IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 August 2024, accepted 6 September 2024, date of publication 11 September 2024,
date of current version 23 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3457800

RESEARCH ARTICLE

Enhancing Manatee Aggregation Counting Through Augmentation and Cross-Domain Learning

MATTEO ZARAMELLA[1], XINGQUAN ZHU[2], (Fellow, IEEE), AND IRENE AMERINI[1], (Member, IEEE)

[1]Department of Computer, Control, and Management Engineering, Sapienza University of Rome, 00185 Rome, Italy
[2]Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

Corresponding author: Matteo Zaramella (zaramella.2025806@studenti.uniroma1.it)

This work was supported in part by the Project PDMI–Accordi Innovazione DM Mise 31/12/2021 and Sapienza University of Rome Project EV2 under Grant 003 009 22; and in part by the U.S. National Science Foundation under Grant IOS-2430224, Grant IIS-2302786, and Grant IIS-2236579.

ABSTRACT In this paper, we propose a novel and enhanced approach for crowd counting within the domain of manatee monitoring, aiming to significantly improve efficiency and accuracy. The proposed model achieves state-of-the-art results in the challenging task of manatee counting, simplifying the work of scientists and experts in the field. Our model not only facilitates the identification and enumeration of manatees in images and videos but also excels in scenarios that pose considerable challenges for human observers. To enhance accurate counting of the manatee aggregation, we introduce a framework with three key innovations to tackle the challenge: a new approach to generate density maps during the training process, an augmented technique to balance the dataset, and a cross-domain solution to enhance overall performance. The proposed two-dimensional Gaussian kernel offers a refined method for creating density maps, providing a more robust foundation for the training phase. Additionally, we built a balanced and augmented dataset, ensuring that the model is exposed to diverse and representative instances, thus improving its generalization capabilities. Furthermore, we incorporate a cross-domain phase pretraining the model utilizing an image dataset of wild animals to initialize the weights and further improve performance. Experiments and comparisons, with respect to previously established CSRNET model presented in Wang et al. (2023), demonstrate noteworthy improvements. Remarkably, our model achieves a Mean Absolute Error (MAE) of nearly half compared to the rival approach, showcasing the substantial advancements achieved through our refined methodology. This progress boosts the reliability of manatee counting in conservation efforts and ecological research.

INDEX TERMS Machine learning, cross-domain learning, convolutional neural networks, crowd counting, manatees.

I. INTRODUCTION

Crowd counting, a widely recognized task, involves the automated counting of individuals, animals, or objects within images or videos. Its importance extends across diverse domains, with applications ranging from public safety and monitoring public spaces to human behavior analysis and video surveillance. The evolution of this task from its pioneering focus on people counting has led to the development of numerous subtasks tailored to count specific animals or objects.

In this paper, we focus on crowd counting with a specific emphasis on manatees. Manatees, characterized by their peaceful behavior, aquatic lifestyle, and herbivorous diet,

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

pose a unique challenge for crowd counting. Notably, the urgency of this task is underscored by the potential threat of manatee extinction in Florida's waters. The alarming statistic of nearly 2000 manatee deaths between 2021 and 2022 [1], attributed to boat collisions and water pollution affecting their seagrass habitat [2], highlights the pressing need for effective monitoring and conservation efforts. Counting manatees becomes intricate due to their characteristics. These gentle creatures, are predominantly gray, making them susceptible to blending into their surroundings, especially when situated near the coast (Fig. 1). The challenge gets harder because of their size changes. In fact, when the image shows a close-up view, they look big, but when it shows a distant view, they look small. Moreover, they often stay in groups, making them hard to count because it's difficult to tell where one animal ends and another begins.

To solve this problem Wang et al. [10] proposed some models trained on a dataset of labeled manatee that represent the state of the art on this task. More specifically the three proposed models were trained on the same dataset, using different density map generation techniques: dots, lines, and Anisotropic Gaussian kernel.

Our solution consists of a new method, more centered on the whole manatee shape and the contrasts with the background. Our method brings three key contributions: the introduction of a novel kernel function designed to represent density maps, the utilization of an augmented dataset to enhance model generalization, and the incorporation of a pretraining phase on a dataset featuring similar background conditions.

More in detail, the introduction of the new kernel function aims to direct the model's attention towards the whole bounding box, consequentially focusing on the entire animal and its differentiation from the background. In addition, the incorporation of a pretraining phase stems from the restricted size of the dataset. By implementing a pretraining phase on a similar dataset to initialize model weights, followed by a fine-tuning phase on the target dataset, performance outcomes are enhanced. Additionally, acknowledging the dataset's imbalance between images, we have chosen to utilize data augmentation techniques. This strategy aims to address the dataset's imbalance, thus facilitating more effective training and ultimately enhancing performance across all categories, regardless of the abundance or scarcity of animal instances within them.

For our pipeline, we employed the CSRNET model [13], a choice motivated by its previous application in manatee-related research as evidenced by the reference to the CSRNET model in the prior study [10]. This deliberate choice not only ensures continuity and comparability with earlier research but also highlights the adaptability and efficacy of the CSRNET model for addressing the challenges posed by manatee detection tasks. The outcomes of our experiments are highly promising, showcasing advanced and state-of-the-art results in the domain of manatee detection. These results underscore the effectiveness of our proposed method in pushing the boundaries of performance for this specific task.

FIGURE 1. Manatees crowd in the Blue Spring State Park, Florida (from Manatee Dataset [33]).

II. RELATED WORKS

Early crowd-counting methods relied on conventional computer vision techniques such as background subtraction [4], blob analysis [3], and feature engineering, but these approaches often struggled with complex scenes, varying lighting conditions, and occlusions. Subsequently, density-based methods emerged as a response to the limitations of traditional approaches. These techniques focus on estimating crowd density maps, utilizing methods like kernel density estimation, Gaussian processes, and Markov random fields [34]. While offering improved performance, these methods still faced challenges with scale variations and crowded scenes.

The advent of deep learning revolutionized crowd counting by enabling automatic feature learning from raw data. Convolutional Neural Networks (CNNs) and their variants, such as VGG [14], ResNet [15], and DenseNet [16], have demonstrated remarkable success in accurately counting crowds. The use of pre-trained models, transfer learning, and fine-tuning further enhanced the adaptability of deep learning approaches to diverse datasets. One of the first deep models for crowd counting was presented in 2015 [5], where they use Convolutional Neural Networks for regression. This method focuses on splitting the image into patches, doing regression on each patch to have the number of people present in each section, and, in the end, summing them to obtain the number of people present in total in the whole image. Another work proposed in the same year by Zhang et al. [18], presents a method based on deep convolutional neural networks (CNNs) to tackle the challenge of crowd counting across diverse scenes. They proposed a convolutional neural network (CNN), trained alternatively with two related learning objectives, crowd density and crowd count. This proposed switchable learning approach is able to obtain a better local optimum for both objectives.

Alternatively, another end-to-end convolutional neural network (CNN) architecture was proposed by Chong et al. [19]. It takes a whole image as its input and directly

outputs the counting result, and it takes advantage of contextual information to predict both local and global counts. In particular, they first feed the image to a pre-trained CNN to get a set of high-level features, then the features are mapped to local counting numbers using recurrent network layers with memory cells. Sam et al. [20] proposed another regression model based on switching convolutional neural networks to leverage variation of crowd density within an image to improve the accuracy and localization of the predicted crowd count. The independent CNN regressors are designed to have different receptive fields and a switch classifier is trained to split the image in patches and relay each crowd scene patch to the best CNN regressor.

MTCNN model, which is based on the work proposed by Zhang et al. [6], uses density map prediction and it is one of the first models to introduce a multi-column structure, that refers to the three cascaded stages, each with its own Convolutional Neural Network (CNN) architecture, working together to achieve robust and accurate results. It became very popular also because it doesn't split the image in patches, but analyzes the whole image and returns the predicted density map.

Features available for crowd discrimination largely depend on the crowd density to the extent that people are only seen as blobs in a highly dense scene. This problem is faced by Sam et al. [28], where presented a growing CNN that can progressively increase its capacity to account for the wide variability seen in crowd scenes. The model starts from a base CNN density regressor, which is trained in equivalence on all types of crowd images. In order to adapt to the huge diversity, two child regressors are created, which are exact copies of the base CNN. A differential training procedure divides the dataset into two clusters and fine-tunes the child networks on their respective specialties. Consequently, the child regressors become experts on certain types of crowds. The child networks are again split recursively, creating two experts at every division. This hierarchical training leads to a CNN tree, where the child regressors are more fine experts than any of their parents. An additional interesting work is presented by Ma et al. [29], who instead of building a new model, used the Bayesian loss. Actually, they used a Bayesian loss function for crowd counting, improving the accuracy by utilizing point-level annotations to better guide the learning process for predicting crowd densities.

PCC Net, by Gao et al. [21], represents a novel method for crowd counting that accounts for perspective distortions in images. They introduce the PCC Net, a spatial convolutional network designed to handle the variations in crowd density and perspective distortions commonly found in real-world scenarios. Their approach improves the accuracy of crowd counts by effectively modeling the spatial relationships and perspective changes within an image. In fact, they designed a perspective module to encode the perspective changes in four directions, namely Down, Up, towards the Left, and Right.

An additional intriguing approach based on VGG-16 is proposed by Sindagi and Patel [23]. It utilizes VGG-16 as a feature extractor and an inverse attention mechanism to effectively identify and count individuals in crowded scenes. By focusing on regions with low attention, the model achieves improved accuracy in estimating crowd densities, even in challenging scenarios.

One extra challenge in crowd counting is the varying scales at which people appear, depending on their distance from the camera. To address this issue, Varior et al. [24] proposed a novel multi-branch scaleaware attention network that exploits the hierarchical structure of convolutional neural networks and generates, in a single forward pass, multi-scale density predictions from different layers of the architecture. To aggregate these maps into a final prediction, they present a new soft attention mechanism that learns a set of gating masks. An alternative proposal for the scale problem is presented in Liu et al. [27], where a novel Deep Structured Scale Integration Network (DSSINet) is presented. This new model addresses the scale variation of people by using structured feature representation learning and hierarchically structured loss function optimization. A supplementary model, the Attentional Neural Field (ANF), is proposed in the work by Zhang et al. [25]. The ANF is an encoder-decoder network composed of conditional random fields (CRFs) and an attention mechanism. More in specific, conditional random fields (CRFs) are present to aggregate multi-scale features, to build more informative representations, and to better model pair-wise potentials in CRFs incorporate a non-local attention mechanism implemented as inter- and intra-layer attentions to expand the receptive field to the entire image respectively within the same layer and across different layers, which captures long-range dependencies to conquer huge scale variations. A further innovative network was proposed by Liu et al. [26], the Cross-stage Refinement Network (CRNet). It can refine predicted density maps progressively based on hierarchical multi-level density priors. In particular, CRNet is composed of several fully convolutional networks stacked together recursively, so that the previous output is the next input, and each of them serves to utilize the previous density output to gradually correct prediction errors of crowd areas and refine the predicted density maps at different stages. Another creative approach was presented by Bai et al. [22], which utilized an adaptive dilated convolutional network combined with a self-correction supervision mechanism. This method addresses the issue of varying crowd densities by adaptively adjusting the dilation rates of convolutions, allowing for more accurate feature extraction across different scales. The self-correction supervision further refines the counting accuracy by iteratively correcting the network's predictions.

Recently other state-of-the-art models were published, each one obtaining great results in different types of datasets, like DSNet [7], SASNet [8], and TransCrowd [9]. The DSNet model, presented by Dai et al. [7], is composed of

the initial ten layers of VGG-16, along with three dense dilated convolution blocks (DDCBs) featuring dense residual connections (DRCs). Additionally, three convolutional layers are employed for the regression of crowd density maps. The purpose of integrating the dilated convolution blocks with dense residual connections is to enhance scale diversity and broaden the receptive fields of features, enabling the model to effectively address variations on large scales and achieve precise estimation of density maps. The SASNet presented in the work proposed by Song et al. [8], is based instead, on a U-shaped backbone for feature extraction, to capture diverse feature representations at multiple levels for a given image. These features are input into an attention layer to generate multi-level confidence maps and density maps. In the final step, guided by the multi-level confidence maps, the density maps are integrated at different levels through a weighted average to derive the final result. TransCrowd, presented by Liang et. al. [9], takes the initial image, splits it into patches of a fixed size, and each patch is subjected to linear embedding along with position embeddings. The resulting sequence of feature embeddings is then passed through a Transformer encoder, after it, a regression head is employed to generate the count prediction. One more transformer-based solution is developed by Lin et al. [30]. They proposed a graph-modulated transformer to enhance the network by adjusting the attention and input node features respectively based on two different types of graphs. Firstly, an attention graph is proposed to diverse attention maps to attend to complementary information, built upon the dissimilarities between patches. Secondly, a feature-based centrality encoding is proposed to discover the centrality positions or importance of nodes.

Another trending field in crowd counting is semi-supervised and unsupervised methods, due to the fact that supervised crowd counting relies heavily on costly manual labeling, which is difficult and expensive, especially in dense scenes. Many works have been published on this topic, like Crowdclip proposed by Liang et al. [31] and the work presented by Ding et al. [32]. Crowdclip is based on the idea that there is a natural mapping between crowd patches and count text In fact, it uses the CLIP pre-trained vision-language model, adjusting the image encoder by using text prompts that rank crowd images based on ordinal relationships. Instead in the Unsupervised Cross-Domain work, they propose a cross-domain learning network to learn the domain gaps in an unsupervised learning manner. More in-depth it firsts explicitly measures the distances between the source domain features and the target domain features and aligns the marginal distribution of their features and then removes domain-specific information from the extracted features and promote the mapping performances of the network.

In the realm of manatee counting, the state-of-the-art model is presented in the paper written by Wang et al. [10]. This paper presents three different CSRNet models [13], trained using three distinct methodologies for generating density maps: dot representation, line representation, and an Anisotropic Gaussian Kernel. However, each of these approaches utilizes a 1-dimensional representation of the manatee, which proves to be quite constraining. While the Anisotropic Gaussian Kernel aims to represent the manatee in a 2-D space, it has limitations because one dimension consistently dominates, creating a slightly thicker line, that essentially enlarges a little bit the 1-dimensional representation. Furthermore, the dataset is labeled to optimize this 1-dimensional encoding, although at the expense of losing significant information.

For our study, we utilized the CSRNet model [13], presented by Wang et al. [10] as one of the state-of-the-art models for the manatee counting task, furthermore facilitating easy comparison of the obtained results.

III. SYSTEM DESIGN

The typical method to approach a crowd-counting task, applicable to various counting fields, involves starting to generate density maps. This is achieved through encoding techniques such as dots, lines, or occasionally kernel functions, ensuring that the sum of the density map pixels corresponds to the total count of objects in the image. The model is then trained by inputting the original image containing the objects to be counted and expecting a density map as output. This initial approach is rudimentary, but it can be enhanced by adding two additional stages to the procedure and employing a 2-D kernel function for better density map creation.

In this paper, we propose a novel and more robust pipeline. Our enhancements focus on three main areas: density map generation, data augmentation, and cross-domain learning application. This methodology transcends specific domains, making it adaptable to a wide range of counting tasks beyond just manatees. Additionally, we include a detailed model section to describe the functionality and implementation of our pipeline.

A. DENSITY MAP GENERATION

Density maps are crafted by leveraging ground truth data. We introduced an innovative kernel function for the generation of density maps, prioritizing the complete coverage of bounding boxes. This adjustment enables the model to shift its focus from a confined area in the middle of the animal, like the dot or the line representations, to include the entire body. Furthermore, the kernel function addresses the differences in contrasts and color variations between the object and its background. Our 2-D kernel function, rooted in a Gaussian kernel, builds a shape for each bounding box with higher values at central points and diminishing values towards the edges. This design facilitates a nuanced representation, ensuring that the sum normalizes to one (Fig. 2). Furthermore, this approach results in a more evenly distributed arrangement of pixels on the density map, simplifying the model's learning process.

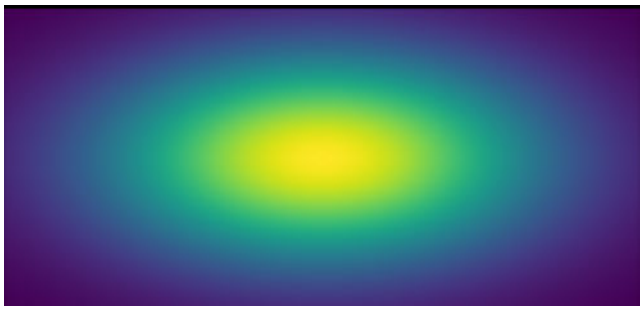

FIGURE 2. Example of our kernel function applied on a bounding box(300 × 80 pixels). This image shows the spreading of the values, higher in the center and near zero in the corners.

We center our attention on datasets containing bounding boxes, where we consider the height and width of each bounding box as hyperparameters. These bounding boxes come in rectangular shapes, with dimensions that vary. Consequently, we analyze the height and width of each bounding box to devise a tailored representation that fully captures its characteristics.

So, let width and height be parameters of the bounding box, μ be the mean vector, σ be the standard deviation vector, and (x, y) be the coordinates of the grid points, the 2D Gaussian distribution is given by the Equation 1.

$$\text{Gaussian}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right) \tag{1}$$

where:

- μ_x and μ_y are the mean values in the x and y directions, respectively.

- σ_x and σ_y are the standard deviation values in the x and y directions, respectively.

The mean vector μ is given by the Equation 2.

$$\mu = \left[\frac{\text{width}}{2}, \frac{\text{height}}{2}\right] \tag{2}$$

The standard deviation vector σ is given by the Equation 3.

$$\sigma = \left[\frac{\text{width}}{4}, \frac{\text{height}}{4}\right] \tag{3}$$

Once calculated the Gaussian, we normalize it as show in Equation 4.

$$\text{Gaussian}(x, y) = \frac{\text{Gaussian}(x, y)}{\sum_{x=0}^{width-1} \sum_{y=0}^{height-1} \text{Gaussian}(x, y)} \tag{4}$$

This normalization ensures that the sum of all values in the Gaussian distribution becomes 1, making it a probability distribution over the defined grid.

Applying this formula to each annotated bounding box results in a rectangular representation containing multiple oval shapes. Each oval has pixels with uniform values, which decrease from the center outward, creating a kind of big faded ellipse as illustrated in Fig. 2. After normalization, all the pixels that compose the representation sum to one, maintaining the core idea of crowd-counting density maps. This codification emphasizes the center of the bounding box, where pixel values are highest, but also includes the entire bounding box, capturing the full body of the animal and its distinction from the background. The variation in pixel values strikes a good balance for the model, focusing primarily on the central part to identify the animal while also considering the whole body. This helps the model better locate the animal and increases the accuracy of the count, reducing potential false positives or hallucinations.

B. DATA AUGMENTATION

To achieve optimal results in predicting density maps across all scenarios, whether working with numerous or just a few, it's crucial to ensure balance in the training set. This entails having a roughly equal number of images representing both scenarios. Achieving such balance necessitates employing data augmentation techniques, a well-established strategy in computer vision [35].

Data augmentation involves generating new data from existing ones. In the realm of images, this typically involves applying various transformations such as random changes in colors, brightness, rotation, and other factors to create new images. This approach serves two main purposes: increasing the number of samples within the dataset and enhancing the diversity of training samples, consequently leading to a more robust model.

For our specific application, we propose employing random adjustments in brightness and contrast as part of our data augmentation techniques. Notably, we avoided stretching the density maps or modifying the images' angle, as altering these aspects would create new images not of real-world occurrences, leading the model to learn irrelevant information.

Consequently, to avoid introducing artifacts in our data, the decision to adopt brightness and contrast changes for improvement. These changes guarantee that the shape of the animal is preserved and, with a careful choice of hyperparameter, that no artifacts are introduced.

More in specific, the brightness adjustment increases or decreases the brightness of an image by adding or subtracting a constant value from each pixel's RGB value.

Denote $I(x, y)$ the intensity (brightness) of the pixel at position (x, y) and c as the constant value added to adjust brightness (this value should be random in a predefined interval). Let's underline that positive c will increase brightness, while negative values will decrease it.

The formula for brightness adjustment is shown in the Equation 5.

$$I'(x, y) = I(x, y) + c \tag{5}$$

The contrast adjustment, instead, changes the difference in intensity between pixels, making the image more or less vivid.

Denote $I(x, y)$ the intensity (brightness) of the pixel at position (x, y), m as the mean intensity of the image, and f as the contrast factor, where $f > 1$ increases contrast and $0 < f < 1$ decreases contrast (this value should be random in a predefined interval). The formula for contrast is shown in the Equation 6.

$$I'(x, y) = (I(x, y) - m) \times f + m \tag{6}$$

Here, $(I(x, y) - m)$ adjusts the intensity relative to the mean intensity m, then it's scaled by the contrast factor f and finally, the mean intensity m is added back.

In our specific case, the random values c and f are chosen between 0.5 and 1.5. Such settings can keep the random field big enough to generate different data and not generate biases with duplicate images, but still limited to guarantee also that no artifacts are introduced.

It is imperative to ensure these techniques are applied ethically, specifically when dealing with sensitive ecological research involving manatees. Special care is taken to ensure that augmentation does not detract from the manatees' representation or contribute to biases that could negatively impact conservation efforts. Our methods are checked to avoid any potential harm to the understanding and preservation of manatees.

C. MODEL ARCHITECTURE

After preparing the data and generating the density maps, a pivotal decision, to make the pipeline work, lies in selecting the appropriate model. Typically, in crowd-counting tasks, the model is divided into two key components: feature extraction and density map prediction.

The feature extraction stage is extremely important as it is tasked with capturing crucial information from the image. Typically, a pre-trained CNN architecture like VGG [14], ResNet [15], or MobileNet [17] is employed for this purpose, given their proficiency in extracting hierarchical features from images. Once trained, this segment should isolate the features relevant to the object of interest, enabling the model to concentrate solely on counting the desired objects.

On the other hand, the second component focuses on predicting the density map. It begins with convolutional layers to further process the features obtained from the first stage, facilitating the model in learning spatial patterns and crowd relationships. Subsequently, upsampling layers, such as transposed convolutions or bilinear upsampling, are employed to gradually enhance the resolution of the feature maps. The network ends with a regression layer responsible for predicting the density map. Typically, this layer consists of a single convolutional layer followed by an activation function like ReLU to ensure non-negative density values.

In this process, starting from the features extracted in the initial phase, the model constructs a density map that accentuates the objects of interest. Specifically, the density map is constructed by predicting the positions of the objects and assigning higher values to their locations, while assigning 0 to areas where objects are absent. In our approach, utilizing the 2-D Gaussian kernel function (presented in Section III-A), the model aims to predict a Gaussian shape resembling the object's bounding box, with higher values concentrated at the center and gradually decreasing towards the edges. To generate this density map, the model considers the dimensions of the original image and predicts the value of the density for each pixel.

In this project, we selected the CSRNet network [13] current state-of-the-art for counting manatees, as demonstrated by Wang et al. [10], moreover, this decision enabled us to compare our results with existing research. The CSRNet comprises two main components, as outlined above: initially, the first layers of the VGG16 net [14] finetuned on our dataset, serve as a feature extractor, subsequently, convolutional layers and final bilinear interpolation are employed to generate density estimation and maintain alignment with the original image dimensions.

To delve deeper, CSRNet takes as inputs the images containing crowd scenes captured by cameras. These images undergo the first ten Convolutional Neural Network (CNN) layers for feature extraction, and, through this process, the model identifies patterns and features relevant to crowd density. Subsequently, the network estimates crowd density across different image regions based on the extracted features, and, in the end, it predicts the density at each pixel, generating density maps as outputs.

During the training and testing phases, the model's outputs are compared with ground truth density maps generated using the 2-D Gaussian kernel function to evaluate the level of the predictions in case of test and to improve them in case of training (this process is shown graphically in Fig. 3). This is often made using the Mean Squared Error function or the Mean Absolute Error function.

1) LOSS AND EVALUATION FUNCTIONS

A crucial aspect, following model selection, involves determining the appropriate loss and evaluation functions. In line with the methodology detailed in the Wang et al. [10], we opted for the Mean Squared Error (MSE) as our loss function and the Mean Absolute Error (MAE) as the evaluation metric. These functions are widely employed for both loss calculation and evaluation in crowd-counting tasks, ensuring the robustness and comparability of our results.

To elaborate, the Mean Squared Error (MSE) quantifies the average squared difference between predicted and actual values, and it's calculated as shown in the Equation 7.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{7}$$

where n represents the number of images, Y_i denotes the actual density maps for the i^{th} image, and $\hat{Y}_i$ stands for the predicted density map for the i^{th} image.

On the other hand, the Mean Absolute Error (MAE) provides an average magnitude of errors between predicted and actual values, calculated as the mean of absolute

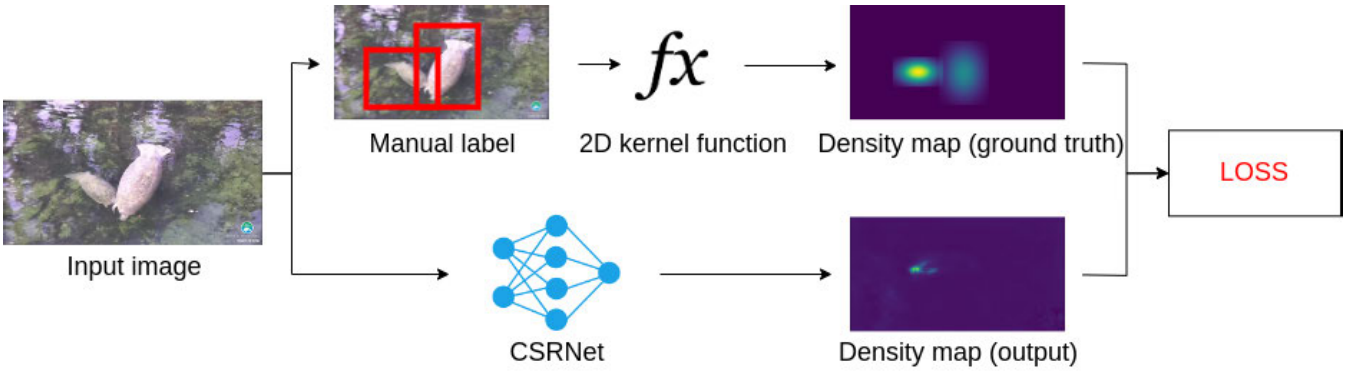

FIGURE 3. Visual representation of the training process, showing the density map generated from the annotations (top) and predicted by the model (bottom).

differences shown in the Equation 8.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{8}$$

Again, n stands for the number of images, while Y_i and $\hat{Y}_i$ represent the actual and predicted density maps for the i^{th} image, respectively.

D. CROSS-DOMAIN LEARNING

In the end, we implemented a cross-domain strategy, specifically focusing on domain adaptation techniques, which are widely utilized to address the challenge of limited labeled data in the target domain. By leveraging labeled data from a related source domain, this approach diminishes the need for extensive labeled data specific to the target domain.

Domain adaptation involves transferring knowledge from one domain to another, often distinct but connected. Consequently, our methodology follows a two-step process: initially, we pre-train the model on a source dataset closely aligned with the target domain, and then, we fine-tune it on the target dataset.

More in specific, the process begins with a labeled dataset, where the model is trained to learn patterns and features. After the pretraining phase, the model is finetuned using the target dataset. When applying the pre-trained model to a different domain, such as the target dataset, disparities emerge due to discrepancies between the two, which is known as the ''domain shift'' problem. For this reason, it is imperative to choose a source dataset with similar backgrounds to those in the target dataset, as backgrounds significantly influence model learning. The pretraining phase it's a very important factor, most of all, in setting the weights of the first part of the model, the feature extraction part (as presented in Section III-C). In contrast, object shapes in the source domain hold less significance, as they represent only a fraction of the images and can be easily learned by the model. As a consequence of this, in the source domain, the model gets used to the dataset and starts localizing and counting the first examples (also if with different shapes). In this phase it's not very important for the accuracy reached but the feature extraction part, where the model learns how to recognize the animals from the background. Once the model has reached several epochs of pretraining or, if set, a threshold of accuracy, it is finetuned in the target dataset. In this second phase, the model applies what was learned in the source domain to the new dataset. So it does not start from scratch but adapts the knowledge that it has to the new domain, becoming more expert to localize and count the requested target animal.

Formally, the source domain is denoted by S with $S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, where x_i^s denotes input data, y_i^s denotes corresponding labels, and n_s is the sample count. Similarly, the target domain is denoted by T with $T = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$, where x_i^t represents input data, y_i^t represents labels, and n_t is the sample count in the target domain dataset. Following pretraining on the source domain S, we identify the optimal model θ' (with minimal error) and fine-tune it on the target domain T to obtain the final model θ^*, as illustrated in Fig.4.

IV. EXPERIMENTS AND RESULTS

In this section, we will conduct experiments in three distinct areas. First, we will focus on manatees: our objective is to train and test our model using the Manatee dataset. Second, we will perform a cross-domain analysis, running an ablation study and comparing results from two different source domains. Finally, we will carry out generalization experiments, where we apply our pretrained model to a similar dataset, the Whale dataset, without fine-tuning, and present the inference results.

A. EXPERIMENTS ON MANATEE DATASET

Our approach to the Manatee counting task involves the method outlined in Section III on the Manatee dataset [33]. In contrast to previous approaches, such as those discussed in Wang et al. [10], which primarily focused on narrow sections of the manatee often confined to small bounding boxes in the middle of its body, our methodology takes a different path. To ensure accuracy and comprehensiveness, we carefully re-annotated the entire dataset. This meticulous manual labeling process resulted in novel and more reliable ground truth labels, prioritizing the capture of the entire animal. These

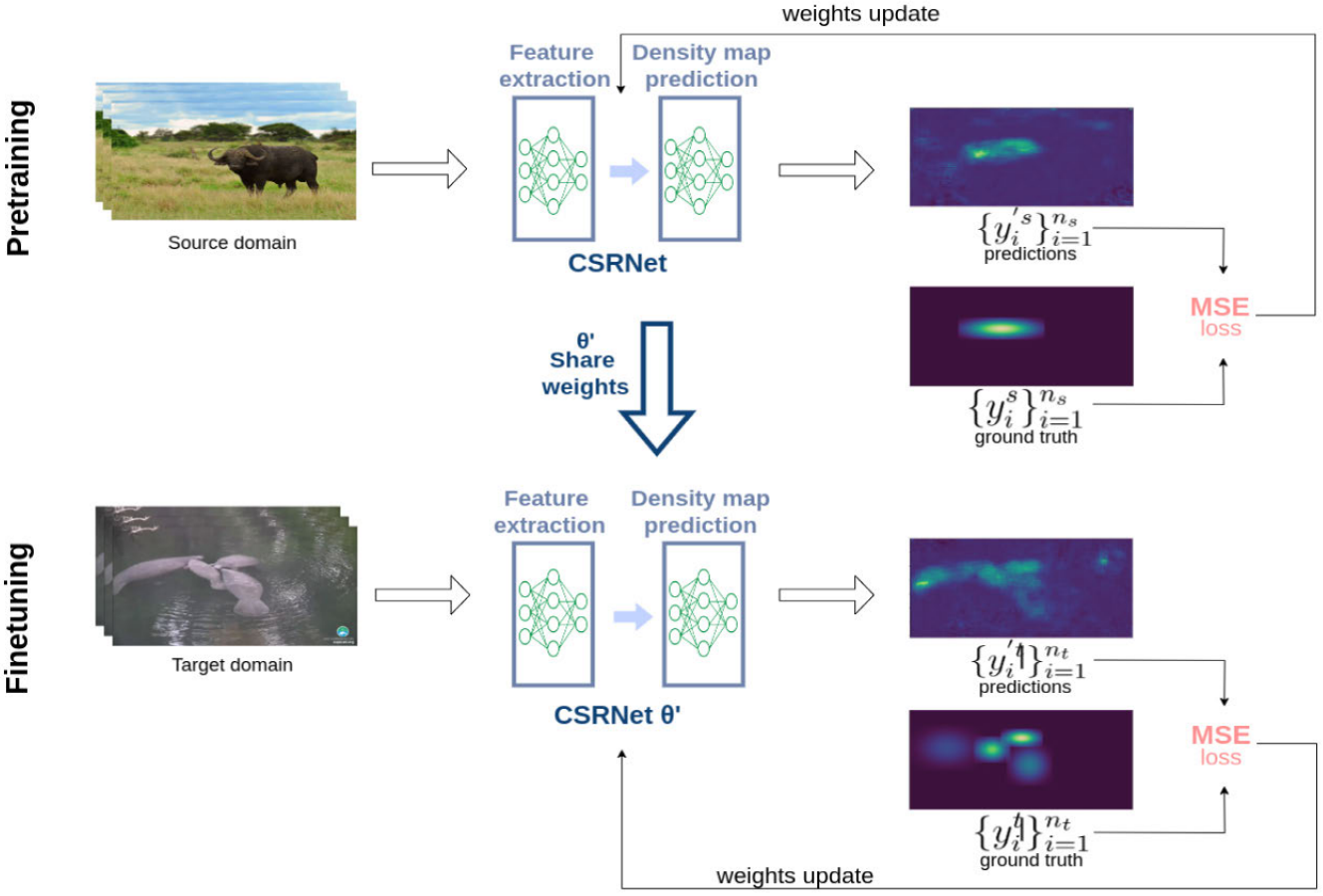

FIGURE 4. Domain adaptation scheme: pretrain the model (CSRNet) on source domain and fine-tune it on target domain.

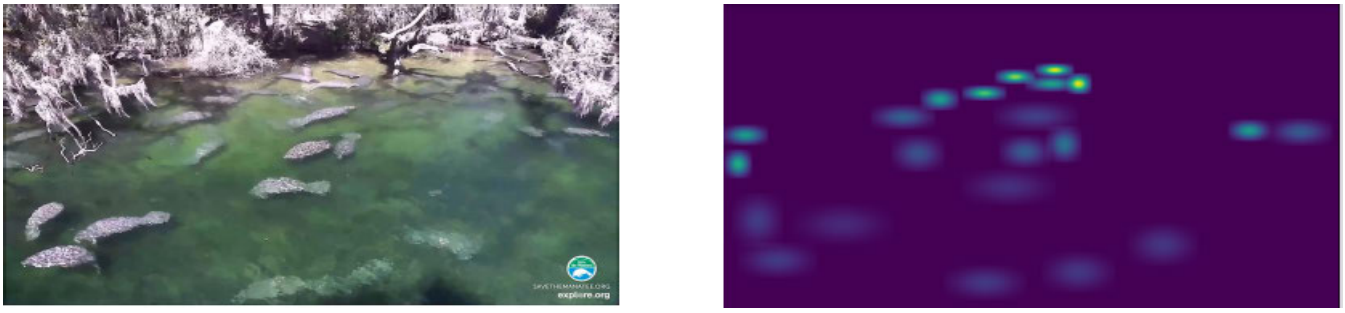

FIGURE 5. Original image (left) and density map generated using our 2-D kernel function (right).

refined labels then formed the basis for generating our new density maps, employing the 2-D kernel function described in Section III-A (an example of a generated density map is shown in Fig. 5).

In terms of the model selection, we opted for the CSRNET, as explained in Section III-C, and we selected Mean Squared Error (MSE) as the loss function and Mean Absolute Error (MAE) as the evaluation function, as described in Section III-C1.

Following the data preparation and model selection, including the choice of model, dataset, loss function, and evaluation criteria, we initiated the training phase.

The Manatee dataset [33] is made of 784 images in total, so its quantity remains insufficient to effectively train a deep model. To address this limitation, we adopted the cross-domain technique outlined in Section III-D, dividing the training process into two stages: pretraining and fine-tuning.

For the pretraining phase, we opted, as source domain, for the African Wildlife dataset from Kaggle [12], which showcases various wild animals such as buffalos, elephants, rhinos, and zebras (Fig. 6). Although these animals possess distinct shapes compared to manatees, they share a commonality: they all dwell in natural settings with

FIGURE 6. Images from the African Wildlife dataset from Kaggle [12].

backgrounds resembling those present in the Manatee dataset [33].

This is a very important aspect for the domain shift problem, so the adaptation of the pretrained model on this domain to the target Manatee Dataset. This issue is managed with the similar background images in both datasets and a higher number of training epochs in the target domain, to adapt it better to it.

This choice is intended to improve the model's flexibility and efficacy when fine-tuning the distinct features of manatee images. So we selected 1504 random images from the African Wildlife dataset, resized them in 1280×720, and split them into 1203 for Training, 150 for Validation, and 151 for Test set.

During the pretraining stage, the model underwent training on the African Wildlife dataset [12] for a total of 550 epochs, achieving a Mean Absolute Error (MAE) of 3.98 on the validation set. Given the nature of the African Wildlife dataset, which predominantly features images with sparse animal populations, this is considered a good outcome. Typically, crowd-counting models excel with denser image data; however, our primary interest wasn't in maximizing performance on this task. Instead, our focus shifted towards the results of the fine-tuning stage.

Considering now the Manatee dataset [33] for the finetuning phase comprises 784 images of manatees captured from an aerial perspective above the water, and these images are distributed into Training, Validation, and Test sets, respectively 80% training and 10% valid and test.

A significant issue arose from the fact that the Training set consisted of 627 images, with only 247 (39.39% of the total train images) featuring fewer than 5 manatees. The mismatch posed a challenge for the model, limiting its accuracy in predicting the number of manatees in images containing only one or a few animals, due to the higher difficulties for crowd-counting models to count when only a few samples are present. To address this limitation, we implemented data augmentation techniques, following the proposed pipeline in Section III-B, applying random changes in brightness and contrast (as shown in Fig. 7), utilizing PIL library [11], to augment the samples with less than 5 manatees present. As illustrated in Fig. 7, these random alterations not only augment the dataset but also improve the visibility of manatees in the images.

We constructed three distinct datasets for our experiment: the original Manatee dataset [33], the Manatee50 adding 50% of augmented images containing less than 5 animals, and the Manatee100 with 100% augmented images containing less than 5 animals. The dataset dimensions are presented in Table 1 together with the results of the finetuning phase on each of them after 750 epochs. For better comprehension and to align our results to the one presented by Wang et al. [10], we divided the data into three primary groups: images with a Low number of manatees (less than 5), Medium manatee density images (between 5 and 20 manatees), and High manatee density images (more than 20 manatees).

Table 1 highlights the challenges faced by the model finetuned on the original Manatee dataset, particularly in accurately identifying instances with few animals, such as in the case of low manatee counts. Conversely, when trained on the Manatee100, the model reached the convergence point faster, after only 327 epochs, but it exhibits a strong bias towards cases with low manatee counts, achieving high accuracy due to the abundance of samples (56,52% of the training set), but neglecting other groups.

Therefore, the most effective alternative appears to be Manatee50 with only half of the images featuring a few animals augmented (it has a number of images with less than 5 manatees equal to 49,33% of the training set). This dataset strikes a balance between the representation of each group, resulting in a more robust training process. Indeed, the model trained on this dataset achieves the best overall performance by focusing on all groups equally and reaching an MAE of 1,69.

Table 1 illustrates significant and unexpected disparities, particularly within the Medium and High groups, between the original Manatee dataset and Manatee50. These differences stem from the implicit instability observed during model training for crowd-counting tasks. Specifically, in our training process, we opted to preserve models with the lowest Mean Absolute Error (MAE) across the Validation set. Consequently, despite some models achieving even lower MAE for the Low group, they were not retained because the overall MAE on the validation set did not surpass the best achieved. This decision, coupled with the training

FIGURE 7. Examples of augmented images changing brightness and contrast (on the left the original image and on the right the augmented one).

TABLE 1. The table shows: the dimension of the Training set, the number of Low augmented samples, and results obtained on the Test set in Low (less than 5 manatees images), Medium (between 5 and 20 manatees images), and High (more than 20 manatees per image) using CSRNet model trained for 750 epochs respectively on original Manatee dataset, 50% (dataset augmented with 50% of low image data on dataset) and 100% (dataset augmented with 100% of low image data on dataset).

Dataset	Training images	Augmented	MAE Low	MAE Medium	MAE High	MAE Total
original	627	0	0.34	0.97	3.24	1.87
Manatee50	750	123	0.32	0.78	2.18	1.68
Manatee100	874	247	0.10	1.43	3.40	1.94

instability, led to results that were not consistently predictable across various training instances. On the other hand, the best MAE reached by the Manatee50 shows how balanced it is between the groups and how much this aspect is important to reach the best performances. The Mean Absolute Error reached by the model training on Manatee50 is remarkably impressive, indicating that our model's predictions are off by slightly more than one manatee and a half per image on average. Indeed, as depicted in the last two rows of Fig. 8, the predicted density maps closely match the ground truths.

Moreover, for a comprehensive analysis of our achieved results, we evaluated our top-performing model (fine-tuned on Manatee50) on the test set, specifying the predicted number of manatees for each group. The results, presented in Table 2, demonstrate consistently low Mean Absolute Error (MAE) across all groups. Higher MAE values observed in the high manatee density group are solely attributed to the larger number of animals in those images.

Examining the alignment between predicted and actual manatee counts reveals a close correspondence across all groups, affirming the model's proficiency in predicting the number of manatees. Despite the imbalanced distribution of images per group in the test set, introduced by a randomized Train-Validation-Test split to enhance reliability, balancing the dataset by adjusting each group to include 19 images does not alter the results. The model maintains its performance, yielding a new test set error of 1.69, underscoring the model's consistency and reinforcing its status as state-of-the-art in this task.

TABLE 2. Manatee counting results with respect to different densities using the CSRNet model trained on Manatee50. Low, Medium, High denote different levels of ground-truth manatee density in each image.

Group	MAE	Tot classified	Tot present	Num images
Low	0.32	79,47	71	26
Medium	0.78	316,39	343	34
High	2.18	576,51	618	19
Total	1.68	972,37	1032	79

In the end, we compared our outcomes with those obtained by Wang et al. [10]. As illustrated in Table 3, our model achieved an improved MAE for each group as well as for the entire test set, achieving a total MAE of 1.68. In contrast,

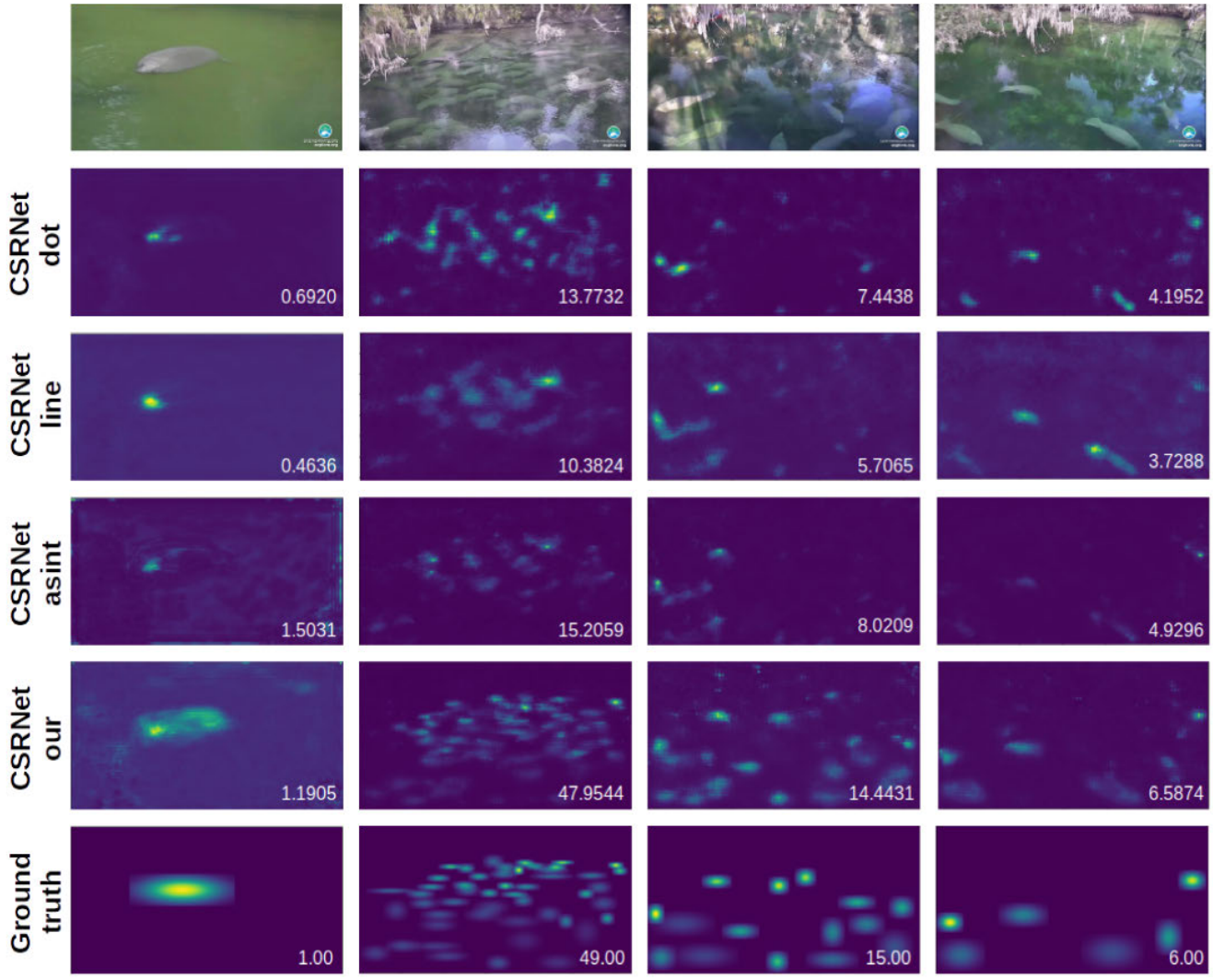

FIGURE 8. Comparison of density maps from the test set using the state-of-the-art CSRNet models (dot, lines, and Anisotropic Gaussian), our and the ground truth generated with the 2-D Gaussian kernel function.

TABLE 3. Experiment comparisons between the proposed method *vs*. a state-of-the-art manatee counting method published by Wang et al. [10]. Notation: *dots*: density maps created with dots, *lines*: density maps created with lines, *anisotropy*: density maps created with anisotropy Gaussian kernel.

METHOD	MAE Low	MAE Med	MAE High	MAE Total
MCNN (dots)	3.64	3.51	26.98	11.37
SANet (lines)	1.51	2.26	4.19	2.65
VGG (anisotropy)	3.40	2.55	15.31	7.08
MARUNet (anisotropy)	3.32	3.83	25.06	9.56
CSRNet (dots)	1.34	2.98	4.97	3.10
CSRNet (lines)	2.25	3.66	5.03	3.65
CSRNet (anisotropy)	1.58	2.94	4.50	3.01
CSRNet with our pipeline	0.32	0.78	2.18	**1.68**

previous state-of-the-art models typically yielded MAE values of around 3, indicating a substantial improvement of half the Mean Absolute Error and establishing a new benchmark for the manatee counting task. This improvement is also shown in Fig. 8, where is visible the difference between the predictions made by our CSRNet model and the preview ones. Our model, not only predicts counts that closely match the actual number of manatees present but also generates density maps that closely resemble the shape of the ground truth. This significant enhancement is attributed to the developed pipeline and the novel techniques applied in data processing. Notably, the most significant advancement introduced is the formulation of the 2-D kernel function, which facilitates a more comprehensive representation of the animals by spreading the values across the image. This allows the model to gain a more accurate understanding of

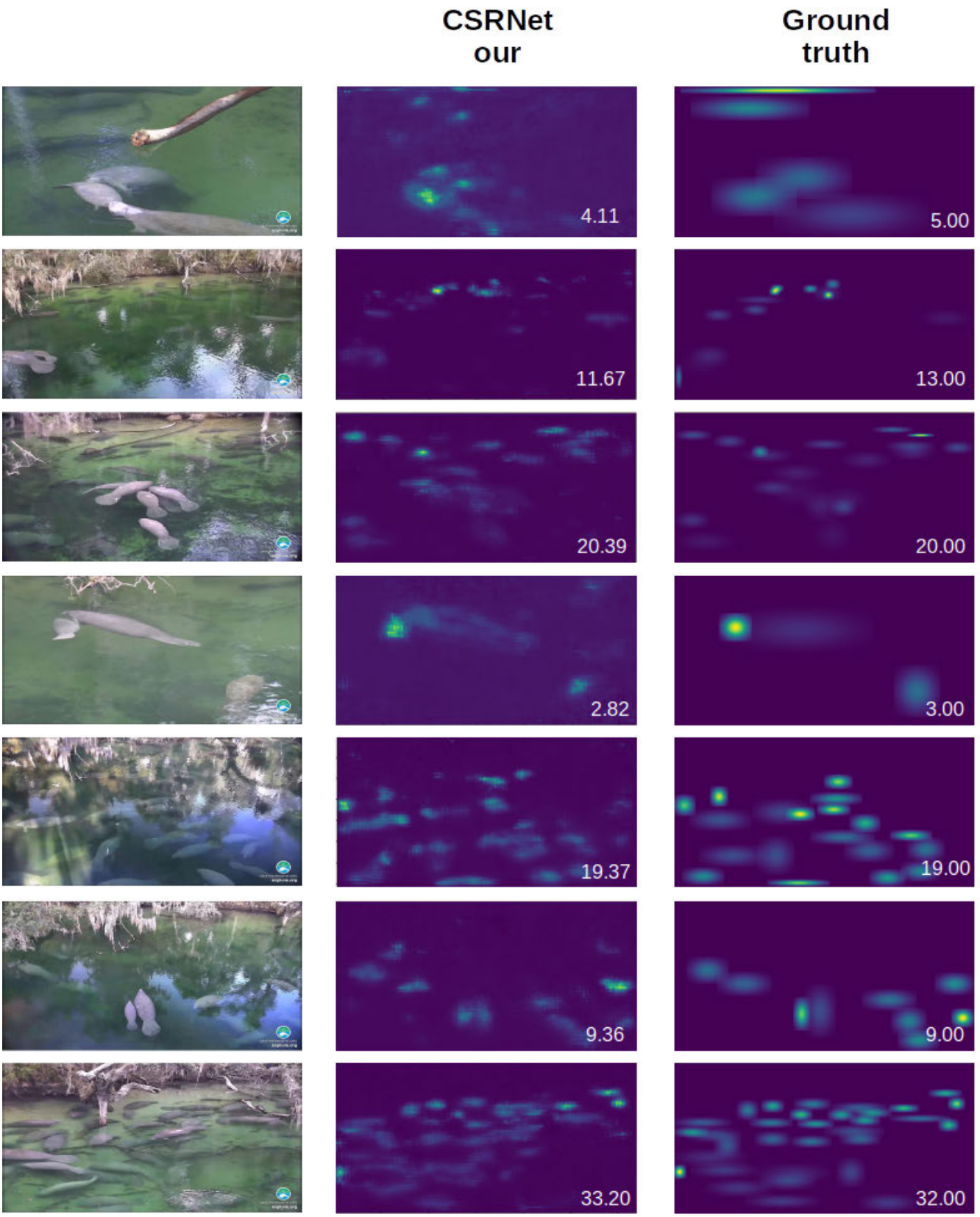

FIGURE 9. Comparison of density maps from the test set using our CSRNet and the ground truth generated with the 2-D Gaussian kernel function.

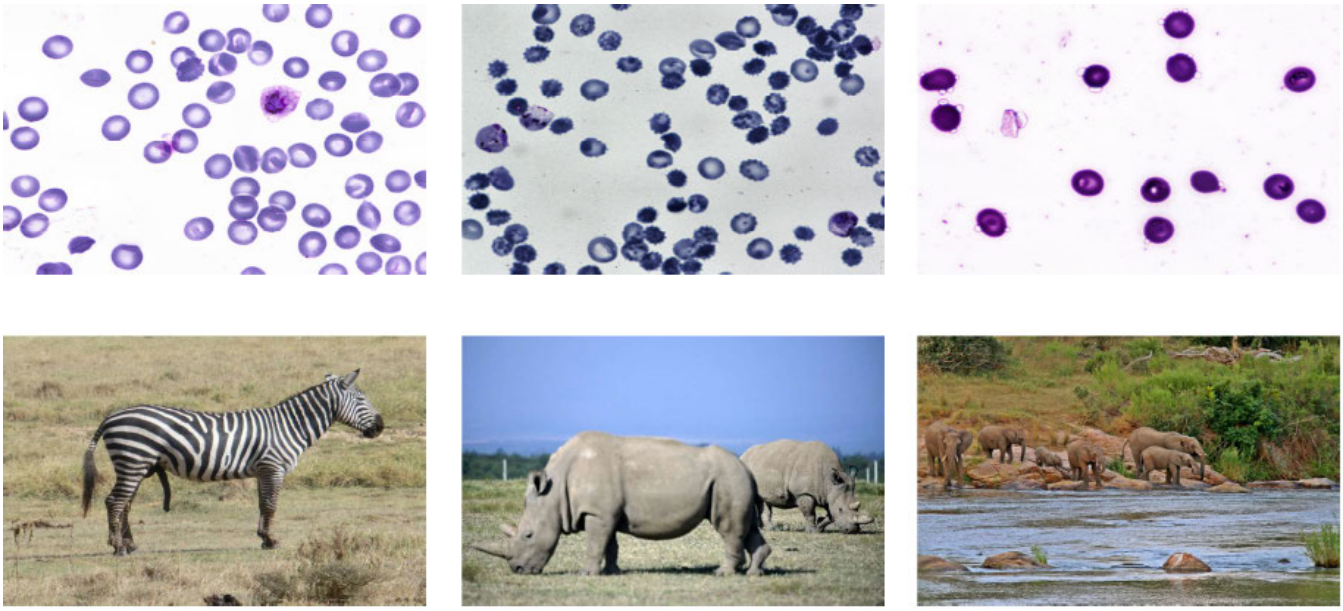

FIGURE 10. Examples of the Malaria Dataset [36] (top row) and the Africa Wildlife Dataset [12] (bottom row).

TABLE 4. The table shows the results achieved from the CSRNet on the Validation set after the pretraining phase (500 epochs) on the corresponding dataset (second column: "Pretraining MAE") and after the finetuning phase (750 epochs) on Manatee50 (third column: "MAE on Manatee50").

Dataset	Pretraining MAE	MAE on Manatee50
Malaria [36]	5.02	2.89
African Wild Life [12]	3.98	1.40

the content of the bounding box. Additional examples to show the accuracy of our model are reported in Fig. 9.

B. ABLATION STUDY ON CROSS-DOMAIN

The project faced major difficulties due to its cross-domain learning. At first, we tried to train the model using a dataset that included objects that have shapes resembling manatees. We believed that the object's shape would significantly impact the model's capacity to generalize well. However, after carrying out the experiments, we realized that this method did not yield the expected results. The key factor in adjusting the model weights correctly during the fine-tuning stage on the target domain was found to be the resemblance in environment and background between the source and target domains. This recognition changed our attention from the form of the items to the situation in which they were seen, emphasizing the significance of training the model on datasets that closely resembled the environmental conditions of the desired domain. This change was necessary to enhance performance during the fine-tuning process and ultimately enhance the accuracy and reliability of the model in spotting manatees in their natural environment.

Two different datasets, the Malaria Dataset [36] and the African Wildlife Dataset [12], are selected for validation. As shown in Fig. 10, these datasets contain very different images. The Malaria Dataset includes images of cells, which can partially resemble manatees in shape since they are round and sparse. However, the background is white, making it quite different from the natural environment in the Manatee Dataset. In contrast, the African Wildlife Dataset features animals with shapes very different from manatees, but the environment is quite similar, sharing similar colors and being set in the wild.

The training phase was done by pretraining the CSRNet model on each one of the two datasets. As shown in Table 4, it reached an MAE (Mean Absolute Error) on the Validation set of 5.02 in the Malaria Dataset and 3.98 in the African WildLife after 500 epochs. The large difference in the Mean Absolute Error is due to the fact that in the Malaria dataset, there are a lot more cells to identify and predict rather than animals in the African one, for this reason, the MAE is higher. After the first phase, we tried to finetune the pre-trained model in the target Dataset, the Augmented Manatee50 one. After finetuning for 750 epochs, the model pre-trained on the Malaria dataset reached a Mean Absolute Error on the Validation set of 2.89, and the one pre-trained on African WildLife of 1.40, as depicted in Table 4. This huge difference shows that the hypothesis done before was correct, that is to say, that the shape of the object is less important or in most cases even irrelevant compared to the importance of the background. So to have a good source domain to apply cross-domain learning it's absolutely mandatory that the background is the same or very similar compared to the one present in the source domain. For this reason for our main experiment, we chose to use the African WildLife Dataset.

C. GENERALIZATION STUDY

To fully evaluate the strength of the proposed pipeline, we tested our pre-trained model, trained on the Manatee

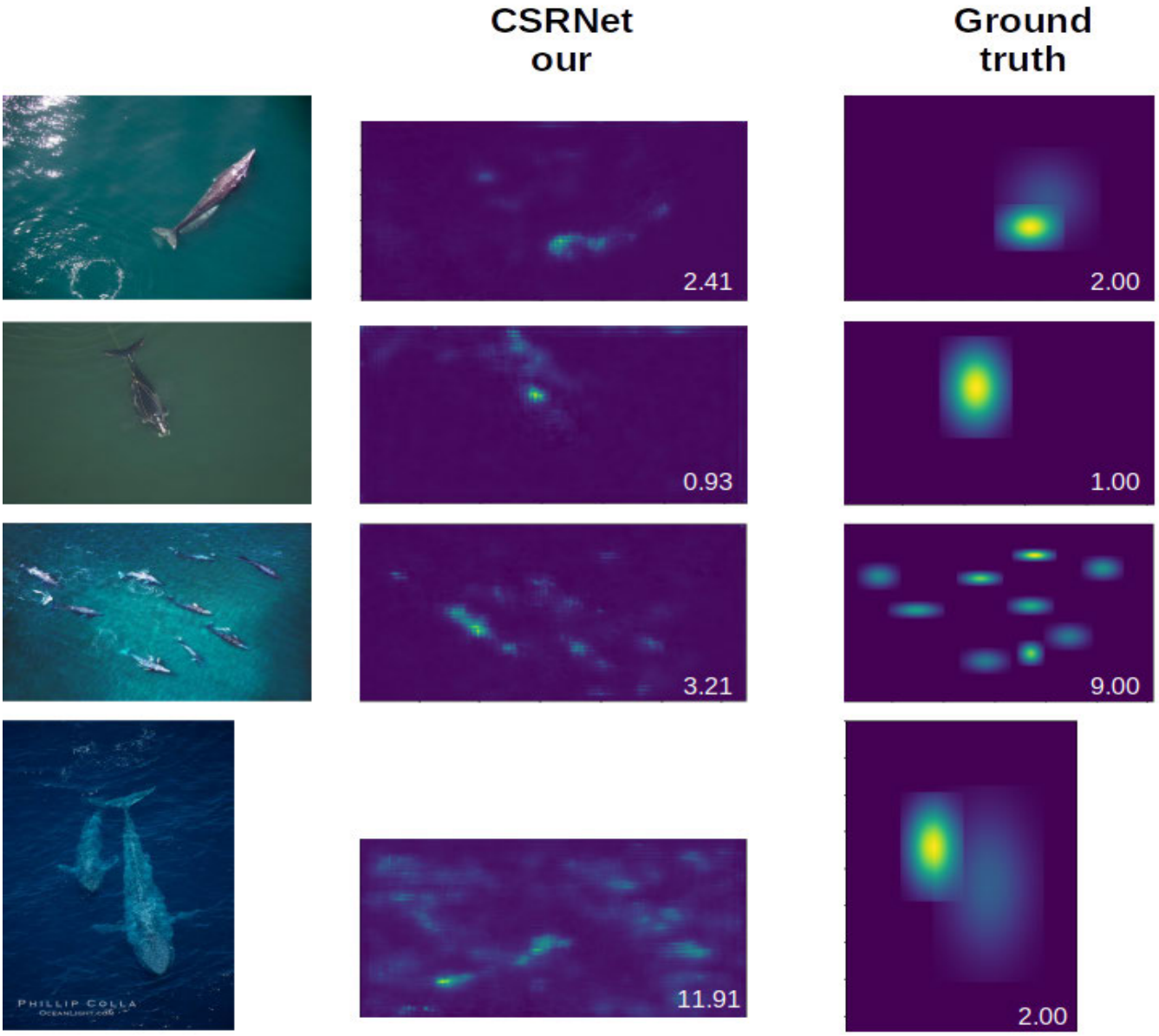

FIGURE 11. Prediction of density maps on the test set of Whale Dataset [37] using our pretrained CSRNet and the ground truth generated with the 2-D Gaussian kernel function (only for visual purposes).

dataset, on a similar domain: the Whale dataset [37]. This test aimed to evaluate the generalization of the model's acquired knowledge to a similar environment. The Whale dataset was chosen due to the similar top-down shape of manatees and whales, providing a comparable visual perspective. However, some images in the Whale dataset are significantly different in backgrounds, as they were taken in the open sea, far from the coast and most of the samples have only a few animals in the scene, imposing challenges to our model.

Due to the limited number of samples in the Whale dataset, we used its original training set to create the test set, which comprised 77 images. The images were reshaped to match the dimensions of those in the pre-trained model. After running the inference, the model achieved a Mean Absolute Error (MAE) of 2.06, demonstrating a good capability to adapt to new tasks. The model performed well on images with backgrounds similar to those in the Manatee dataset but struggled with images that had significantly different backgrounds. Although a model trained specifically on the Whale dataset would likely achieve better performance, our pre-trained model still can serve as a strong baseline, showing good generalization.

The results in Fig. 11 show that the model performs very well on images similar to those in the Manatee dataset, particularly in the first two rows. Specifically, the second image, which has the exact same background as the original target dataset, is predicted almost perfectly. On the other hand, images with significantly different backgrounds are

predicted less accurately. For example, in the last two rows, the model predicts nearly 12 animals where there are only two. The background plays a crucial role and can easily mislead the model, resulting in incorrect predictions. Additionally, in Fig. 11, we have included the density map of the original image to provide a clearer visual understanding.

V. DISCUSSION

The current state-of-the-art models for this task, presented by Wang et al. [10], are trained on the Manatee Dataset, focusing solely on localizing the center of the animal. These models predict density maps generated using dot, line, or anisotropic Gaussian notations. Each notation creates a codification respectively with a dot, line, or slightly thicker line at the center of the bounding box to identify the animal. While Wang et al. achieved good overall results, encoding only the center of the bounding boxes led to a significant loss of information. Another limitation is introduced by the dataset in fact the Manatee Dataset is made by a limited amount of images, which doesn't allow the model to generalize well with new samples.

In this work, we proposed a pipeline to mitigate this loss of information and address the limited data available. Specifically, we introduced a new kernel function to encode the entire bounding boxes without losing information, ensuring the model focuses on the whole body of the animal. Additionally, we re-labeled the dataset to create more precise bounding boxes that encompass the entire animal. Furthermore, to overcome the limited amount of data, we implemented a cross-domain phase and applied data augmentation techniques to train a more robust model.

Ultimately, when comparing our proposed model with previous state-of-the-art models, we achieved nearly half the Mean Absolute Error on the test set, showing improvements across all images, whether containing many or few samples. These results demonstrate that our pipeline not only better encodes all information during density map creation but also effectively generalizes knowledge from the source domain to the Manatee Dataset. Consequently, we consider our model to be the new state-of-the-art for the manatee counting task.

VI. CONCLUSION

This paper introduces an innovative pipeline designed to enhance the performance of crowd-counting models, particularly in the task of counting manatees. The pipeline consists of three key steps: generating density maps using a novel 2-D kernel function, data augmentation for dataset balancing, and cross-domain techniques to improve accuracy. More density map predictions are available at the link on the GitHub platform https://github.com/Matteozara/Manatee_count.git together with the code to test and train the model.

While these steps have demonstrated promising results, particularly in accurately predicting crowded images, challenges persist, notably in accurately counting images with fewer and larger objects. Future efforts should focus on generating images with fewer animals to better train the model, possibly through fine-tuning Generative Adversarial Networks (GANs) or Diffusion Models.

Additionally, there is a need to reduce the model's dimensions to enhance its speed and efficiency, thus facilitating real-time applications. This optimization would make the model more accessible and user-friendly for scientists working in the field of manatee or other animal conservation.

REFERENCES

[1] *Fish and Wildlife Service to Consider Restoring Manatee's Endangered Status*. Accessed: Oct. 11, 2023. [Online]. Available: https://insideclimatenews.org/news/11102023/fish-and-wildlife-service-to-consider-restoring-manatees-endangered-status/

[2] *Will Florida Manatees Be Listed as an Endangered Species Again? Feds to Review Data*. Accessed: Oct. 11, 2023. [Online]. Available: https://www.tampabay.com/news/environment/2023/10/11/will-florida-manatees-be-listed-an-endangered-species-again-feds-review-data/

[3] S. Yoshinaga, A. Shimada, and R.-I. Taniguchi, "Real-time people counting using blob descriptor," *Proc.-Social Behav. Sci.*, vol. 2, no. 1, pp. 143–152, 2010, doi: 10.1016/j.sbspro.2010.01.028.

[4] J. Luo, J. Wang, H. Xu, and H. Lu, "Real-time people counting for indoor scenes," *Signal Process.*, vol. 124, pp. 27–35, Jul. 2016, doi: 10.1016/j.sigpro.2015.10.036.

[5] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, QLD, Australia, Oct. 2015, pp. 1299–1302, doi: 10.1145/2733373.2806337.

[6] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597, doi: 10.1109/CVPR.2016.70.

[7] F. Dai, H. Liu, Y. Ma, X. Zhang, and Q. Zhao, "Dense scale network for crowd counting," in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021.

[8] Q. Song, C. Wang, Y. Wang, Y. Tai, C. Wang, J. Li, J. Wu, and J. Ma, "To choose or to fuse? Scale selection for crowd counting," in *Proc. 35th AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 2576–2583.

[9] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "TransCrowd: Weakly-supervised crowd counting with transformers," *Sci. China Inf. Sci.*, vol. 65, no. 6, Apr. 2022, doi: 10.1007/s11432-021-3445-y.

[10] Z. Wang, Y. Pang, C. Ulus, and X. Zhu, "Counting manatee aggregations using deep neural networks and Anisotropic Gaussian kernel," *Sci. Rep.*, vol. 13, no. 1, p. 19793, Nov. 2023, doi: 10.1038/s41598-023-45507-3.

[11] Python Softw. Found., Wilmington, NC, USA. (2010). *Python Imaging Library (PIL) Documentation*. [Online]. Available: https://pillow.readthedocs.io/en/stable/

[12] B. Ferreira. *African Wildlife*. Kaggle. Accessed: May 11, 2024. [Online]. Available: https://www.kaggle.com/datasets/biancaferreira/african-wildlife

[13] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," 2018, *arXiv:1802.10062*.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[18] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 833–841, doi: 10.1109/CVPR.2015.7298684.

[19] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1215–1219, doi: 10.1109/ICIP.2016.7532551.

[20] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4031–4039.
[21] J. Gao, Q. Wang, and X. Li, "PCC Net: Perspective crowd counting via spatial convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3486–3498, Oct. 2020.
[22] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4593–4602.
[23] V. A. Sindagi and V. M. Patel, "Inverse attention guided deep crowd counting network," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.
[24] R. Rama Varior, B. Shuai, J. Tighe, and D. Modolo, "Multi-scale attention network for crowd counting," 2019, *arXiv:1901.06026*.
[25] A. Zhang, L. Yue, J. Shen, F. Zhu, X. Zhen, X. Cao, and L. Shao, "Attentional neural fields for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5713–5722, doi: 10.1109/ICCV.2019.00581.
[26] Y. Liu, Q. Wen, H. Chen, W. Liu, J. Qin, G. Han, and S. He, "Crowd counting via cross-stage refinement networks," *IEEE Trans. Image Process.*, vol. 29, pp. 6800–6812, 2020, doi: 10.1109/TIP.2020.2994410.
[27] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 1774–1783, doi: 10.1109/ICCV.2019.00186.
[28] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3618–3626, doi: 10.1109/CVPR.2018.00381.
[29] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6141–6150.
[30] H. Lin, Z. Ma, X. Hong, Q. Shangguan, and D. Meng, "GramFormer: Learning crowd counting via graph-modulated transformer," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 4, pp. 3395–3403.
[31] D. Liang, J. Xie, Z. Zou, X. Ye, W. Xu, and X. Bai, "CrowdCLIP: Unsupervised crowd counting via vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2893–2903.
[32] G. Ding, D. Yang, T. Wang, S. Wang, and Y. Zhang, "Crowd counting via unsupervised cross-domain feature adaptation," *IEEE Trans. Multimedia*, vol. 25, pp. 4665–4678, 2023, doi: 10.1109/TMM.2022.3180222.
[33] M. Zaramella, X. Zhu, I. Amerini, and P. Russo. (2024). *Manatee Counting Dataset*. [Online]. Available: https://drive.google.com/drive/folders/1rct4zK_7jNQDNe8XiIlTK4sSG9fto7Vo?usp=drive_link
[34] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL, USA: CRC Press, 2005.
[35] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
[36] K. Mader. (2019). *Malaria Bounding Boxes*. [Online]. Available: https://www.kaggle.com/datasets/kmader/malaria-bounding-boxes
[37] University of Texas at San Antonio. (2023). *Whales Dataset*. [Online]. Available: https://universe.roboflow.com/university-of-texas-at-san-antonio/whales-hksa5

MATTEO ZARAMELLA received the bachelor's degree in computer engineering from the University of Padova. He is currently pursuing the master's degree in artificial intelligence and robotics with the Sapienza University of Rome. He is currently working as a Junior Artificial Intelligence Researcher with Alcor Laboratory. In 2023, joined as a Research Exchange with Florida Atlantic University (FAU), FL, USA and also he did a Junior Research job at the Sapienza University of Rome, in 2023, and won the Best Paper Award at the SUMAC workshop, ACMMM 2023 for his paper "Why Don't You Speak?: A Smartphone Application to Engage Museum Visitors Through Deepfakes Creation."

XINGQUAN ZHU (Fellow, IEEE) received the Ph.D. degree in computer science from Fudan University. He is currently working as a Full Professor with the Department of Electrical Engineering and Computer Science, Florida Atlantic University (FAU). Since 2000, he has published more than 350 referred journal and conference papers in these areas, including four Best Paper Awards and three Best Student Paper Awards. His research interests include data mining, machine learning, and biomedical data analytics. He is a Steering Committee Member of the International Conference on Scientific and Statistical Database Management (SSDBM). He received the FAU Researcher of the Year Award, in 2024, the IEEE ICDM Outstanding Service Award, in 2023, the FAU College of Engineering and Computer Science Senior Faculty Research Award, in 2024 and 2019, and the National Engineers' Council Outstanding Engineering Achievement Merit Award, in 2019. He is a General Co-Chair of the 2024 IEEE International Conference on Knowledge Graph (ICKG), a Program Committee Co-Chair of the 2023 International Conference on Computational Data and Social Networks (CSoNet), a Program Committee Co-Chair of the 2022 IEEE International Conference on Data Mining (ICDM), and an Associate Editor of the ACM Transactions on Knowledge Discovery from Data, in 2017.

IRENE AMERINI (Member, IEEE) received the Ph.D. degree in computer engineering, multimedia, and telecommunication from the University of Florence, Italy, in 2010. She is currently working as an Associate Professor with the Department of Computer, Control, and Management Engineering A. Ruberti, Sapienza University of Rome, Italy. Her research interests include digital image processing, computer vision and multimedia forensics. She is a member of the IEEE Information Forensics and Security Technical Committee; the EURASIP TAC Biometrics, Data Forensics, and Security; and the IAPR TC6-Computational Forensics Committee.

Open Access funding provided by 'Università degli Studi di Roma ''La Sapienza'' 2' within the CRUI CARE Agreement