", agreement number 22CETC19-ICN1785

or even potentially dangerous situations, e.g. lost things, suspicious behavior, civil unrest, etc. [5], [6].

3) Marketing and analytics. In marketing, crowd segmentation can be used for analyzing customers' behavior when they wander, seek something, or stop near a merchant's place in malls and fairs. In such context, segmentation can be a helping tool to improve goods placing, estimate marketing strategies efficiency, and improve customer service.
4) Transport management. In urban planning and traffic management, crowd segmentation can be used for pedestrian traffic optimization, crowd prevention, as well as planning more efficient pedestrian and transport routes.
5) Social behavior research. Nowadays, crowd segmentation is effectively used to research social behavior, e. g. by analyzing the dynamics of interaction between people in different scenes.

However, semantic segmentation of crowd images is a challenging task due to the complex and dynamic nature of crowd scenes, which often exhibit variations in density, scale, occlusions, and illumination conditions. Accurate segmentation of individual objects within a crowd is crucial for the above-mentioned applications. Traditional methods for crowd segmentation rely on handcrafted features and manual annotation, which are labor-intensive and often fail to capture the diverse characteristics of crowd scenes.

In recent years, semantic technologies are widely used in various computer vision applications such as knowledge based computer vision [7]–[9], video and images annotation [10], and video retrieval [11], [12]. In fact, they have the potential to significantly enhance the capabilities of crowd segmentation and attention estimation systems by incorporating semantic understanding into the analysis process. By leveraging semantic technologies, such as ontologies and knowledge graphs [13], [14], as well as semantic reasoning mechanisms [9], [15], researchers and practitioners can improve the accuracy, efficiency, and interpretability of crowd segmentation and attention estimation algorithms.

One key aspect of applying semantic technologies to crowd segmentation is the incorporation of domain-specific knowledge about crowd behavior, scene context, and environmental factors [16], [17]. By encoding this knowledge into formal ontologies or knowledge graphs, segmentation algorithms can better understand the semantics of crowd scenes, leading to more robust and context-aware segmentation results [18], [19]. Furthermore, semantic reasoning mechanisms can be used to infer higher-level semantic concepts from low-level segmentation outputs, enabling the identification of complex crowd behaviors and interactions [9].

In context of OSTIS systems, deep learning techniques have shown remarkable success in various computer vision tasks, including semantic segmentation. Convolutional Neural Networks (CNNs) have emerged as powerful tools for learning discriminative features directly from data, enabling end-to-end training for semantic segmentation tasks [21]–[24].

II. Main Semantic Segmentation Techniques Survey

Semantic segmentation, the task of assigning semantic labels to each pixel in an image, has witnessed significant advancements in recent years driven by deep learning techniques. In this subsection, we provide an overview of the main semantic segmentation techniques, focusing on classical and deep learning-based approaches.

A. Sliding Window

A simple semantic segmentation method using a sliding window involves sequentially applying the window of different sizes to the entire image [25], [26]. This process consists of several stages like setting the size of the window, applying it to the image, classifying the extracted features, and building the semantic map. Such an approach, however, possesses multiple cons some of them being:

1) High computational complexity. When working with high-resolution images, step-by-step window displacements lead to excessive iterations, during which several time-consuming operations are performed.
2) The lack of a global context. As far as each region is processed independently, such a technique grasps little to no connection between regions. It might result in a fragmented representation of the object and a lack of understanding of the whole picture.
3) Different size objects predicament. If the image depicts multiple objects of interest with different sizes, then the fixed-sized window might struggle with extracting features from some of them. Dynamic resizing of the window is likely to pose extra computational difficulties.
4) Objects overlapping. When there are some overlapping objects on the image, the sliding window method is likely to give poorly highlighted edges, especially if the objects are close to each other or have the same size.
5) Sensitivity to the parameters. Several parameters like window size or the step value should be fine-tuned precisely. Otherwise, the result might get worse dramatically.

The method can be used in remote monitoring when an observer is so distant that the scene can be considered as infinitely distanced from them. Another suitable condition to use the approach is the equality of sizes of interesting objects so one could predetermine the window size.

B. Fully Convolutional Networks

Fully convolutional networks (FCN) are the type of neural network designed for semantic segmentation tasks. Instead of using fully connected layers, FCNs use convolutional layers. It allows processing input images of arbitrary size and generating segmentation maps with the same size [24]. The main concept of FCN is replacing fully connected layers with convolutional ones to obtain a segmentation map of the same size as the initial image. Besides, some intermediate layers and skip connections between them could be used to improve the segmentation and get more detailed information.

Fully convolutional networks possess the following drawbacks:

1) Ineffectiveness with objects having different sizes. FCNs might concentrate more on larger objects, neglecting smaller ones.
2) The spatial information loss. Using maximal pooling layers and image upscaling might lead to spatial information loss, especially if some features are neglected in the convolutional layers.
   Furthermore, FCNs are characterized by a vast amount of parameters. As far as FCNs use convolutional layers, the number of parameters might significantly exceed compared to simple models. This issue raises even more drawbacks:
3) Computational complexity. Estimating the parameters and fine-tuning the FCN requires vast time and computational resources.
4) Training data requirements. A large number of parameters arise need in big training data which must be well-prepared to avoid network overfitting which causes poor ability of the model to make general results.
5) FCNs are prone to overfitting.

C. Convolutional Neural Networks

Convolutional neural networks (CNN) are one of the key frameworks in image processing and computer vision. They combine such tools as image convolution, image pooling, feature extracting, and classification based on those features [27]. The main idea is to use multiple layers of different types: convolutional (to extract features), pooling (to solve image size-related issues), and fully connected layers (to classify the image based on the extracted features). CNNs have established their place in computer vision and image processing thanks to many advantages like effectiveness in extracting semantic, morphological, and spatial features, their ability to process images of different sizes, as well as their ability to identify the image context.

D. Conditional Random Fields

Conditional random fields (CRF) is a statistical model effectively used with CNNs to refine semantic segmentation maps. A CRF takes part in postprocessing the results of a CNN prediction to refine and improve the segmentation spatial structure [28], [29]. The common way to use CRF in image semantic segmentation features the next stages:

1) Receiving predictions from the CNN. The prediction includes a segmentation map with probabilities for each pixel belonging to each considered class.
2) Preparing the features for the CRF. The probabilities from the segmentation map are used to form features given to the input of the CRF. Spatial coordinates of separate pixels or objects may also be such features.
3) Applying the CRF to refine the segmentation. The CRF uses context and spatial data from the CNN to refine the segmentation. CRF usually models interconnections between neighbor pixels and implements that information into the semantic map.
4) MAP optimization. The CRF uses the MAP method (Maximum A Posteriori) to tune its parameters to maximize the *a posteriori* probability for each pixel to fit in the appropriate class.

CRFs provide context information based on the information on interconnections between pixels. That allows us to improve edge detection and highlighting object details. Besides, CRF might reduce noise and smooth predictions which is extremely important in applications where high-quality object separation is crucial.

III. Methodology

In the research, we consider the following task of crowd semantic segmentation. Based on various characteristics (e. g. crowd density, people's spatial distribution, their visual texture), crowds can be classified as dense and sparse. Different approaches can be employed to determine if the given crowd is dense or sparse, e.g. manual annotation, crowd density maps, computer vision methods considering the texture of the image, as well as social force models [30], [31]. In this paper, we use a semi-automatic approach to generate ground truth maps for binary (non-crowded and crowd regions) and ternary (non-crowded, sparsely crowded, and densely crowded areas) semantic segmentation. This features the following steps:

- An annotated crowd images dataset is used (Fig. 1a). For each image, the annotations present the locations of the labels assigned to each individual's head.
- Based on the labels' locations, a 2D binary array is assigned to each image where 0 indicates the absence of a person's head, and 1 stands for a label.
- The array is Gaussian blurred to obtain density maps (Fig. 1b).

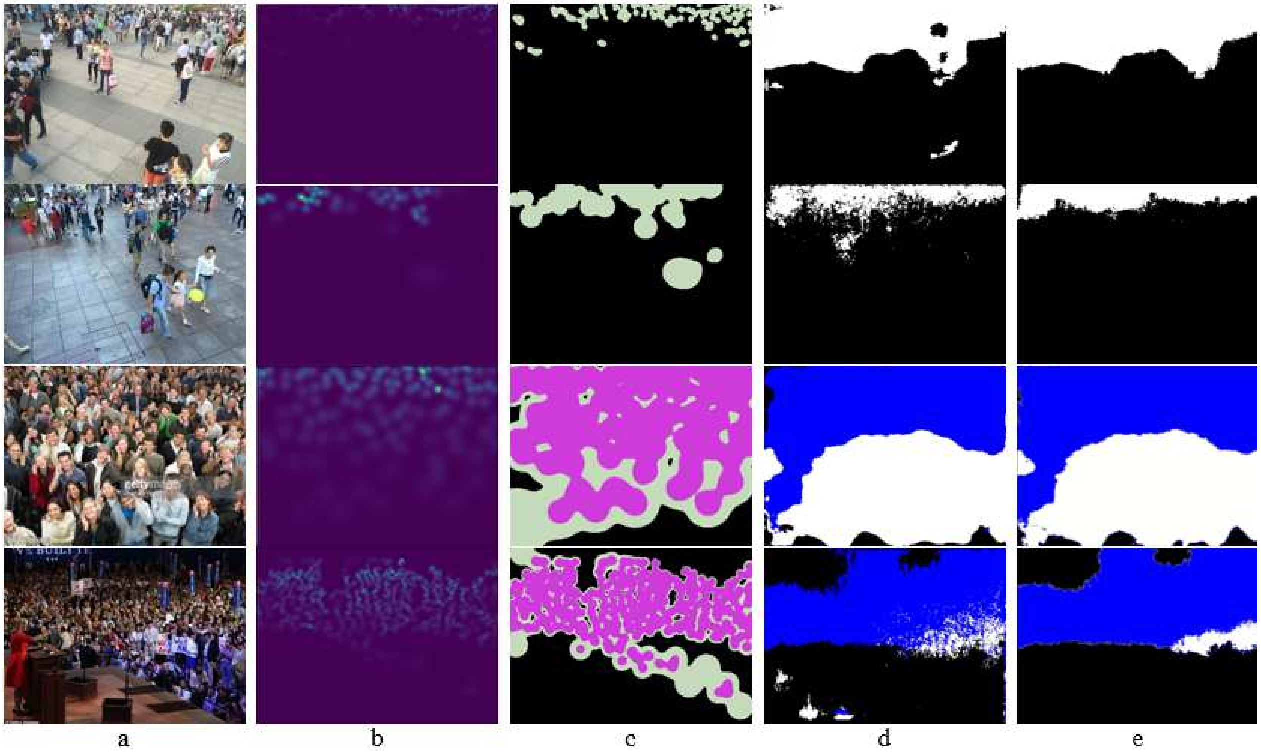

Figure 1. A crowd image (a), the corresponding density map (b), ground truth segmentation (purple is dense crowd, green is sparse crowd, and black is non-crowded areas) (c), predicted segmentation after employing UNet (blue is dense crowd, white is sparse crowd, and black is non-crowded areas) (d), and the segmentation refined after using CRF (e). Each row presents an example of binary and ternary prediction (first two rows and last two rows respectively) using either dice (first and third lines) or focal (second and fourth) loss function

- The resulting density maps are segmented into two or three areas based on thresholding values (Fig. 1c).

Based on the analysis given above, we decided to use a CNN + CRF model for the task. We use UNet as the network to calculate initial predictions. The initial images and the corresponding ground truth segmentation maps are divided into training, validating, and testing samples to train the UNet neural network which effectively takes advantage of the depicted objects' semantic, morphological, and spatial characteristics. The neural network gives an initial segmentation map (Fig. 1d). After obtaining the initial segmentation map, a CRF is used to refine it. As a result, we get the final crowd segmentation based on the individuals' head location (Fig. 1e).

After obtaining the final segmentation maps, some metrics based on the relations between ground truth and obtained maps are calculated to evaluate the final prediction accuracy. Based on such metrics, we can compare the impact of various parameters on the final semantic segmentation.

A. ShanghaiTech Dataset

For the experiment, we use a highly recognized ShanghaiTech dataset [32]. It consists of two parts. Part A features 482 crowd images taken from the Internet. In each image, there are from 33 to 3138 individuals, and the majority of the images represent dense crowds. Part B consists of 716 images. The images contain mainly sparse crowds from 9 to 576 people. Hence, we decided to use the B part to train the model for binary segmentation (crowd and non-crowded regions), and the A part for ternary segmentation (dense crowd, sparse crowd, no crowd). In both parts, 100 images form the training sample, 100 images — the validating sample and other ones are used to test the trained model.

To generate ground truth segmentation, we build density maps first. We do so by using Gaussian blurring (Fig. 1b). After that, we segment density maps based on thresholding values. We use two thresholds:

$$\mu_1 = 0.001M,$$
$$\mu_2 = 0.01M,$$

where M stands for the maximal value in the considered density map. For binary segmentation, only μ_2 is used (Fig. 1c).

B. UNet

UNet, one of the deep learning networks with an encoder-decoder architecture, is a popular neural network architecture designed for semantic segmentation tasks, particularly in biomedical image segmentation [21]. It makes maximal use of feature maps in full scales for accurate segmentation and efficient network architecture with fewer parameters.

The architecture is characterized by its U-shaped design, consisting of a contracting path (encoder) followed by an expansive path (decoder) allowing for precise localization while capturing contextual information (Fig. 2). The EfficientNetB3 was used as the encoder [33], [34]. It consists of convolutional (Fig. 2, blue blocks) and bottleneck layers (Fig. 2, yellow blocks) [35]. In the middle column of Fig. 2, dashed arrows denote skip connections between layers in the encoder and the decoder which help the decoder part recover spatial information lost during downsampling in the encoder part. In the research, we used Segmentation Models, a Python library with Neural Networks for Image Segmentation based on well-known Keras and TensorFlow frameworks [36].

The input images must have 3 channels (e. g. RGB) and have the same size. To effectively use the network, we resize initial images and the corresponding ground truth segmentations to 256x256. This allows us to get the results quickly without significant quality loss. However, other sizes could also be used, e. g. 224x224 which is a standard input size for EfficientNetB3, as well as bigger sizes. Also, the following parameters were used during the training: batch size is 5, optimizer function is Adam, the learning rate is 0.0001, the activation function on the last layer is softmax, and the number of epochs is 50.

Besides, we consider two loss functions that take into consideration the spatial characteristics of segments. Dice loss is widely used as a metric showing how much two images are similar to each other [23], [37], [38]:

$$GDL = 1 - 2\frac{\sum\limits_{l=1}^{C} w_l \sum\limits_{n=1}^{N} r_{ln} p_{ln} + \varepsilon}{\sum\limits_{l=1}^{C} w_l \sum\limits_{n=1}^{N} (r_{ln} + p_{ln}) + \varepsilon},$$

where N is the number of pixels on each of two compared images, C is the number of classes, $p_l n$ and $r_l n$ are probabilities for the n-th pixel from both images to be in the l-th class, w_l is a normalizing coefficient, ε is a term to avoid division by zero. In the paper, value $\varepsilon = 10^{-7}$ is used.

Focal loss is the metric widely used in image classification and segmentation tasks [37], [39]. It derives from the cross entropy concept and addresses the one-stage object detection scenario in which there is an extreme imbalance between foreground and background classes during training. The loss function is calculated according to the formula:

$$FL(p) = -(1-p)^{\gamma} \log(p),$$

where p is the model's estimated probability for a pixel to belong to a certain class, and γ is the focusing parameter to down-weight easy examples and focus on training on hard ones. We used the default value which is $\gamma = 2$.

Both functions are suitable for binary and ternary semantic segmentation. Besides, they require only a few parameters to define, so the models don't become too hard to tune.

C. CRF

In the research, we use PyDenseCRF, a Cython-based Python wrapper for a fully connected CRF with a highly efficient approximate inference algorithm implemented in which the pairwise edge potentials are defined by a linear combination of Gaussian kernels [28], [40]. Concepts of appearance and smoothness are used in the network to calculate *a posteriori* probabilities for each pixel belonging to each class. Appearance is the property of a segmentation map to have nearby pixels of the same color likely belonging to the same class. In smooth models, large classes must absorb small isolated regions nearby. The formalization of both concepts can be expressed by the formula:

$$k(\vec{f_i}, \vec{f_j}) = w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + \\ + w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right),$$

where p_i and p_j – positions of two pixels, I_i and I_j – their colors, $\vec{f_i}$ and $\vec{f_j}$ – their features vectors, k – the similarity function to be maximized, $w^{(1)}$ and $w^{(2)}$ – linear combination weights, θ_α, θ_β, θ_γ are initially defined parameters. In the research, we used the following values: $\theta_\alpha = 10$, $\theta_\beta = 20$, $\theta_\gamma = 1$.

D. The Results Processing

After evaluating the final segmentation maps, some metrics functions are calculated to perform pixel-wise comparison ground truth maps with obtained results.

For binary segmentation, we calculated four values for each pair:

$$\begin{aligned} TP &= |\{(i,j) : p_{ij} = 1 \land \hat{p}_{ij} = 1\}|, \\ FP &= |\{(i,j) : p_{ij} = 1 \land \hat{p}_{ij} = 0\}|, \\ TN &= |\{(i,j) : p_{ij} = 0 \land \hat{p}_{ij} = 0\}|, \\ FN &= |\{(i,j) : p_{ij} = 0 \land \hat{p}_{ij} = 1\}|, \end{aligned}$$

where (i, j) stands for the position of two pixels to be compared, $\hat{p}_{ij}$ is the value for the ground truth pixel (1 stands for crowded region, and 0 means the pixel belongs to non-crowded area), and p_{ij} is the value for the pixel on predicted map. After that, accuracy, crowd predictive value, and non-crowded predictive value are calculated:

$$\begin{aligned} acc &= \frac{TP+TN}{TP+FP+TN+FN}, \\ cpv &= \frac{TP}{TP+FP}, \\ npv &= \frac{TN}{TN+FN}. \end{aligned}$$

Besides, for each image, we calculated the number of annotation labels that fell into each class (according to ground truth and predicted segmentation maps):

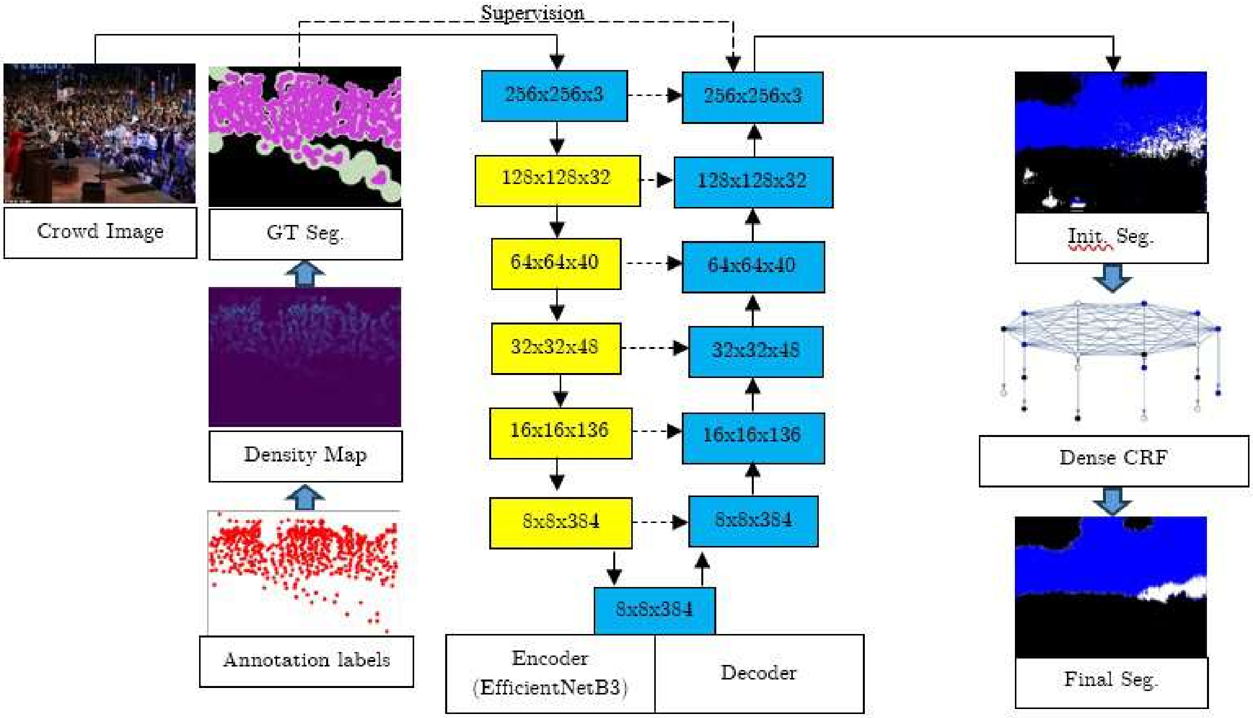

Figure 2. The crowd segmentation prediction framework consisting of image preprocessing (left column), the UNet network (middle column), and dense CRF to refine the predicted segmentation (right column)

$$\begin{aligned} N_0 &= \left|\{i : p_{h_i} = 0\}\right| \\ \hat{N}_0 &= \left|\{i : \hat{p}_{h_i} = 0\}\right|, \\ N_1 &= \left|\{i : p_{h_i} = 1\}\right|, \\ \hat{N}_1 &= \left|\{i : \hat{p}_{h_i} = 1\}\right|, \end{aligned}$$

where N_i stands for the number of annotation labels felt in the class labeled by the number i, and h_i is the location of the i-th annotation label. Then, to estimate the rate of the crowd detection, we calculated the $N_1/\hat{N}_1$.

For the ternary segmentation, the following characteristics are calculated:

$$\begin{aligned} DD &= \left|\{(i,j) : p_{ij} = 2 \wedge \hat{p}_{ij} = 2\}\right|, \\ DS &= \left|\{(i,j) : p_{ij} = 2 \wedge \hat{p}_{ij} = 1\}\right|, \\ DN &= \left|\{(i,j) : p_{ij} = 2 \wedge \hat{p}_{ij} = 0\}\right|, \\ SD &= \left|\{(i,j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 2\}\right|, \\ SS &= \left|\{(i,j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 1\}\right|, \\ SN &= \left|\{(i,j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 0\}\right|, \\ ND &= \left|\{(i,j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 2\}\right|, \\ NS &= \left|\{(i,j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 1\}\right|, \\ NN &= \left|\{(i,j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 0\}\right|, \end{aligned}$$

where 0 stands for non-crowded pixels, 1 is sparse crowd, and 2 is dense crowd. Based on these values, we calculate metrics that we call accuracy, dense crowd predictive value, crowd predictive value, and non-crowded predictive value:

$$\begin{aligned} acc &= \tfrac{DD+SS+NN}{DD+DS+DN+SD+SS+SN+ND+NS+NN}, \\ dpv &= \tfrac{DD}{DD+DS+DN}, \\ cpv &= \tfrac{DD+DS+SD+SS}{DD+DS+DN+SD+SS+SN}, \\ npv &= \tfrac{NN}{ND+NS+NN}. \end{aligned}$$

In the same manner, as for binary segmentation, we calculate the number of annotation labels corresponding to all three predicted classes:

$$\begin{aligned} N_0 &= \left|\{i : \hat{p}_{h_i} = 0\}\right|, \\ \hat{N}_0 &= \left|\{i : \hat{p}_{h_i} = 0\}\right|, \\ N_1 &= \left|\{i : \hat{p}_{h_i} = 1\}\right|, \\ \hat{N}_1 &= \left|\{i : \hat{p}_{h_i} = 1\}\right|, \\ N_2 &= \left|\{i : \hat{p}_{h_i} = 2\}\right|. \\ \hat{N}_2 &= \left|\{i : \hat{p}_{h_i} = 2\}\right|. \end{aligned}$$

Based on such characteristics, we can calculate the rate of dense crowd detection equalling $N_2/\hat{N}_2$.

E. Crowd Dense Semantics

After obtaining and refining the segmentation maps (Fig. 3a), we clusterize the people in crowd images based on their positions provided as annotations for the considered dataset. For this, we divide all annotation points according to the corresponding connected areas of resultant segmentations (Fig. 3b) and clusterize each group separately (Fig. 3c). We don't clusterize points falling into non-crowded areas. Any clustering method for 2D points can be used. We decided to use DBSCAN as it demonstrated its benefit in various researches in logistics, spatial analysis, and behavior patterns detection [41], [42].

After the clustering, all points are divided into multiple clusters represented as convex hulls of the points inside them (Fig. 3c). Each cluster is characterized by its location, density, people count, as well as spatial connections

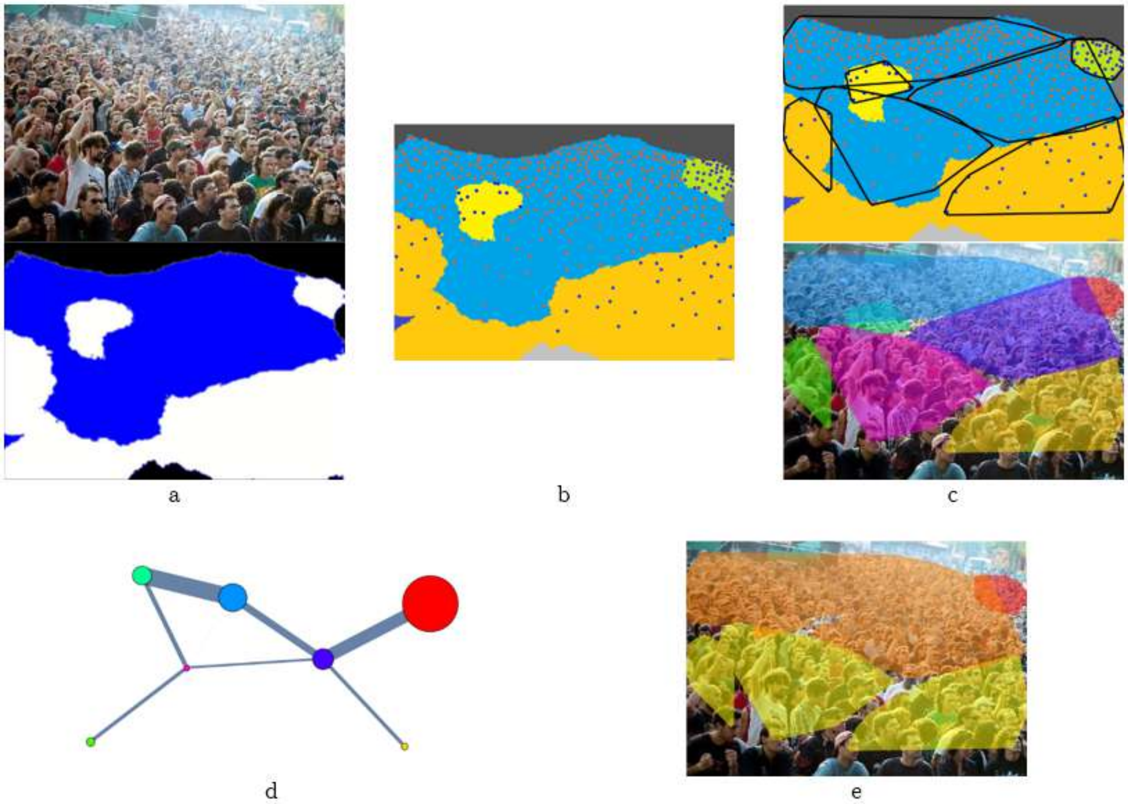

Figure 3. Semantics in a crowd: the initial image and the segmentation map for it (a), connected areas in the segmentation map (b), clusterization of annotation labels within connected areas (c), the clusters connectivity graph (d), crowd semantic clusters (e)

with other clusters. All clusters and such connections are presented as the graph where each vertex is assigned to a cluster, and two vertices are adjacent if only two clusters overlap or share a border (Fig. 3d). The size of a vertex is proportional to the density of the crowd in the corresponding cluster, and the thickness of each edge is proportional to the area shared between two adjacent clusters. Based on the separate clusters' properties and the graph's connectivity, we can interpret semantics for the depicted crowd in the image. In Fig. 3e, the red cluster is a dense crowd (more than 1 individual per 1000 pixels), the orange one is a regular crowd (0.5-1.5 individuals), and the yellow clusters present a sparse crowd (less than 0.7 individuals). Different clusters can share their borders, overlap, or even be a part of another cluster. Considering such facts, we can evaluate the crowd semantics.

IV. Results and Discussion

A. Segmentation Evaluation

After training neural networks and obtaining predicted segmentation maps for images in testing samples (Fig. 1, 3a), we received the results presented in Table I. Statistics on calculated accuracies and predictive values through the testing samples (from ShanghaiTech dataset's part B for binary segmentation and part A for ternary segmentation) are presented there.

As we can see, the CRF refinement significantly improves the ternary segmentations in terms of overall quality (acc.) and dense crowd detection (DPV). In other cases, it didn't demonstrate better results. Comparing dice and focal loss functions' results, we can conclude that the neural network with focal loss can predict crowd regions slightly better (according to the CPV). Nevertheless, dice loss allows us to predict dense crowds slightly better (DPV). Also, binary segmentation gave better results than ternary one in terms of overall performance. However, it takes place due to its ability to detect non-crowded areas (NPV) whereas the ternary segmentation model is more successful in detecting crowded areas (CPV).

Statistics on crowd detection rates are presented in Table II. Here we can see the poor performance of the model using the focal loss function in detecting sparse crowds (low values of $N_1/\hat{N}_1$), but for dense crowds detection, both dice and focal functions results are approximately equal (dice function demonstrates slightly better results though). On average, the model is prone to overlook some parts of sparse and dense crowds (from 3 % to 22% according to average and median results).

Table I
Accuracy and Predictive Values

	UNet Segmentation				**CRF Segmentation**			
	Acc.	CPV	DPV	NPV	Acc.	CPV	DPV	NPV
Binary Segmentation, Dice Lose								
Minimal	0.2629	0.0108	-	0.3144	0.2556	0.0086	-	0.3141
Average	0.7427	0.3962	-	0.8984	0.7440	0.3978	-	0.8987
Median	0.7580	0.3933	-	0.9464	0.7589	0.3939	-	0.9479
Maximal	0.9243	0.8968	-	1	0.9217	0.9001	-	1
Binary Segmentation, Focal Lose								
Minimal	0.3379	0	-	0.3330	0.3389	0	-	0.3322
Average	0.8308	0.4023	-	0.8513	0.8098	0.4185	-	0.8292
Median	0.8671	0.3960	-	0.8891	0.8412	0.4110	-	0.8635
Maximal	0.9785	0.9897	-	0.9965	0.9780	1	-	0.9950
Ternary Segmentation, Dice Lose								
Minimal	0.0685	0.1310	0	0	0.0622	0.1271	0	0
Average	**0.3175**	0.6830	**0.3746**	0.5127	**0.4498**	0.6819	**0.5169**	0.5132
Median	**0.3067**	0.7412	**0.3245**	0.5017	**0.4386**	0.7447	**0.5616**	0.5132
Maximal	**0.7331**	0.9992	0.9992	1	**0.9168**	0.9983	0.9947	1
Ternary Segmentation, Focal Lose								
Minimal	0.0053	0.0466	0	0	0.0534	0.0899	0.0002	0
Average	**0.3178**	0.7028	**0.2437**	0.4929	**0.5026**	0.7014	**0.4890**	0.4765
Median	**0.3012**	0.8016	**0.0771**	0.4913	**0.4952**	0.7785	**0.4911**	0.4693
Maximal	**0.7775**	0.9998	1	1	**0.9381**	1	0.9911	1

Sometimes the rates $N_1/\hat{N}_1$ and $N_2/\hat{N}_2$ are greater than 1. It may indicate situations where separate persons are detected as crowds and sparse crowds are recognized as dense crowd regions. According to our calculations, from 15% to 25% images have such values. Hence, the model might both under- and overestimate the crowd density. Also, data from Table II prove again using the CRF is crucial for ternary segmentation prediction.

Table II
Crowd Detection Rates

	UNet segmentation		**CRF segmentation**	
	Dice	Focal	Dice	Focal
Binary Segmentation ($N_1/\hat{N}_1$)				
Minimal	0.0097	0	0	0
Average	0.7896	0.1719	0.7909	0.1687
Median	0.8454	0.1189	0.8473	0.0872
Maximal	1.3235	0.8868	1.3235	0.9057
Ternary Segmentation ($N_2/\hat{N}_2$)				
Minimal	0	0	0.0402	0
Average	**0.5298**	**0.0199**	**0.8785**	**0.7896**
Median	**0.5215**	**0**	**0.9704**	**0.8517**
Maximal	1.2971	1	1.2971	1.2971

B. Crowd Semantics

Most of the ShanghaiTech datasets images are captured by a camera observing a nearby scene from above. Hence, the typical clusterization consists of distant dense clusters, closer regular clusters, and near sparse clusters (Fig. 4a). However, the crowd in an image can be divided vertically if there is a tall object like a pole or a flag (Fig. 4b).

Sometimes the pattern doesn't hold, which could indicate a group with interest. Some examples where we can detect a people's interest include:

- Multiple clusters with equal density spanning most of the image (Fig. 4c). Those usually present a uniform crowd with regular attention.
- A small cluster within or near a bigger one of the different type (Fig. 4, d-e). Those situations usually present a concentration of interest in particular groups within or near the crowd.
- A significant overlapping between clusters of different types. (Fig. 4f). This one can indicate a spreading interest or joining the people groups. Real-time surveillance systems must detect such actions to prevent any dire situations.
- Elongated clusters presenting regulate or dense crowds may indicate the presence of a queue in the region (Fig. 4g, the bottom orange cluster). If the density is high enough, some extraordinary situations might take place like queue crushes or evacuation panic which must be dealt with immediately.
- A wide sparse cluster at the bottom of the image might indicate a group of people that is very close to the observer (Fig. 4, d, h). Overlapping between the close cluster and other, distant ones is another feature of such a situation. Depending on the people's behavior, such a close group might be considered an outlier or an interest group, especially when it grows or approaches the observer.

V. Conclusions

This paper presents an approach for semantic segmentation of dense and sparse crowd images, addressing the critical

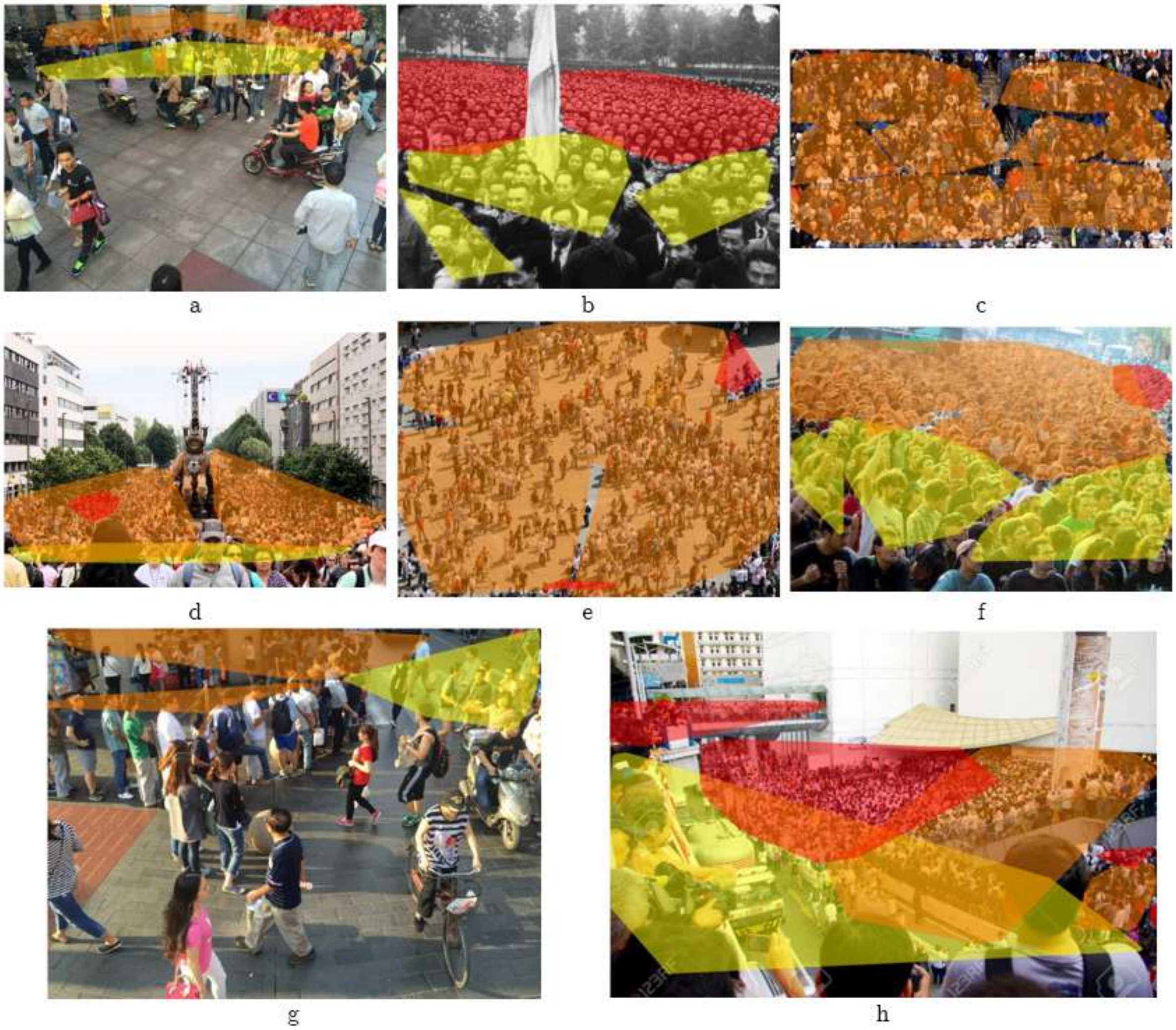

Figure 4. Semantic segmentation of crowds of different types: regular crowds with no attention (a, b), uniform crowd with regular attention (c), diverse crowd containing groups with increased interest (d, e), diverse crowd with a spreading group with interest (f), crowd with a queue (g), crowd with a close cluster (h)

need for accurate crowd analysis in various applications such as crowd management, surveillance, and urban planning. Our proposed method leverages a combination of UNet and CRF networks, augmented by a semi-automatic labeling technique based on Gaussian blur and thresholding methods to generate ground truth maps. Furthermore, we highlight some typical crowd behavior patterns based on clustering the people groups by their density and interconnections between them. Indicating those patterns is important for understanding crowd structures and dynamics as well as establishing crowd management and safety.

Through extensive experimentation and evaluation, we have demonstrated the effectiveness of our approach in accurately segmenting crowd images, particularly in binary segmentation tasks distinguishing crowded from non-crowded regions. While our model excels in binary segmentation, we acknowledge the challenges encountered in ternary segmentation tasks involving dense crowds, sparse crowds, and non-crowded areas. Despite this, our model shows promising results in crowd detection regardless of crowd density. Besides, we prove the necessity of CRF refinement to get better results in ternary segmentation.

References

[1] K. Khan et al., "Crowd Counting Using End-to-End Semantic Image Segmentation," *Electronics*, 2021, vol. 10, no. 11, #1293, doi: 10.3390/electronics10111293.

[2] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, L. Lin, "Crowd counting with deep structured scale integration network," in *2019 IEEE International Conference on Computer Vision (CVF)*, pp. 1774-1783.

[3] S. Sholtanyuk, A. Leunikau, "Lightweight Deep Neural Networks for Dense Crowd Counting Estimation," in *Pattern Recognition and Information Processing (PRIP'2021).* United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, 2021, pp. 61–64.

[4] S. Sholtanyuk, "Finding The Optimal Segmentation of a Crowd Image with Watershed Method," in *Information Systems and Technologies (CSIST'22).* Part 2, Belarusian State University, Minsk, 2022, pp. 217-223. Available at: https://elib.bsu.by/handle/123456789/288544.

[5] F. Abdullah, A. Jalal, "Semantic Segmentation Based Crowd Tracking and Anomaly Detection via Neuro-Fuzzy Classifier in Smart Surveillance System". *Arabian Journal for Science and Engineering*, 2023, vol. 48, no. 2, pp. 2173-2190.

[6] M. Gruosso, N. Capece, U. Erra, "Human Segmentation in Surveillance Video with Deep Learning", *Multimedia Tools and Applications*, 2021, vol. 80, no. 1, pp. 1175-1199.

[7] A. Kroshchanka, E. Mikhno, M. Kovalev, V. Zahariev, A. Zagorskij, "Semantic Analysis of the Video Stream Based on Neuro-Symbolic Artificial Intelligence," in *Open Semantic Technologies for Intelligent Systems*

(OSTIS-2021). Belarusian State University of Informatics and Radioelectronics, Minsk, 2021, pp. 193-204.

[8] A. Kroshchanka, V. Golovko, E. Mikhno, M. Kovalev, V. Zahariev, and A. Zagorskij, “A Neural-Symbolic Approach to Computer Vision,” in *Open Semantic Technologies for Intelligent Systems*, eds.: V. Golenkov, V. Krasnoproshin, V. Golovko, and D. Shunkevich, Cham: Springer International Publishing, 2022, pp. 282–309, doi: 10.1007/978-3-031-15882-7_15.

[9] L. Greco, P. Ritrovato, M. Vento, “On the use of semantic technologies for video analytics,” *J Ambient Intell Human Comput*, 2021, vol. 12, pp. 567–587, doi: 10.1007/s12652-020-02021-y.

[10] P. Anderson, B. Fernando, M. Johnson, S. Gould, “SPICE: Semantic Propositional Image Caption Evaluation,” in *Computer Vision–ECCV 2016*, eds.: B. Leibe, J. Matas, N. Sebe, M. Welling, Cham: Springer International Publishing, 2016, pp. 382-398, doi: 10.1007/978-3-319-46454-1_24.

[11] M. H. T. De Boer, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo, W. Kraaij, “Semantic Reasoning in Zero Example Video Event Retrieval,” *ACM Trans. Multimedia Comput. Commun. Appl.*, 2017, vol. 13, no. 4, Article 60, 17 p., doi: 10.1145/3131288.

[12] Z. Feng, Z. Zeng, C. Guo, Z. Li, “Exploiting Visual Semantic Reasoning for Video-Text Retrieval,” *arXiv preprint, arXiv:2006.08889,* 2020.

[13] S. Munir, S.I. Jami, S. Wasi, “Towards the Modelling of Veillance based Citizen Profiling using Knowledge Graphs,” *Open Computer Science*, 2021, vol. 11, no. 1, pp. 294-304, doi: 10.1515/comp-2020-0209.

[14] X. Guo, M. Gao, G. Zou, A. Bruno, A. Chehri, G. Jeon, “Object Counting via Group and Graph Attention Network,” in *IEEE Transactions on Neural Networks and Learning Systems*, 2023, pp. 1-12, doi: 10.1109/TNNLS.2023.3336894.

[15] L. Greco, P. Ritrovato, A. Saggese, M. Vento, “Improving reliability of people tracking by adding semantic reasoning,” in *IEEE conference on advanced video and signal based surveillance (AVSS)*, IEEE, 2016, pp. 194-199.

[16] K. Humphrey, G. Underwood, G., “Domain knowledge moderates the influence of visual saliency in scene recognition,” *British Journal of Psychology*, 2009, vol. 100, no. 2, pp. 377-398, doi: 10.1348/000712608X344780.

[17] B. Chen, Z. Yan, K. Li, P. Li, B. Wang, W. Zuo, L. Zhang, “Variational Attention: Propagating Domain-Specific Knowledge for Multi-Domain Learning in Crowd Counting,” in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16065-16075.

[18] V. A. Sindagi, V. M. Patel, “HA-CCN: Hierarchical Attention-Based Crowd Counting Network,” in *IEEE Transactions on Image Processing*, 2020, vol. 29, pp. 323-335, doi: 10.1109/TIP.2019.2928634.

[19] W. Liu, M. Salzmann, P. Fua, “Context-Aware Crowd Counting,” in *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5099-5108.

[20] J. Wang, Z. Chen, Y. Wu, “Action recognition with multiscale spatio-temporal contexts,” in *CVPR 2011*. IEEE, June 2011, pp. 3185-3192.

[21] O. Ronneberger, P. Fischer, T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234-241.

[22] V. Badrinarayanan, A. Kendall, R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 12, pp. 2481-2495.

[23] F.I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, “ResUNet-a: a Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data,” *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, vol. 162, pp. 94-114.

[24] J. Long, E. Shelhamer, T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440.

[25] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, J. Malik, “Semantic Segmentation Using Regions and Parts,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2012, pp. 3378-3385.

[26] R. Girshick, J. Donahue, T. Darrell, J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[27] K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv preprint, arXiv:1409.1556*, 2014.

[28] P. Krähenbühl, V. Koltun, “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials,” *Advances in Neural Information Processing Systems*, 2011, vol. 24, pp. 109-117.

[29] X. He, R.S. Zemel, M.A. Carreira-Perpinán, “Multiscale Conditional Random Fields for Image Labeling,” in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2004, doi: 10.1109/CVPR.2004.1315232.

[30] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, 2005, vol. 1, pp. 886-893.

[31] D. Helbing, P. Molnár, “Social force model for pedestrian dynamics,” *Physical Review E*, 1995, vol. 51, no. 5, pp. 4282-4286.

[32] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589-597.

[33] M. Tan, Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *International conference on machine learning*. PMLR, May 2019, pp. 6105-6114.

[34] S. He et al., “An Image Inpainting-Based Data Augmentation Method for Improved Sclerosed Glomerular Identification Performance with The Segmentation Model EfficientNetB3-UNet,” *Scientific Reports*, 2024, vol. 14, no. 1, #1033.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520.

[36] P. Iakubovskii, “Segmentation Models,” *GitHub repository*. Available at: https://github.com/qubvel/segmentation_models (accessed 2024, Mar)

[37] S. Jadon, “A Survey of Loss Functions for Semantic Segmentation,” in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, Oct. 2020, pp. 1-7.

[38] F. Milletari, N. Navab, S.A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” in *2016 International Conference on 3D Vision (3DV)*. IEEE, Oct. 2016, pp. 565-571.

[39] M.S. Hossain, J.M. Betts, A.P. Paplinski, “Dual Focal Loss to Address Class Imbalance in Semantic Segmentation,” *Neurocomputing*, 2021, vol. 462, pp. 69-87.

[40] L. Beyer, “PyDenseCRF,” *GitHub repository*. Available at: https://github.com/lucasb-eyer/pydensecrf (accessed 2024, Mar).

[41] M. Chen, S. Banitaan, M. Maleki, Y. Li, “Pedestrian Group Detection with K-Means and DBSCAN Clustering Methods,” in *2022 IEEE International Conference on Electro Information Technology (eIT)*. Mankato, MN, USA, 2022, pp. 1-6, doi: 10.1109/eIT53891.2022.9813918.

[42] A. Boumchich, J. Picaut, E. Bocher, “Using a Clustering Method to Detect Spatial Events in a Smartphone-Based Crowd-Sourced Database for Environmental Noise Assessment,” *Sensors*, 2022, vol. 22, #8832, doi: 10.3390/s22228832.

АВТОМАТИЗАЦИЯ ОЦЕНКИ ВНИМАНИЯ СКОПЛЕНИЙ ЛЮДЕЙ НА ОСНОВЕ ПОЛУАВТОМАТИЧЕСКОЙ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ СЕТЕЙ UNET И CRF

Шолтанюк С. В., Малёнкин Я. О.,
Лэй Б., Недзьведь А. М.

Семантическая сегментация изображений скоплений людей играет ключевую роль в различных приложениях, таких как управление толпой, наблюдение и городское планирование. В данной статье предложен подход к семантической сегментаций изображений с плотной и разреженной толпой на основе полуавтоматической разметки, использующий комбинацию UNet и условных случайных полей (CRF).

Представлена методика генерации карт сегментации для изображений скоплений людей. Сеть UNet используется для первоначальной, грубой сегментации, после которой следует её уточнение с использованием CRF. Результаты экспериментов показали, что модель лучше выполняет бинарную сегментацию (области, занятые толпой, и области, свободные от толпы), нежели тернарную сегментацию (области плотной толпы, разреженной толпы, и области, свободные от толпы). Однако, в ходе тернарной сегментации получились лучшие результаты по сегментации толпы в целом (без учёта типа толпы). Кроме того, показана значимость уточнения сегментации при помощи CRF в задаче тернарной сегментации толпы.

Также на основе предложенной модели сегментации выделены некоторые закономерности поведения скоплений людей. Они различаются по типу внимания людей, связями внутри скоплений людей и между ними, а также вероятностью возникновения чрезвычайных ситуаций.

Received 25.03.2024