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 domainspecific 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 connection 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.

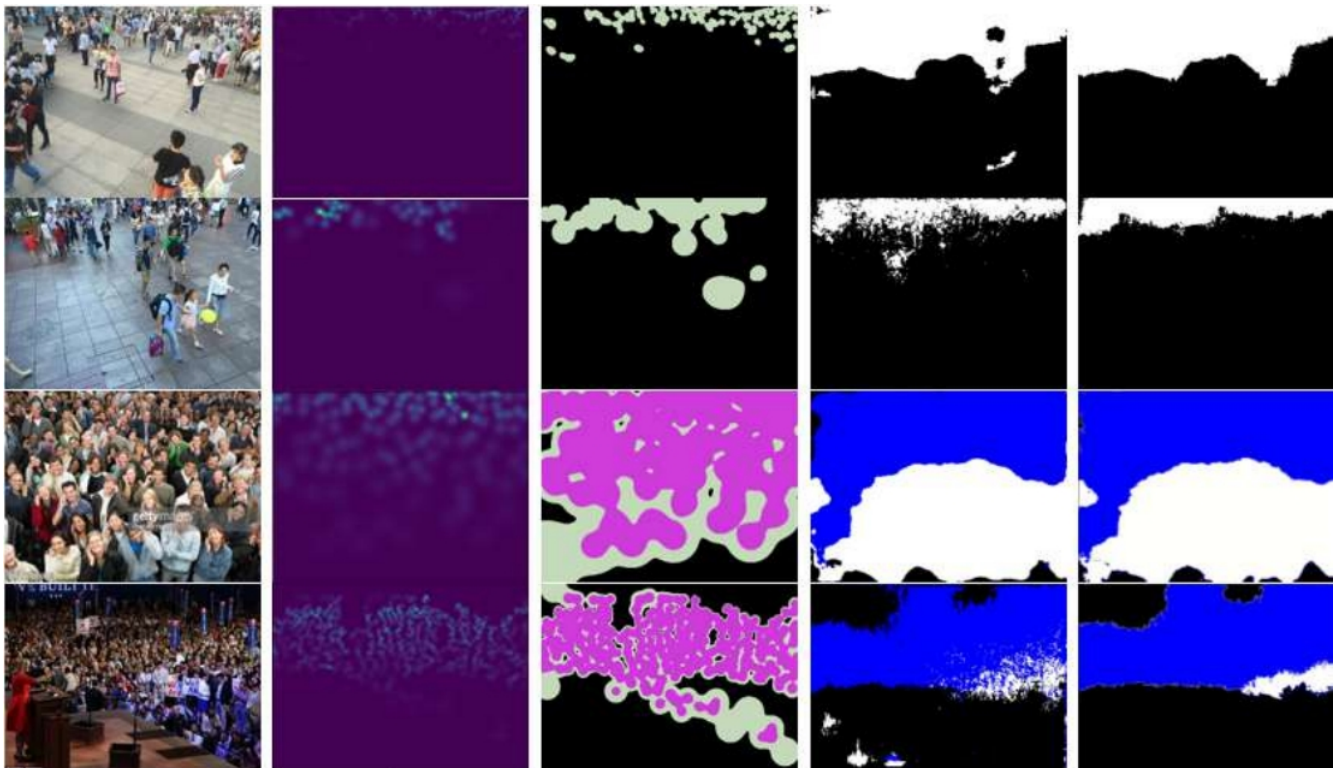
III. Methodology

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).



a b c d e

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 $\mu 2$ 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