Localization in the Crowd with Topological Constraints

Shahira Abousamra[1], **Minh Hoai**[1], **Dimitris Samaras**[1], **Chao Chen**[2]
[1]Department of Computer Science, Stony Brook University, USA
[2]Department of Biomedical Informatics, Stony Brook University, USA
{sabousamra, minhhoai, samaras}@cs.stonybrook.edu, chao.chen.1@stonybrook.edu

Abstract

We address the problem of crowd localization, i.e., the prediction of dots corresponding to people in a crowded scene. Due to various challenges, a localization method is prone to spatial semantic errors, i.e., predicting multiple dots within a same person or collapsing multiple dots in a cluttered region. We propose a topological approach targeting these semantic errors. We introduce a topological constraint that teaches the model to reason about the spatial arrangement of dots. To enforce this constraint, we define a persistence loss based on the theory of persistent homology. The loss compares the topographic landscape of the likelihood map and the topology of the ground truth. Topological reasoning improves the quality of the localization algorithm especially near cluttered regions. On multiple public benchmarks, our method outperforms previous localization methods. Additionally, we demonstrate the potential of our method in improving the performance in the crowd counting task.

Introduction

Localization of people or objects, i.e., identifying the location of each instance, in a crowded scene is an important problem for many fields. Localization of people, animals, or biological cells provides detailed spatial information that can be crucial in journalism (McPhail and McCarthy 2004), ecology (Elphick 2008) or cancer research (Barua et al. 2018). A high quality localization algorithm naturally solves the popular crowd counting problem, i.e., counting the number of people in a crowded scene (Idrees et al. 2018). Furthermore, the rich spatial pattern can be used in many other tasks, e.g., initialization of tracking algorithms (Ren et al. 2018), animal population studies (Elphick 2008), tumor microenvironment analyses (Aukerman et al. 2020), and monitoring of social distancing (Yang et al. 2020).

Despite many proposed methods (Zhao, Nevatia, and Wu 2008; Ge and Collins 2009; Liu, Weng, and Mu 2019; Babu Sam et al. 2020), localization remains a challenging task. Aside from fundamental challenges of a crowded scene such as perspective, occlusion, and cluttering, one key issue is the limitation of annotation. Due to the large number of target instances, the ground truth annotation is usually provided in the form of dots located inside the instances (Fig. 1(a)). These dots only provide limited information. A dot can be arbitrarily perturbed as long as it is within the target instance, which can be of very different scales. As a consequence, the dot features are not specific. Without sufficient supervision, we cannot decide the boundary between instances. Thus, it is very hard to prevent spatial semantic errors, i.e., predicting multiple dots within a same person (false positives) or collapsing the dots of multiple persons in a cluttered area (false negatives).

In this paper, we propose a novel topological approach for the localization problem. We treat the problem as predicting a binary mask, called the Topological Map (Fig. 1(b)), whose connected components one-to-one correspond to the target dots. The number of components in the predicted mask should be the same as the number of ground truth dots. This spatial semantic constraint is indeed *topological*. During training we enforce such a "topological constraint" locally, i.e., the topology should be correct within each randomly sampled patch. The topological constraint teaches the model to reason about spatial arrangement of dots and avoids incorrect phantom dots and collapsing dots. This significantly improves the localization method quality, especially near dense regions. See Fig. 1(b), (c) and (d).

To enforce the topological constraint, we introduce a novel loss, called *persistence loss*, based on the theory of persistent homology (Edelsbrunner and Harer 2010). Instead of directly computing the topology of the predicted binary mask, we inspect the underlying likelihood map, i.e., the sigmoid layer output of the neural network. The persistent homology algorithm captures the topographic landscape features of the likelihood map, namely, modes and their saliency. Our persistence loss compares these modes and the true topology. Within a sample patch, if there are k true dots, the persistence loss promotes the saliency of the top k modes and penalizes the saliency of the remaining modes. This way it ensures that there are exactly k modes in the likelihood landscape, all of which are salient. A 0.5-thresholding of such a topologically correct likelihood map gives a binary mask with exactly k components, as desired.

We evaluate our method on various benchmarks and show that our proposed method, TopoCount, outperforms previous

(a) Input Image and Ground Truth Dot Map (b) TopoCount Result (c) Without Topological Constraint (d) Closeup Comparison on 3 Patches (e) Density Map

Figure 1: (a) a sample image and the ground truth (GT) dots; (b) the localization result of our method (TopoCount); (c) without the topological constraints, topological errors may happen and the localization quality is impacted; (d) closeup view of some specific patches. Without the topological constraints, the prediction often misses dots or collapses nearby dots (R1). It can also create phantom dots (R2 and R3). TopoCount successfully avoids such errors; (e) a density map from a SOTA counting method (Ma et al. 2019). The density map loses important topological information and cannot recover spatial arrangement of dots.

localization methods in various localization metrics. [1]

Application to crowd counting. We further demonstrate the power of our localization method by applying it to a closely related problem, *crowd counting*. For the counting problem, training images are also annotated with dots, but the task is simpler; one only needs to predict the total number of instances in each image. State-of-the-art (SOTA) counting algorithms, such as (Liu, Salzmann, and Fua 2019; Ma et al. 2019; Jiang et al. 2020; Wang et al. 2020a), learn to approximate a density map of the crowd whose integral gives the total count in the image. The learnt density function, even with accurate counting number, can significantly lose the topological characterization of the target density, especially near the dense population region. See Fig. 1(e).

To solve the counting problem, we can directly apply the localization algorithm and count the number of dots in the output map. However, this is not necessarily ideal for the task. Counting is an easier problem than localization. It has been shown that relaxing the output to a density function is the most effective strategy, although it will lose the exact locations of people or objects.

To achieve the best counting performance, we incorporate our localization result as complimentary information for density-based counting algorithms. By introducing our TopoCount results as additional input to SOTA counting algorithms (Ma et al. 2019; Liu, Salzmann, and Fua 2019), we improve their counting performance by 7 to 28 % on several public benchmarks. This further demonstrates the power of the spatial configuration information that we obtain through the topological reasoning.

In summary, our technical contribution is three-fold.

- We propose a topological constraint to address the topological errors in crowd localization.
- To enforce the constraint, we propose a novel persistence loss based on the theory of persistent homology. Our method achieves SOTA localization performance.
- We further integrate the topology-constrained localization algorithm into density-based counting algorithms to improve the performance of SOTA counting methods.

[1]The code and a full version of this paper can be found at https://github.com/TopoXLab/TopoCount.

Related Work

We discuss various localization approaches; some of which learn the localization algorithm jointly with the counting model. Babu Sam et al. (2020) learn to predict bounding boxes of human heads by fusing multi-scale features. The model is trained with cross entropy loss over the whole image and special focus on selected high error regions. Liu et al. (2018) performs detection using Faster RCNN (Ren et al. 2015). Faster RCNN has been shown to not scale well with the increasing occlusion and clutter in crowd counting benchmarks (Wang et al. 2020c). Liu, Weng, and Mu (2019) also learn to predict a localization map as a binary mask. They use a weighted cross-entropy loss to compensate for the unbalanced foreground/background pixel populations. In dense regions, the localization is further improved by recurrent zooming. However, all these methods are not explicitly modeling the topology of dots and thus cannot avoid topological errors (phantom dots and dots collapsing).

A related method is by (Laradji et al. 2018). It formulates the problem as a semantic segmentation problem. Blobs of the segmentation mask correspond to the target object instances. A blob is split if it contains multiple true dots and is suppressed if it does not contain any true dot. This method is not robust to the perturbation of dot locations; a blob that barely misses its corresponding true dot will be completely suppressed. On the contrary, our method leverages the deformation-invariance of topological structures arising from dots, and thus can handle the dot perturbation robustly.

(Ranjan, Le, and Hoai 2018; Cao et al. 2018; Li, Zhang, and Chen 2018; Liu, Salzmann, and Fua 2019; Ma et al. 2019; Jiang et al. 2020; Ranjan et al. 2020; Wang et al. 2020a). These methods train a neural network to generate a density function, the integral of which represents the estimated object count (Lempitsky and Zisserman 2010). The ground truth density functions are generated by Gaussian kernels centered at the dot locations. While these methods excel at counting, the smoothed density functions lose the detailed topological information of the original dots, especially in dense areas (see Fig. 1(e)). As a consequence, localization maps derived from the estimated density maps, e.g., via integer programming (Ma, Lei Yu, and Chan 2015) or via multi-scale representation of the density function (Idrees et al. 2018), are also of limited quality.

Topological information has been used in various learning and vision tasks. Examples include but are not limited to shape analysis (Reininghaus et al. 2015; Carriere, Cuturi, and Oudot 2017), graph learning (Hofer et al. 2017; Zhao and Wang 2019; Zhao et al. 2020), clustering (Ni et al. 2017; Chazal et al. 2013), learning with label noise (Wu et al. 2020) and image segmentation (Wu et al. 2017; Mosinska et al. 2018; Chan et al. 2017; Waggoner et al. 2015). Persistent-homology-based objective functions have been used for image segmentation (Hu et al. 2019; Clough et al. 2019), generative adversarial networks (GANs) (Wang et al. 2020b), graphics (Poulenard, Skraba, and Ovsjanikov 2018) and machine learning model regularization (Hofer et al. 2019; Chen et al. 2019). To the best of our knowledge, our method is the first to exploit topological information in crowd localization and counting tasks, and to use a topology-informed loss to solve the corresponding topological constraint problem.

Method: TopoCount

We formulate the localization problem as a structured prediction problem. Given training images labeled with *dot annotations*, i.e., sets of dots representing persons (Fig. 1(a)), we train our model to predict a binary mask. Each connected component in the mask represents one person. We take the centers of the connected components as the predicted dots. For training, we expand the dot annotations of training images into dot masks by a slight dilation of each dot, but with the condition that the expanded dots do not overlap. We call this "dot mask" the *ground truth dot map*.

To train a model to predict this binary ground truth dot map, we use a U-Net type architecture with a per-pixel loss. The output of the model after the Sigmoid activation is called the *topological likelihood map*. During inference, a final thresholding step is applied to the likelihood to generate the binary mask. We call the mask the *topological dot map* as it is required to have the same topology as the ground truth dot map. Fig. 2 shows our overall architecture.

The rest of this section is organized as follows. We first introduce the topological constraint for the topological dot map. Next, we formalize the persistence loss that is used to enforce the topological constraint. Afterwards, we provide details of the architecture and training. Finally, we discuss

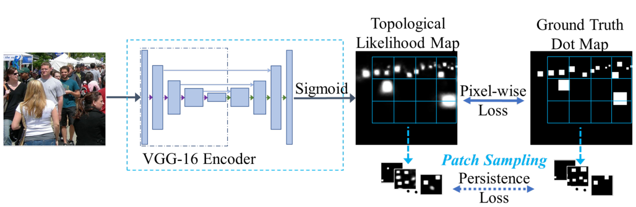

Figure 2: TopoCount has a U-Net style architecture with a VGG-16 encoder. During training a pixel-wise loss (DICE loss) is applied on the whole image and persistence loss is applied on sampled patches.

how to incorporate our method into SOTA counting algorithms to improve their counting performance.

Topological Constraint for Localization

For the localization problem, a major challenge is the perturbation of dot annotation. In the training dot annotation, a dot can be at an arbitrary location of a person and can correspond to different parts of a human body. Therefore it is hard to control the spatial arrangement of the predicted dots. As illustrated in Fig. 1(d), a model without special design can easily predict multiple "phantom dots" at different body parts of the same person. At cluttered regions, the model can exhibit "dots collapsing". To address these semantic errors, we must teach the model to learn the spatial contextual information and to reason about the interactions between dots. A model needs to know that nearby dots are mutually exclusive if there are no clear boundary between them. It should also encourage more dots at cluttered regions. To teach the model this spatial reasoning of dots, we define a *topological constraint* for the predicted topological dot map y:

Definition 1 (Topological constraint for localization) *Within any local patch of size $h \times w$, the Betti number of dimension zero; i.e, the number of connected components, of y equals to the number of ground truth dots.*

This constraint allows us to encode the spatial arrangement of dots effectively without being too specific about their locations. This way the model can avoid the topological errors such as phantom dots and dots collapsing, while being robust to perturbation of dot annotation. Next, we introduce a novel training loss to enforce this topological constraint.

Persistence Loss

Directly enforcing the topological constraint in training is challenging. The number of connected components and the number of dots within each patch are discrete values and their difference is non-differentiable. We introduce a novel differentiable loss called *persistence loss*, based on the persistence homology theory (Edelsbrunner, Letscher, and Zomorodian 2000; Edelsbrunner and Harer 2010). The key idea is that instead of inspecting the topology of the binary topological dot map, we use the continuous-valued likelihood map of the network f. We consider f as a terrain function and consider the landscape features of the terrain. These features provide important structural information. In particular, we focus on the modes (i.e., local maxima) of f. As illustrated in Fig. 3(b)(c), a salient mode of f, after thresholding, will become a connected component in the predicted topological dot map. A weak mode will miss the cutoff value and disappear in the dot map.

The persistence loss captures the saliency of modes and decides to enhance/suppress these modes depending on the ground truth topology. Given a patch with c many ground truth dots, our persistence loss enforces the likelihood f to only have c many salient modes, and thus c connected components in y. It reinforces the total saliency of the top c modes of f, and suppresses the saliency of the rest. The saliency of each mode, m, is measured by its *persistence*, $\mathrm{Pers}(m)$, which will be defined shortly. As an example, in

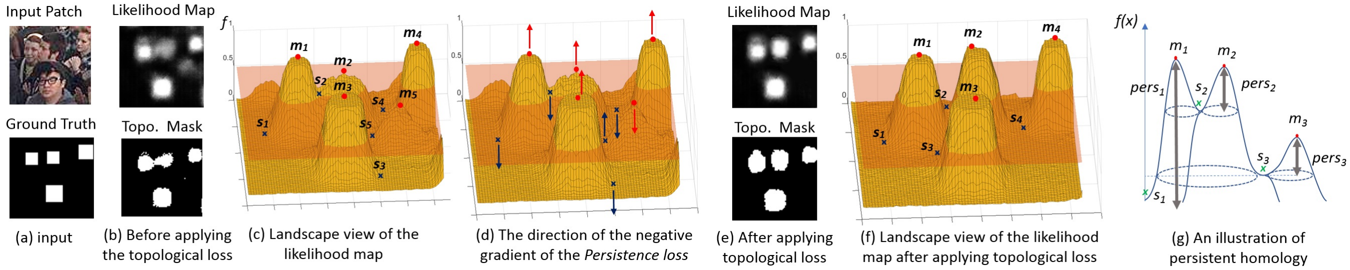

Figure 3: (a) An example image patch and its ground truth dot map (with 4 true dots). (b) The likelihood map and prediction mask without topological constraint. (c) A landscape view of the likelihood function. There are 5 modes (red dots) and their paired saddles(blue crosses). The top 4 salient ones (m_1 to m_4) are matched to the ground truth dots. The 5th, m_5, is not matched. The thresholding excludes the weak mode (m_5) in the predicted mask. But m_1 and m_2 are merged in the thresholded result because the saddle point between them (s_2) is above the cutoff value. (d) Optimizing the Persistence loss will suppress m_5 by reducing $f(m_5)$. Meanwhile, it will enhance the saliency of m_2 by increasing $f(m_2)$ and decreasing $f(s_2)$. When $f(s_2)$ is below the threshold, m_1 and m_2 are separated into two components in the final prediction. (e) The likelihood and the prediction mask with topological constraint (persistence loss). The collapsing of m_1 and m_2 is avoided. (f) A landscape view of the likelihood function in (e). Only 4 modes remain. (g) An illustration of persistent homology. Three modes are paired with saddles at which their attractive basins merge with others. The differences $f(m_i) - f(s_i)$, $i = 1, 2, 3$ are their persistence.

Fig. 3(c), f has 5 salient modes. If $c = 4$, the persistence loss will suppress the mode with the least persistence, in this case m_5, and will reinforce the other 4 modes. As a consequence, the mode m_2 is enhanced and is separated from m_1, avoiding a mode collapsing issue. Formally:

Definition 2 (Persistence Loss) *Given a patch, δ, with c ground truth dots, denote by $\mathcal{M}_c$ the top c salient modes, and $\overline{\mathcal{M}_c}$ the remaining modes of f. The persistence loss of f at the patch δ is*

$$\mathcal{L}_{Pers}(f, \delta) = -\sum\nolimits_{m \in \mathcal{M}_c} \mathrm{Pers}(m) + \sum\nolimits_{m \in \overline{\mathcal{M}_c}} \mathrm{Pers}(m) \tag{1}$$

Minimizing this loss is equivalent to maximizing the saliency of the top c modes and minimizing the saliency of the rest. Consequently, the function will only have c salient modes, corresponding to c components in the predicted mask, Fig. 3(e)(f). Next we formalize the mode saliency, called *persistence*, and derive the gradient of the loss.

Saliency/persistence of a mode. For a mode m (local maximum), its *basin of attraction* is the region of all points from which a gradient ascent will converge to m. Intuitively, the persistence of m, measuring its "relative height", is the difference between its height $f(m)$ and the level $f(s)$ at which its basin of attraction meets that of another higher mode. See Fig. 3(g) for an illustration.

In implementation, the saliency/persistence of each mode is computed by capturing its local maximum and corresponding saddle point. To find each mode m_i and its corresponding saddle point s_i where the component of m_i dies, we use a merging tree algorithm (Edelsbrunner and Harer 2010; Ni et al. 2017).

This algorithm is almost linear. The complexity is $O(n \log n + n\alpha(n))$, where n is the patch size. The $O(n \log n)$ term is due to the sorting of all pixels. $O(n\alpha(n))$ is the complexity for the union-find algorithm for merging connected components. $\alpha(n)$ is the inverse Ackermann's function, which is almost constant in practice. The algorithm will detect all critical points, i.e. modes and saddle points, at different thresholds and pair them properly corresponding to all topological features of the function/landscape.

Having obtained the critical points of the likelihood function using the above algorithm, we apply the persistence loss as follows: For each component c_i, denote by m_i its birth maximum and by s_i its death saddle critical points. The persistence of c_i is $\mathrm{Pers}(m_i) = f(m_i) - f(s_i)$. We sort all modes (or maximum-saddle pairs) according to their persistence. The persistence loss in Eq. (1) can be rewritten as

$$\begin{aligned} \mathcal{L}_{Pers}(f, \delta) = \quad & -\sum\nolimits_{m_i \in \mathcal{M}_c} (f(m_i) - f(s_i)) \\ & +\sum\nolimits_{m_i \in \overline{\mathcal{M}_c}} (f(m_i) - f(s_i)) \end{aligned} \tag{2}$$

When we take the negative gradient of the loss, for each of the top c modes, we will improve its saliency by increasing the function value at the maximum, $f(m_i)$, and decreasing the function value at its saddle $f(s_i)$. But for each other mode that we intend to suppress, the negative gradient will suppress the maximum's value and increase the saddle point's value. An important assumption in this setting is that the critical points, m_i and s_i, are constant when taking the gradient. This is true if we assume a discretized domain and a piecewise linear function f. For this discretized function, within a small neighborhood, the ordering of pixels in function value f remains constant. Therefore the algorithm output of the persistent computation will give the same set of mode-saddle pairs. This ensures that s_i and m_i's for all modes remain constant. The gradient of the loss w.r.t. the network weights, W, $\nabla_W \mathcal{L}_{Pers}(f, \delta) = -\sum_{m_i \in \mathcal{M}_c} \left(\frac{\partial f(m_i)}{\partial W} - \frac{\partial f(s_i)}{\partial W} \right) + \sum_{m_i \in \overline{\mathcal{M}_c}} \left(\frac{\partial f(m_i)}{\partial W} - \frac{\partial f(s_i)}{\partial W} \right)$.

TopoCount: Model Architecture and Training

TopoCount computes the topological map that has the same topology as the dot annotation. To enable the model to learn

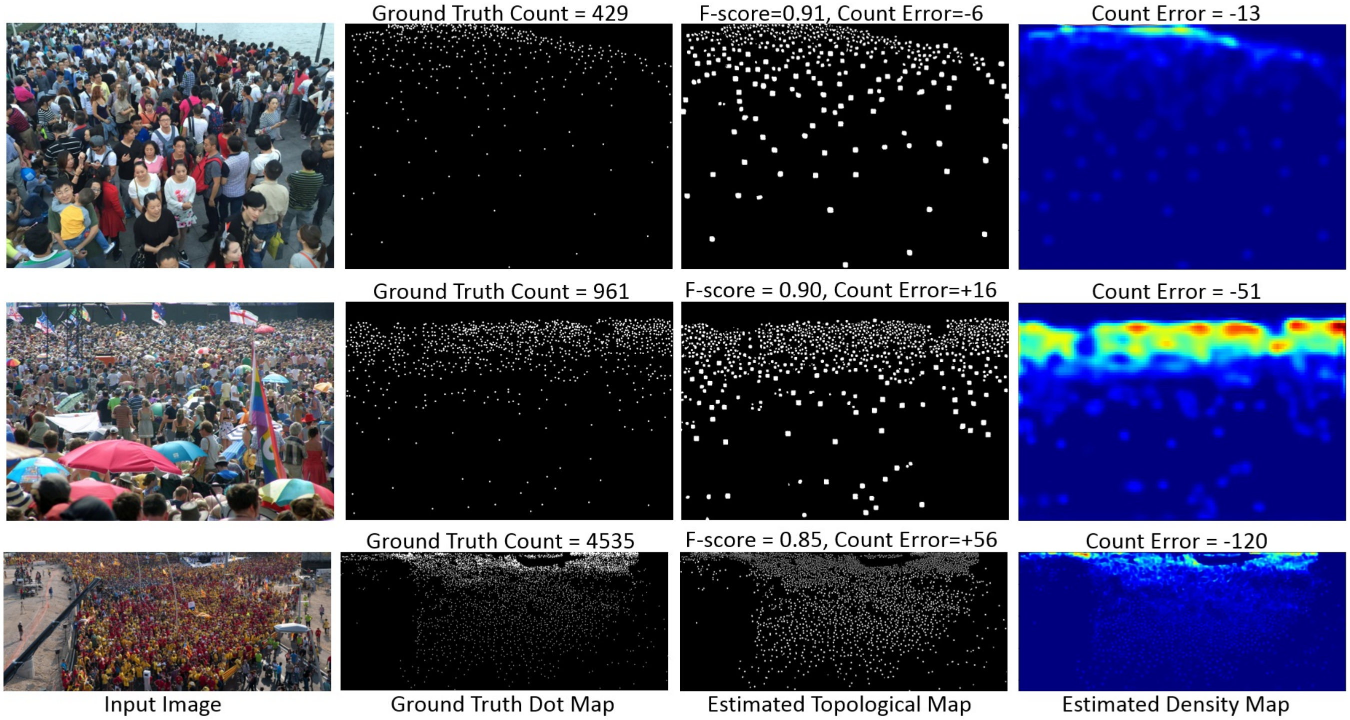

Figure 4: Sample results from different density crowd images. The columns represent the original image, ground truth, topological map by TopoCount, and the estimated density map by the integration of Bayesian + TopoCount.

to predict the dots quickly, we provide per-pixel supervision using DICE loss (Sudre et al. 2017). The DICE loss ($\mathcal{L}_{DICE}$) given Ground truth (G) and Estimation (E) is: $\mathcal{L}_{DICE}(G,E) = 1 - 2 \times \frac{(\sum G \circ E)+1}{(\sum G^2 + \sum E^2)+1}$, where $\circ$ is the Hadamard product.

More precisely, the model is trained with the loss:

$$\mathcal{L} = \mathcal{L}_{DICE} + \lambda_{pers}\mathcal{L}_{Pers} \tag{3}$$

in which λ_{pers} adjusts the weight of the persistence loss. An ablation study on the weight λ_{pers} is reported in the experiments. To provide more balanced samples for the per-pixel loss, we dilate the original dot annotation (treated as a supervision mask) slightly, but ensure that the dilation does not change its topology. The masks of two nearby dots stop dilating if they are about to overlap and impose false topological information. The size of the dilated dots is not related to the scale of the objects. These dilated dot masks from ground truth are used for training. Note that the persistence loss is applied to the likelihood map of the model, f.

Model Architecture Details. We use a UNet (Ronneberger, P.Fischer, and Brox 2015) style architecture with a VGG-16 encoder (Simonyan and Zisserman 2015). The VGG-16 backbone excludes the fully connected layers and has ≈ 15 million trainable parameters. There are skip connections between the corresponding encoding and decoding path blocks at all levels except for the first. The skip connection between the first encoder block and last decoder block is pruned to avoid overfitting on low level features, e.g., simple repeated patterns that often occur in crowd areas. The final output is the raw topological map, a Sigmoid activation is applied to generate the likelihood map. See Fig 2.

Integration with Counting Methods

In this section, we discuss how to apply our localization method to the task of crowd counting. In crowd counting, one is given the same dots annotation as the localization task. The goal is to learn to predict the total number of person or objects in an image. A straightforward idea is to use the localization map predicted by TopoCount and directly count the number of dots. Empirically, this solution is already on par with state-of-the-art methods (Table 5).

Here we present a second solution that is better suited for the counting task. We combine our localization algorithm with existing density-estimation methods (Ma et al. 2019; Liu, Salzmann, and Fua 2019) to obtain better counting performance. It has been shown that for the counting task, predicting a density map instead of the actual dots is the best strategy, especially in extremely dense or sparse regions. We argue that high quality localization maps provide additional spatial configuration information that can further improve the density-estimation counting algorithms.

To combine TopoCount with density-estimation counting methods, we use the raw output of a pre-trained TopoCount (the dot map and the topological likelihood map) as two additional channels concatenated to the RGB image. The five-channel 'images' are used as the input for a density-estimation model. The existing density-estimation model has to be adjusted to account for the change in the number of input channels. The density estimation model is usually initialized with weights from a pre-trained network on ImageNet. We keep the architecture and pre-trained weights of the density-estimation model the same everywhere except for the input layer. We modify the input layer so it accepts five channels and randomly initialize its weights. The density-estimation network is then trained end-to-end.

Our method is agnostic of the density-estimation model. As a proof-of-concept, we integrate TopoCount with two popular density-estimation counting methods: Bayesian (Ma et al. 2019) and CAN (Liu, Salzmann, and Fua 2019). We will show that the integration of the localization result learnt with topological constraint significantly boosts the performance of SOTA counting algorithms (7 to 28%, see Section). This further demonstrates the power of the spatial configuration we obtain through the topological reasoning.

The downside of the combined approach is that it only outputs density maps. The density maps, although better approximates the counts, cannot provide high quality localization information. This is the price one has to pay in order to achieve better counting performance.

Experiments

We validate our method on popular crowd counting benchmarks including ShanghaiTech parts A and B (Zhang et al. 2016), UCF CC 50 (Idrees et al. 2013), UCF QNRF (Idrees et al. 2018), JHU++ (Sindagi, Yasarla, and Patel 2020), and NWPU Challenge (Wang et al. 2020c).

For the localization task, our method is superior compared to other methods. Moreover, we show that the localization results of our method benefits the counting task.

Training Details. We train our *TopoCount* with the dilated ground truth dot mask. The dilation is by default up to 7 pixels. For JHU++ and NWPU, which are provided with head box annotation, we use a more accurate dilation guided by the box size, $\max$(7, box width/2, box height/2). In all cases the dilation is no more than half the distance to the nearest neighbor to avoid overlapping of nearby dots.

The window size of the patch for topological constraint controls the level of localization we would like to focus on. Since the scale of persons within an image is highly heterogeneous, varying the window size based on scale sounds intriguing. However the ground truth dot annotation generally do not carry scale information. As a result, we fix the patch size for each dataset. We use 50×50 pixels patches for ShanghaiTech and UCF CC 50, and 100×100 pixels patches for the larger scale datasets UCF QNRF, JHU++, and NWPU to account for the larger scale variation. An ablation study on the patch size selection is reported in the experiments. The persistence loss is applied on grid tiles to enforce topological consistency between corresponding prediction and ground truth tiles/patches. As data augmentation, coordinates of the top left corner of the grid are randomly perturbed. It should be noted that this tiling procedure is only performed during training with the persistence loss and is not performed during inference.

The model is trained with the combined loss $\mathcal{L}$ (Eq. (3)). During the first few epochs the likelihood map is random and is not topologically informative. In the beginning of training we use DICE loss only ($\lambda = 0$). When the model starts to converge to reasonable likelihood maps, we add the persistence loss with $\lambda = 1$. Fig. 4 shows qualitative results.

Localization Performance

We evaluate *TopoCount* on several datasets using (1) localized counting; (2) F1-score matching accuracy; and (3) the NWPU localization challenge metric.

Localized Counting. We evaluate the counting performance within small grid cells and aggregate the error. The Grid Average Mean absolute Errors (GAME) metric (Guerrero-Gómez-Olmedo et al. 2015), $G(L)$, divides the image into 4^L non-overlapping cells. In Table 1, the cell count in the localization-based methods LSC-CNN (Babu Sam et al. 2020) and TopoCount is the sum of predicted dots in the cell. On the other hand, the cell count in the density map estimation methods CSRNet (Li, Zhang, and Chen 2018) and Bayesian (Ma et al. 2019) is the integral of the density map over the cell area. TopoCount achieves the lowest error especially at the finest scale (level L=3), which indicates higher localization accuracy by the predicted dots.

Matching Accuracy. We evaluate matching accuracy in two ways. First, similar to Idrees et al. (2018) on the UCF QNRF dataset, we perform a greedy matching between detected locations and ground truth dots at thresholds varying from 1 to 100 and average the precision, recall, and F-scores over all thresholds. We compare with scores reported in (Idrees et al. 2018) in addition to calculated scores for SOTA localization methods (Babu Sam et al. 2020). Table 2 shows that our method achieves the highest F-score.

Second, similar to (Liu, Weng, and Mu 2019) on ShanghaiTech Part A and UCF QNRF datasets, we impose at each dot annotation an un-normalized Gaussian function parameterized by σ. A true positive is a predicted dot whose response to the Gaussian function is greater than a threshold t. We compare with (Liu, Weng, and Mu 2019) results at $\sigma = 5$ and $\sigma = 20$. Table 3 reports the mean average precision (mAP) and mean average recall (mAR) for $t \in [0.5, 0.95]$, with a step of 0.05. TopoCount achieves the highest scores with a large margin at both the small and large sigma σ.

NWPU-Crowd Online Localization Challenge. NWPU dataset provides dot annotation in addition to box coordinates with specified width w and height h surrounding each head. The online challenge evaluates the F-score with two adaptive matching distance thresholds: $\sigma_l = \sqrt{w^2 + h^2}/2$ and a more strict threshold $\sigma_s = min(w, h)/2$. Table 4 shows the F-score, precision, and recall with the 2 thresholds against the published challenge leaderboard. TopoCount achieves the highest F-score in both thresholds.

Counting Performance

Our localization method can be directly applied to crowd counting task. It performs competitively among SOTA counting methods. In Table 5, we compare TopoCount's overall count in terms of the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) against SOTA counting methods. Our method achieves SOTA performance on

	ShanghaiTech A			ShanghaiTech B			UCF QNRF		
Model	G(1)	G(2)	G(3)	G(1)	G(2)	G(3)	G(1)	G(2)	G(3)
CSRNet (Li, Zhang, and Chen 2018)	76	113	149	13	21	29	157	187	219
Bayesian (Ma et al. 2019)	75	90	130	**10**	**14**	23	**100**	**117**	150
LSC-CNN (Babu Sam et al. 2020)	70	95	137	**10**	17	27	126	160	206
TopoCount (proposed)	**69**	**81**	**104**	**10**	**14**	**20**	102	119	**148**

Table 1: Grid Average Mean absolute Errors (GAME)

Method	Prec.	Recall	F-score
MCNN (Zhang et al. 2016)	59.93%	63.50%	61.66%
CL-CNN D_∞ (Idrees et al. 2018)	75.8%	59.75%	66.82%
LSC-CNN (Babu Sam et al. 2020)	74.62%	73.50%	74.06%
TopoCount (proposed)	**81.77%**	**78.96%**	**80.34%**

Table 2: Localization accuracy on the UCF QNRF dataset, with metric in (Idrees et al. 2018)

	ShanghaiTech A		UCF QNRF	
	$\sigma = 20$	$\sigma = 5$	$\sigma = 20$	$\sigma = 5$
	mAP/mAR	mAP/mAR	mAP/mAR	mAP/mAR
RAZ_Loc	58.4/74.1	19.7/42.2	28.4/48.3	3.7/14.8
TopoCount	**85.0/82.8**	**56.0/54.8**	**69.0/66.5**	**27.1/26.2**

Table 3: Localization accuracy using metric in (Liu, Weng, and Mu 2019).

Method	F1-m / Pre / Rec (%)
Faster RCNN (Ren et al. 2015)	σ_l : 6.7 / **95.8** / 3.5 σ_s : 6.3 / **89.4** / 3.3
TinyFaces (Hu and Ramanan 2017)	σ_l : 56.7 / 52.9 / 61.1 σ_s : 52.6 / 49.1 / 56.6
VGG+GPR (Gao et al. 2019)	σ_l : 52.5 / 55.8 / 49.6 σ_s : 42.6 / 45.3 / 40.2
RAZ_Loc (Liu, Weng, and Mu 2019)	σ_l : 59.8 / 66.6 / 54.3 σ_s : 51.7 / 57.6 / 47.0
TopoCount (proposed)	σ_l : **69.1** / 69.5 / **68.7** σ_s : **60.1** / 60.5 / **59.8**

Table 4: NWPU-Crowd Localization Challenge Results. F1-m=F1-measure. Refer to (Wang et al. 2020c) for more details.

the new JHU++ large scale dataset and is mostly between second and third place for the other datasets compared to SOTA density-based methods.

Integration with Density-Based Methods. A high quality localization model can be combined with density-based methods to boost their performance. As described in Section , we integrate TopoCount with two SOTA density-based counting algorithms and report the results. Table 6 shows that the integration results in a significant improvement over the individual performance of the density map models. This further demonstrates the high quality of TopoCount localization and suggests that more sophisticated density map models can benefit from high quality localization maps to achieve an even better counting performance.

Ablation Studies

Ablation Study for the Loss Function. On the ShanghaiTech Part A dataset, we compare the performance of TopoCount trained with variations of the loss function in Eq. 3: (1) per-pixel weighted Binary Cross Entropy (BCE) loss as in (Liu, Weng, and Mu 2019) with empirically chosen weight of 5 to account for the amount of class imbalance in the ground truth dot maps, (2) per-pixel DICE loss only (i.e $\lambda = 0$ in Eq. 3), and (3) a combined per-pixel DICE loss and Persistence loss with $\lambda \in \{0.5, 1, 1.5, 2\}$. The results in Table 7 show the training with BCE loss gives the largest error. With $\lambda = 0$, i.e., DICE without the persistence loss, the error is lower. The error is further lowered with the introduction of the persistence loss. Varying λ between 0.5 and 2.0 the results are more robust and comparable, with the best performance at $\lambda = 1$ and $\lambda = 1.5$. Consequently, we use $\lambda = 1$ in all our experiments.

Ablation Study for Choosing Persistence Loss Patch Size The window size of the topological constraint patch controls the level of localization we would like to focus on. In the one extreme, when the patch is 1×1, the topological constraint becomes a per-pixel supervision. It helps the model to learn features for dot pixels, but loses the rich topological information within local neighborhoods. It is also not flexible/robust with perturbation. On the other extreme, when the patch is the whole image, the topological information is simply the total count of the image. This information is too high-level and will not help the model to learn efficiently; thus we have all other intermediate level supervisions, such as the density map. A properly chosen patch size will exploit rich spatial relationships within local neighborhoods while being robust to perturbation. In our experiments, we use a patch size of 50×50 pixels for ShanghaiTech and UCF CC 50 datasets, for datasets with larger variation in scale, namely UCF QNRF, JHU++, and NWPU-Crowd datasets, we use a larger patch size of 100×100 pixels. Next we explain how we choose the patch sizes.

To select the patch size for the persistence loss, we train four models on the ShanghaiTech Part A dataset with different patch sizes: 150×150, 100×100, 50×50, and 30×30. We evaluate the models localization accuracy using the GAME metric at scales L = 1 through 3, see Table 8. Training with patch size 30 or 150 yields poor performance. Using patch sizes 50 or 100 gives mostly similar results except at the smallest cell size (L=3) where patch size 50 is the winner, indicating better localization. We thus choose patch size of 50 for the ShanghaiTech and UCF CC 50 experiments.

The datasets UCF QNRF, JHU++, and NWPU-Crowd are

	Shanghai. A		Shanghai. B		UCF CC 50		UCF QNRF		JHU++		NWPU	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
IC-CNN (Ranjan, Le, and Hoai 2018)	68.5	116.2	10.7	16	260.9	365.5	-	-	-	-	-	-
CSRNet (Li, Zhang, and Chen 2018)	68.2	115	10.6	16	266.1	397.5	-	-	85.9	309.2	121.3	387.8
SANet (Cao et al. 2018)	67	104.5	8.4	13.6	258.4	334.9	-	-	91.1	320.4	190.6	491.4
ANF (Zhang et al. 2019)	63.9	99.4	8.3	13.2	250.2	340	110	174	-	-	-	-
RAZ-Net (Liu, Weng, and Mu 2019)	65.1	106.7	8.4	14.1	-	-	116	195	-	-	151.5	634.7
LSC-CNN (Babu Sam et al. 2020)	66.4	117.0	8.1	12.7	225.6	302.7	120.5	218.2	112.7	454.4	-	-
CAN (Liu, Salzmann, and Fua 2019)	62.3	100	7.8	12.2	212.2	**243.7**	107	183	100.1	314.0	106.3	**386.5**
Bayesian (Ma et al. 2019)	62.8	101.8	7.7	12.7	229.3	308.2	89	155	75.0	299.9	105.4	454.2
CG-DRC (Sindagi, Yasarla, and Patel 2020)	60.2	94.0	7.5	12.1	-	-	95.5	164.3	71.0	278.6	-	-
ASNet (Jiang et al. 2020)	**57.8**	**90.1**	-	-	**174.8**	251.6	91.6	159.7	-	-	-	-
DM-Count (Wang et al. 2020a)	59.7	95.7	**7.4**	**11.8**	211.0	291.5	**85.6**	**148.3**	-	-	**88.4**	388.6
TopoCount (proposed)	61.2	104.6	7.8	13.7	184.1	258.3	89	159	**60.9**	**267.4**	107.8	438.5

Table 5: Counting Performance Evaluation

	ShanghaiTech A		ShanghaiTech B		UCF CC 50		UCF QNRF		JHU++	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
CAN (Liu, Salzmann, and Fua 2019)	62.3	100	7.8	**12.2**	212.2	**243.7**	107	183	100.1	314.0
TopoCount + CAN (proposed)	**59.5**	**93.3**	**7.5**	13.2	**190**	249	**99**	**162**	**71.9**	**260.9**
	-4.5%	***-6.7%***	***-2.5%***	*+8.1%*	***-10.5%***	*+2.2%*	***-7.5%***	***-11.5%***	***-28.2%***	***-19.9%***
Bayesian (Ma et al. 2019)	62.8	101.8	7.7	12.7	229.3	308.2	89	155	75.0	299.9
TopoCount + Bayesian (proposed)	**58**	**96.3**	**7.2**	**11.8**	**191**	**257**	**85**	**148**	**61.8**	**262.0**
	-7.6%	***-5.4%***	***-6.5%***	***-7.1%***	***-16.7%***	***-16.6%***	***-4.5%***	***-4.5%***	***-17.6%***	***-12.7%***

Table 6: Integration of TopoCount with density estimation methods.

	BCE	DICE	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
G(3)	122	114	109	**104**	**104**	107

Table 7: Ablation study on the loss function. Compare TopoCount localization score G(3) on the ShanghaiTech Part A dataset when trained with different loss functions: weighted BCE loss, DICE loss ($\lambda = 0$), and DICE loss + λ Persistence loss ($\lambda = 0.5, 1, 1.5, 2$).

different from the other datasets in their wide variation in scale and resolution. We suspect that a patch size of 50 may not be suitable. We experiment with a small subset of randomly selected (N=50) images from the UCF QNRF training data. Again, we train 4 models with different patch sizes: 150, 100, 50, and 30, and evaluate the models localization using GAME. Because the images in this dataset have a higher resolution range, we use L = 1 through 4, see Table 8. We find that a patch size of 150 is more suitable at the coarser cells (L=1, 2) while a patch size of 50 is more suitable at the finer cells (L=3, 4). For training on these datasets, we choose the intermediate patch size of 100.

		G(L)			
Dataset	Patch Size	G(1)	G(2)	G(3)	G(4)
Shanghai. A	150	75.4	89.9	114.2	-
	100	**68.4**	82.0	107.7	-
	50	69.3	**81.6**	**104.9**	-
	30	75.4	86.4	108.2	-
UCF-QNRF	150	**153.5**	**175.1**	208.5	272.3
	100	155.9	177.1	**206.6**	264.4
	50	160.6	179.4	206.8	**260.3**
	30	179.9	195.2	222.1	273.1

Table 8: Comparison of patch size for persistence loss on ShanghaiTech Part A and UCF-QNRF (N=50) using localization score GAME(L)

Conclusion

This paper proposes a novel method for localization in the crowd. We propose a topological constraint and a novel persistence loss based on persistent homology theory. The proposed topological constraint is flexible and suitable for both sparse and dense regions. The proposed method achieves state-of-the-art localization accuracy. The high quality of our results is further demonstrated by the significant boost of the performance of density-based counting algorithms when using our results as additional input. Our method closes the gap between the performance of localization and density map estimation methods; thus paving the way for advanced spatial analysis of crowded scenes in the future.

Acknowledgements

This research was partially supported by US National Science Foundation Awards IIS-1763981, CCF-1855760, IIS-1909038, the SUNY2020 Infrastructure Transportation Security Center, Air Force Research Laboratory (AFRL) DARPA FA8750-19-2-1003, the Partner University Fund, and a gift from Adobe. We thank anonymous reviewers for constructive comments and suggestions.

References

Aukerman, A.; Carrière, M.; Chen, C.; Gardner, K.; Rabadán, R.; and Vanguri, R. 2020. Persistent Homology Based Characterization of the Breast Cancer Immune Microenvironment: A Feasibility Study. In *Proceedings of the International Symposium on Computational Geometry*.

Babu Sam, D.; Peri, S. V.; Narayanan Sundararaman, M.; Kamath, A.; and Radhakrishnan, V. B. 2020. Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Barua, S.; Fang, P.; Sharma, A.; Fujimoto, J.; Wistuba, I.; Rao, A. U. K.; and Lin, S. H. 2018. Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer. *Lung Cancer* 117: 73–79.

Cao, X.; Wang, Z.; Zhao, Y.; and Su, F. 2018. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *Proceedings of the European Conference on Computer Vision*.

Carriere, M.; Cuturi, M.; and Oudot, S. 2017. Sliced wasserstein kernel for persistence diagrams. In *Proceedings of the International Conference on Machine Learning*.

Chan, H. L.; Yan, S.; Lui, L. M.; and Tai, X.-C. 2017. Topology-Preserving Image Segmentation by Beltrami Representation of Shapes. *Journal of Mathematical Imaging and Vision* 60: 401–421.

Chazal, F.; Guibas, L. J.; Oudot, S. Y.; and Skraba, P. 2013. Persistence-based clustering in riemannian manifolds. *Journal of the ACM* 60(6): 41.

Chen, C.; Ni, X.; Bai, Q.; and Wang, Y. 2019. A Topological Regularizer for Classifiers via Persistent Homology. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Clough, J.; Oksuz, I.; Byrne, N.; Schnabel, J.; and King, A. 2019. Explicit Topological Priors for Deep-Learning Based Image Segmentation Using Persistent Homology. In *Proceedings of the International Conference on Information Processing in Medical Imaging*.

Edelsbrunner, H.; and Harer, J. 2010. *Computational Topology - an Introduction.* American Mathematical Society.

Edelsbrunner, H.; Letscher, D.; and Zomorodian, A. 2000. Topological persistence and simplification. In *Proceedings of the Annual Symposium on Foundations of Computer Science*.

Elphick, C. S. 2008. How you count counts: the importance of methods research in applied ecology. *Journal of Applied Ecology* 45(5): 1313–1320.

Gao, J.; Han, T.; Wang, Q.; and Yuan, Y. 2019. Domain-adaptive Crowd Counting via Inter-domain Features Segregation and Gaussian-prior Reconstruction. *CoRR* abs/1912.03677.

Ge, W.; and Collins, R. T. 2009. Marked point processes for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Guerrero-Gómez-Olmedo, R.; Torre-Jiménez, B.; López-Sastre, R.; Maldonado-Bascón, S.; and Oñoro-Rubio, D. 2015. Extremely Overlapping Vehicle Counting. In *Proceedings of Pattern Recognition and Image Analysis, IbPria*.

Hofer, C.; Kwitt, R.; Niethammer, M.; and Dixit, M. 2019. Connectivity-Optimized Representation Learning via Persistent Homology. In *Proceedings of the International Conference on Machine Learning*.

Hofer, C.; Kwitt, R.; Niethammer, M.; and Uhl, A. 2017. Deep learning with topological signatures. In *Proceedings of Advances in Neural Information Processing Systems*.

Hu, P.; and Ramanan, D. 2017. Finding Tiny Faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hu, X.; Li, F.; Samaras, D.; and Chen, C. 2019. Topology-Preserving Deep Image Segmentation. In *Advances in Neural Information Processing Systems*.

Idrees, H.; Saleemi, I.; Seibert, C.; and Shah, M. 2013. Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Máadeed, S.; Rajpoot, N. M.; and Shah, M. 2018. Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. In *Proceedings of the European Conference on Computer Vision*.

Jiang, X.; Zhang, L.; Xu, M.; Zhang, T.; Lv, P.; Zhou, B.; Yang, X.; and Pang, Y. 2020. Attention Scaling for Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Laradji, I. H.; Rostamzadeh, N.; Pinheiro, P. O.; Vázquez, D.; and Schmidt, M. 2018. Where Are the Blobs: Counting by Localization with Point Supervision. In *Proceedings of the European Conference on Computer Vision*.

Lempitsky, V.; and Zisserman, A. 2010. Learning To Count Objects in Images. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Li, Y.; Zhang, X.; and Chen, D. 2018. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, C.; Weng, X.; and Mu, Y. 2019. Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, J.; Gao, C.; Meng, D.; and Hauptmann, A. G. 2018. DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, W.; Salzmann, M.; and Fua, P. 2019. Context-Aware Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ma, Z.; Lei Yu; and Chan, A. B. 2015. Small instance detection by integer programming on object density maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ma, Z.; Wei, X.; Hong, X.; and Gong, Y. 2019. Bayesian Loss for Crowd Count Estimation With Point Supervision. In *Proceedings of the IEEE International Conference on Computer Vision*.

McPhail, C.; and McCarthy, J. 2004. Who Counts and How: Estimating the Size of Protests. *Contexts* 3(3): 12–18.

Mosinska, A.; Márquez-Neila, P.; Kozinski, M.; and Fua, P. 2018. Beyond the Pixel-Wise Loss for Topology-Aware Delineation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ni, X.; Quadrianto, N.; Wang, Y.; and Chen, C. 2017. Composing tree graphical models with persistent homology features for clustering mixed-type data. In *Proceedings of the International Conference on Machine Learning*.

Poulenard, A.; Skraba, P.; and Ovsjanikov, M. 2018. Topological Function Optimization for Continuous Shape Matching. *Computer Graphics Forum* 37(5): 13–25.

Ranjan, V.; Le, H.; and Hoai, M. 2018. Iterative Crowd Counting. In *Proceedings of the European Conference on Computer Vision*.

Ranjan, V.; Wang, B.; Shah, M.; and Hoai, M. 2020. Uncertainty Estimation and Sample Selection for Crowd Counting. In *Proceedings of the Asian Conference on Computer Vision*.

Reininghaus, J.; Huber, S.; Bauer, U.; and Kwitt, R. 2015. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*.

Ren, W.; Kang, D.; Tang, Y.; and Chan, A. B. 2018. Fusing Crowd Density Maps and Visual Object Trackers for People Tracking in Crowd Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ronneberger, O.; P.Fischer; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Sindagi, V. A.; Yasarla, R.; and Patel, V. M. 2020. JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method. *Technical Report* .

Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; and Jorge Cardoso, M. 2017. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Proceedings of Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.

Waggoner, J.; Zhou, Y.; Simmons, J.; Graef, M. D.; and Wang, S. 2015. Topology-Preserving Multi-label Image Segmentation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

Wang, B.; Liu, H.; Samaras, D.; and Hoai, M. 2020a. Distribution Matching for Crowd Counting. In *Proceedings of Advances in Neural Information Processing Systems*.

Wang, F.; Liu, H.; Samaras, D.; and Chen, C. 2020b. Topogan: A topology-aware generative adversarial network. In *European Conference on Computer Vision*, volume 2.

Wang, Q.; Gao, J.; Lin, W.; and Li, X. 2020c. NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Wu, P.; Chen, C.; Wang, Y.; Zhang, S.; Yuan, C.; Qian, Z.; Metaxas, D.; and Axel, L. 2017. Optimal topological cycles and their application in cardiac trabeculae restoration. In *International Conference on Information Processing in Medical Imaging, IPMI*, 80–92.

Wu, P.; Zheng, S.; Goswami, M.; Metaxas, D.; and Chen, C. 2020. A Topological Filter for Learning with Label Noise. *Advances in Neural Information Processing Systems* 33.

Yang, D.; Yurtsever, E.; Renganathan, V.; Redmill, K. A.; and Özgüner, Ü. 2020. A Vision-based Social Distancing and Critical Density Detection System for COVID-19. *CoRR* abs/2007.03578.

Zhang, A.; Yue, L.; Shen, J.; Zhu, F.; Zhen, X.; Cao, X.; and Shao, L. 2019. Attentional Neural Fields for Crowd Counting. In *Proceedings of the IEEE International Conference on Computer Vision*.

Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; and Ma, Y. 2016. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhao, Q.; and Wang, Y. 2019. Learning metrics for persistence-based summaries and applications for graph classification. In *Proceedings of Advances in Neural Information Processing Systems*.

Zhao, Q.; Ye, Z.; Chen, C.; and Wang, Y. 2020. Persistence enhanced graph neural network. In *International Conference on Artificial Intelligence and Statistics*, 2896–2906. PMLR.

Zhao, T.; Nevatia, R.; and Wu, B. 2008. Segmentation and Tracking of Multiple Humans in Crowded Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .