Learning-to-Count by Learning-to-Rank

Adriano C. D'Alessandro, Ali Mahdavi-Amiri, Ghassan Hamarneh
School of Computing Science, Simon Fraser University, Burnaby, Canada
{*acdaless,amahdavi,hamarneh*}*@sfu.ca*

***Abstract*—Object counting methods rely on density maps, which are heatmaps produced by placing Gaussian density over object locations. However, density maps are expensive to collect. To reduce the annotation burden, we propose a form of weak supervision that only requires object-based pairwise image rankings. These annotations can be collected rapidly with a single click per image pair and supply a weak signal for object quantity. However, a model learn to fit spurious patterns that satisfy the ranking constraint but do not rely on the objects. To encourage the network to solve the ranking constraints by localizing objects, we propose adversarial density map estimation. This method regularizes a ranking network's intermediate feature representation such that it corresponds to a plausible density map. We demonstrate the effectiveness of our method on several benchmark object counting datasets, and show results with a performance that approaches that of fully-supervised methods using data that can be collected with a fraction of the annotation burden.**

***Keywords*-Weak Supervision, Object Counting, Ranking**

The object counting problem involves enumerating the number of objects within an image, which has broad applicability across several domains such as wildlife population monitoring [1], crowd analysis [2]–[7], and traffic analysis [8]. Different methods have approached the problem using annotations strategies such as bounding boxes [9], [10], global object counts [11], inter-image ranking [12], and density maps [13]. Density maps tend to provide the best performance on the object counting task [13]–[15]. Density maps provide object counts via localization, but they localize objects with Gaussian blobs rather than bounding boxes.

In addition to performance considerations, each of these annotation strategies carry their own per-annotation cost or labor burden. Figure 2 highlights the relative annotation burden for the most relevant annotation strategies. By including this additional consideration, we can compare annotation strategies by their compromise between test time performance and training time annotation burden. Despite their respective performance, both density map and global count annotations tend to carry a high annotation burden [15]. This provides motivation to seek an alternative annotation type with a favorable compromise between burden and performance. We circumvent the burden of previous annotation types by introducing pairwise inter-image ranking, a simple form of annotation that can be rapidly collected.

Pairwise inter-image ranking is a novel binary valued annotation that orders two different images based on their per-image object counts. These annotations provide a weak signal for object quantity by creating an object-based partial ordering of the available images. Previous work by Liu et al [12] on *intra*-image ranking has demonstrated that image ranking can be an effective training signal for semi-supervised counting problems. This previous work automatically collected *intra*-image ranking annotations by leveraging the fact that the object count of a sub-image crop cannot exceed that of the whole image. Liu et al. used this relationships as an additional unsupervised training signal to improve the results of methods trained using fully-supervised density map. However, while free to collect, Liu et al. demonstrated that a model trained using only *intra*-image ranking annotations performs significantly worse. By comparison, we demonstrate that our proposed *inter*-image ranking annotations perform well on their own, without any additional supervisory signal.

The major technical challenge in exploiting our inter-image ranking annotation formulation is finding a way to extract counts and relevant locations given only rankings. To this end, we propose an adversarial strategy for regularizing the penultimate representation of a ranking network to have the properties of a density map by comparing it to a pseudo-density map distribution. By enforcing that the model must solve the ranking problem using plausible density maps, we argue that this strategy forces the model to solve the ranking sub-task by detecting re-occurring objects that can satisfy all of the ranking constraints. Combined with the weak quantity signal provided by inter-image pairwise ranking annotations, this strategy produces a counting and localization model with acceptable performance as compared with architectures trained using density maps and global counts annotations. In summary, we make the following contributions: (1) We propose object-based pairwise inter-image ranking as a novel, low-cost annotation strategy for weakly supervised counting, and demonstrate that it performs comparably with fully supervised counting methods. (2) We propose adversarial density map regularization, a novel method for enforcing that the network learns an internal representation conforming to the properties of a density map.

I. Previous Work

Object Counting with Limited Data.: Object counting methods perform best when learning from density maps [11], [13], which carry a high annotation burden. Several methods have attempted to eliminate this burden. These methods

can be split into 4 categories – semi-supervised, knowledge transfer, sample selection, and weakly supervised methods.

Semi-supervised counting methods alleviate the annotation burden by including additional unlabelled data; inter-image ranking introduced an unsupervised ranking loss that exploits the fact that any image has as many or more objects than a cropped portion of that image [12], [16]. Other methods [17] have proposed feature learning strategies. Domain transfer methods transfer features between counting problems; one method [7] proposed learning from a multi-modal dataset containing both density maps and global object counts. A recent approach [18] proposed a few-shot learning strategy using exemplar and density map pairs, which could be extended to novel object classes. Active learning methods [19], [20] approach the problem by finding ways to only label important examples. Weakly supervised methods attempt to better utilize global counts as annotations. Recent methods [21], [22] proposed various regularization terms. One such method introduced a soft-sorting loss [21], which involved learning from global object count annotations directly and indirectly. Our method differs from the above weakly-supervised methods by learning exclusively from a weaker signal than global object counts.

Ranking as Supervision.: The use of ranking as a training signal originates within information retrieval research. RankNet [23], a document retrieval network, emerged as the first deep learning approach to ranking. However, this approach has been extended into several computer vision applications. Facial age estimation [24] has benefited from pairwise image rankings to learn the 'amount' of age in an image. Pairwise image ranking has also been used to localize facial attributes [25]. These applications highlight that ranking plays a significant role in computer vision.

II. Pairwise Ranking

A. The Burden of Pairwise Ranking Annotations

Pairwise image ranking is the process of providing a binary ordering annotation r_{ij} for an image pair (x_i, x_j) based on the object counts (c_i, c_j) present in the two images, where $r_{ij} \equiv c_i \geq c_j$. A pairwise ranking training dataset for N pairs of images is given as $\mathcal{D}_{rank} = \{(x_i, x_j), r_{ij}\}^N$. We also define an important relationship between c_i and c_j, which is the ratio between object counts, $\gamma = \min(c_i/c_j, c_j/c_i)$, where $0 \leq \gamma \leq 1$.

Researchers in human psychology have found that humans can rapidly assess which of two groups of objects has the most objects if γ is below a threshold [26], [27]. This is known as Weber-Fechner law, which describes the change in a stimulus necessary for a human to perceive the difference relative to the existing stimulus [28]. For the task of pairwise image ranking, researchers [26], [27] have also found that within 0.75 seconds, untrained adults are capable of determining which of two images has more objects, if γ between object count in the two images is smaller than approximately 9:10 and 10:11 (independent of the absolute count), i.e, if $\gamma < \gamma^*$, where γ^* is the Weber-Fechner ratio with value around 0.9-0.91. We we adopt these established values as an approximation of the upper bound on the Weber-Fechner ratio, as the true Weber-Fechner ratio is likely related to object scale, density, complexity, etc.

Given an approximate upper bound for the Weber-Fechner ratio, we must now decide on how to handle the case where a similar number of objects appear in both images, i.e., the image pair violates $\gamma < \gamma^*$, and the annotation cost is expected to increase. One option is to ask an annotator to count up the objects in both images; however, this is a costly option. Another option is to permit annotator noise by requesting that an annotator guess within a certain time constraint. In the case of random guessing, we will have a 50% chance of acquiring a correct rank annotation, i.e., approximately half of the labels will be correct and the other half will only be incorrect by a small number of objects.

Figure 1 provides an analysis of the distribution of object count ratios γ for pairs of images sampled from popular counting benchmark datasets, which are described further in section IV-A. To produce the distribution of ratios given above for a counting dataset $\{x_i, c_i\}^{N_c}$ with N_c examples, we sample all $\binom{N_c}{2}$ pairs and calculate their ratio γ. The percentage of image pairs that violate $\gamma < \gamma^*$ for the benchmarks is 7.4-22.8%. So, when Weber-Fechner ranking label noise is permitted, we expect that half of those, on average, would be incorrectly annotated, i.e. 3.7-11.4%.

It is well known that label noise is a persistent feature of many popular datasets [29], with the ImageNet test set having an error rate of 5.83% and the CIFAR-100 test-set having an error rate of 5.85%. Given that many popular benchmarks are noisy, we demonstrate in section IV-D that this small amount of ranking annotation noise is tolerable. We are now left to answer the difficult question of how to extract the counts and estimate the locations of objects when presented with a weak rank signal that can be rapidly collected albeit with a small amount of label noise.

B. Beyond Intra-Image Ranking

Previous work on semi-supervised intra-image ranking [12] has demonstrated the value of ranking as a training signal for counting problems. Given that these annotations can be collected rapidly and for free, one may reasonably wonder why we would seek to manually collect ranking annotations. The original intra-image ranking method included two loss terms, a fully-supervised density map regression term and an intra-image ranking term. This is important, as the method explicitly requires (costly) annotations that ground the object counts to the object location.

When paired with fully-supervised density map regression, the ranking loss provides additional useful weak quantity information. However, when used in isolation, the intra-image ranking loss no longer receives information related

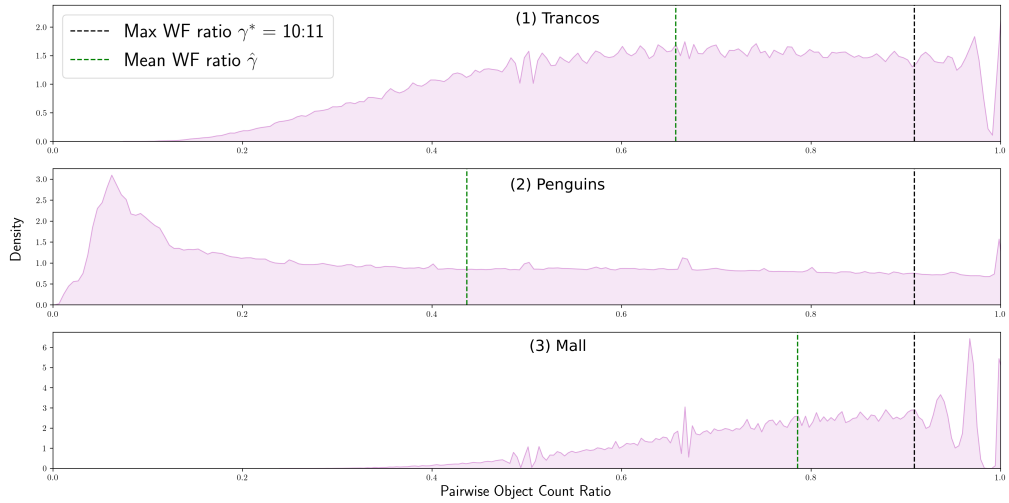

Figure 1. Density plots showing the distribution of object count ratios γ for objects in pairs of images sampled from each respective benchmark dataset: (a-c) Trancos, Penguins, and Mall. Mean ratio $\bar{\gamma}$ and Weber-Fechner ratio γ^* are labelled on each plot, providing a quick summary of the annotation difficulty for each dataset. The percentage of pairs on the right of γ^* is, 12.6% for Trancos, 7.4% for Penguins, and 22.8% for Mall respectively.

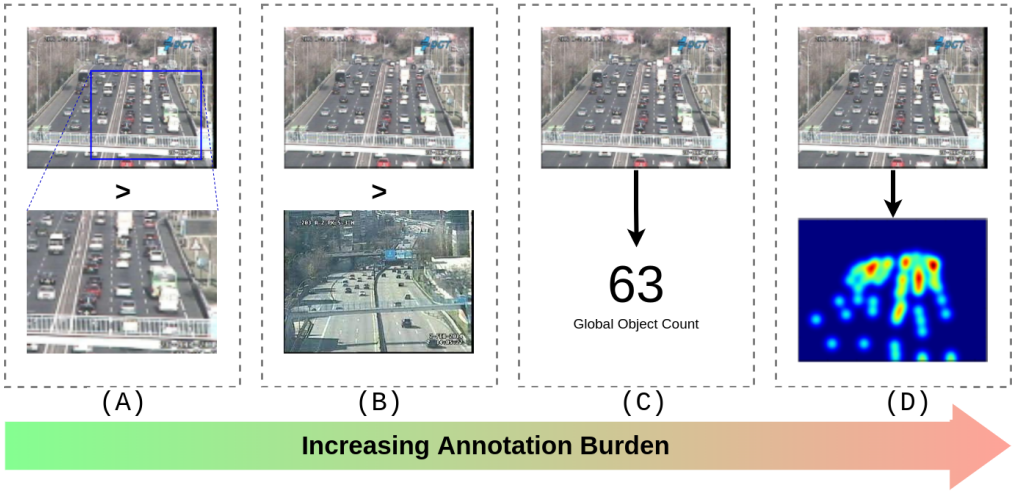

Figure 2. Different types of annotations and their relative burden for training object counting models. **(A)** Intra-image ranking annotations compare object counts in an image and a sub-crop of that image. **(B)** *Ours*. Inter-image ranking annotations are manually collected and compare object counts between a diverse set of image pairs. **(C)** Global object count annotations require a single number per image representing the object count. **(D)** Density map annotations require a single dot per object location, and which is converted into a Gaussian blob.

to object identity. Given that the sub-crop contains a subset of the image-level features contained in the whole image, there are many trivial solutions that can satisfy the intra-image ranking constraints. Figure 2 highlights the difference between inter- and intra-image ranking pairs. Visual inspection of intra-image ranking pairs reveals plausible spurious solutions that satisfy the intra-image ranking constraint. For example, image boundary artifacts such as the trees in figure 2A, are always less likely to appear in the sub-crop and the model may overfit to such trivial features. The authors of [12] empirically verified this observation. They pre-trained a model using only the intra-image ranking examples and then fine-tuned the model on fully-supervised counting examples. They reported that the model pre-trained using only the ranking signal saw a significant increase in error when compared to both their proposed semi-supervised setup and a model pre-trained ImageNet features. Thus, if we wish to learn explicitly from ranking pairs, we argue that a move beyond intra-image ranking is necessary. We solve this problem by introducing inter-image ranking pairs.

III. Method

The goal of our method is to develop a model which can extract object counts given only pairwise image rankings.

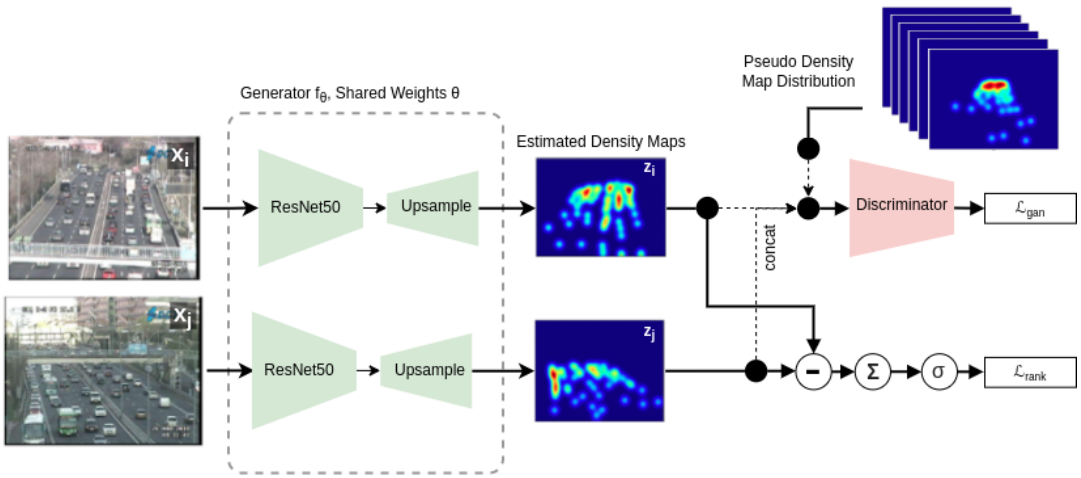

Figure 3. Method overview. Pairs of images with count based ranking labels are used to train a neural network. Each image x_i and x_j are passed to the generator model f_θ and their respective outputs are used to calculate $\mathcal{L}_{rank}$, which provides the weak object quantity training signal. In addition to this, we include an adversarial density map generation loss, $\mathcal{L}_{gan}$, which encourages the output of the generator model to have the properties of a density map.

These annotations contain a weak signal for object quantity, which we use to train a neural network, $f(x_i;\theta)$, outlined in Figure 3. However, this target alone is not enough to learn a representation from which we can extract object counts. First, this target does not require that a model learn features which correspond to whole object counts. Second, there are potentially trivial solutions to the ranking constraints that the model can exploit. We solve this problem by proposing adversarial density map estimation, a strategy which structures the intermediate representation of f_θ to have the properties of a density map. By solving the image ranking constraint using plausible density map proposals, our model learns to not only count objects but also localize those objects within an image.

A. Ranking Network

The purpose of the network $f(x_i;\theta)$ is to extract the underlying weak object quantity signal from the pairwise image ranking annotations $\mathcal{D}_{rank}$. As outlined in Figure 3, our base counting model f_θ receives two images as input, x_i and x_j, and outputs representations $z_i, z_j \in [0,1]^{w\times h}$. We calculate the object counts by taking the integral of each representation, such that $c_i = \sum_{k,l}(z_i)_{kl}$. We then follow the strategy proposed in [23] and model the probability that $c_i \geq c_j$ by approximating the true distribution as follows:

$$p_{rank} = P(r_{ij}|x_i, x_j;\theta) = \sigma(\sum_{k,l}(z_i)_{kl} - (z_j)_{kl}), \quad (1)$$

where σ is the sigmoid operation and (k,l) are the indices for the representations z_i and z_j. Here, we benefit from the fact that when the difference between z_i and z_j is positive, the sigmoid operation outputs a value greater than 0.50. Whereas when the difference is negative, the sigmoid operation outputs a value less than 0.50. This allows us to model which of two images in a pair has more objects by inspecting the magnitudes of the sum over z_i and z_j. Thus, by optimizing θ using the following loss function:

$$\mathcal{L}_{\texttt{rank}} = -E_{p_{data}}[\log(p_{rank})], \quad (2)$$

the model must learn to minimize the number of pairwise inversions (from the ground truth distribution p_{data}) among all ranking examples in the training dataset, which creates a partial ordering of all the images by object count. We also create a version p_{data} corrupted by simulated annotator noise by randomly selecting half of the ground truth labels r_{ij} which violate γ^* and flipping their labels. In the next section, we propose an approach to explicitly connect the output representation to object locations.

B. Adversarial Density Map Generation

Previous empirical results have demonstrated the value of density maps as a location-based annotation for counting problems. Density maps are structured such that they place Gaussian density where objects occur and integrate to the global count. These properties are useful because they explicitly connect the detection and counting task. While we do not have access to density maps, we argue that optimizing the counting network f_θ to propose density maps as intermediate representations while solving the pairwise image ranking problem captures some of the useful properties of density maps. We explore a strategy for structuring the output representation z_i to have these properties.

We first establish a pseudo point map distribution from which we can randomly sample point maps $z_{\texttt{point}} \in \{0,1\}^{w\times h}$. We convolve $z_{\texttt{point}}$ with K_δ, a 2D kernel with std. dev. $\delta = 1.5$, to generate a pseudo density map:

$$\tilde{z}_{\texttt{dmap}} = K_\delta * z_{\texttt{point}}. \quad (3)$$

Given these pseudo density maps, we establish an adversarial training objective that penalizes the network output z_i when it deviates from the properties of a density map.

Adversarial training [30] is a widely adopted technique for modeling the underlying generating distribution that explains a dataset. Our training strategy involves optimizing two neural networks, our counting network f (termed the generator) and a discriminator D. The generator is tasked with generating samples that appear as though they are sampled from the underlying pseudo density map distribution. The discriminator is tasked with evaluating whether a sample came from the pseudo density map distribution or the generator's distribution. The generator is optimized using feedback from the discriminator. We use the LS-GAN objective function [31], which is given as:

$$\mathcal{L}_{gan}^{f} = -E_x \left[D(f_\theta(x_i) - 1)^2 \right], \tag{4}$$

for the generator, and:

$$\begin{aligned} \mathcal{L}_{gan}^{D} = &- E_{\tilde{z}_{\texttt{dmap}}} \left[(D(\tilde{z}_{\texttt{dmap}}) - 1)^2 \right] \\ &+ E_x \left[D(f_\theta(x_i))^2 \right]. \end{aligned} \tag{5}$$

for the discriminator.

To produce the pseudo point map distribution, we uniformly sample a total count, c_{pseudo}, for the number of Gaussian blobs in a particular density map:

$$c_{pseudo} \sim \mathcal{U}_{\{0,N_c\}}, \tag{6}$$

where N_c is a hyper-parameter roughly corresponding to the estimated maximum object count for the dataset. Then, we uniformly sample c_{pseudo} co-ordinates:

$$i, j \sim \mathcal{U}_{[0,w]\times[0,h]}, \tag{7}$$

which gives us $\texttt{z}_{\texttt{point}}$ by setting all points (i, j) to 1, and all other points to 0.

IV. Experiments

A. Datasets

We benchmark our results on three sparse object counting datasets, which we consider to be datasets with fewer than 50 objects per image. Intuitively, this setting will be the most challenging, as there will be a greater potential for problematic spurious background features which can satisfy the ranking constraint. By specifically increasing the be the problem more challenging. TRANCOS [8] is a vehicle counting dataset containing 1,244 images of 46,796 highly occluded vehicles in traffic. Penguins [1] is an animal counting dataset containing around 82,000 images of penguin colonies. Each image contains several dot maps from different annotators. We use the mixed setting for training and evaluation. The Mall dataset [2]–[5] is a crowd counting data containing 2,000 images of over 60,000 pedestrians in a shopping malls. All of these datasets are challenging benchmarks as they contain highly occluded objects with a variety of environmental conditions and scales.

B. Sampling Image Ranking Data

There are presently no well-established image ranking datasets available for weakly supervised object counting benchmarking. Given this, all image ranking datasets used for evaluating our experiments must be curated. We experiment with the object counting datasets outlined above, and reformulate all of the available datasets as ranking datasets as follows. Given a counting dataset

$$\mathcal{D}_{count} = \{x_i \in R^{h,w,d}, c_i \in N\}_{i=0}^{N_p},$$

where d is the number of channels, c_i is the object count in image x_i, and N_p is the number of counting examples, we sample $N = 2,000$ image pairs (x_i, x_j) from $\mathcal{D}_{count}$. We then calculate their pairwise ranking as $r_{ij} = c_i \geq c_j$. This provides us with a curated ranking dataset:

$$\mathcal{D}_{rank} = \{(x_i, x_j)_n, (r_{ij})_n \equiv (c_i \geq c_j)_n\}_{n=1}^{N}.$$

For our experiments, we impose no constraints on the sampling procedure and generate the training ranking dataset by simply uniformly sampling examples from the training dataset. We simulate label noise by randomly selecting half of the examples in $\mathcal{D}_{rank}$ that violate the Weber-Fechner ratio, and flip their label. Given a dataset which simulates annotator error for difficult examples, we demonstrate that our model can even tolerate this noise.

C. Implementation Details

We use ResNet50 [32] as the base architecture for the counting model f_θ. We then generate a density map z_i by up sampling features using transposed convolutions [33]. To construct the ranking network, we pass two images through f_θ and then we take the sum of the difference between the respective outputs. The density map estimates, z_i and z_j, are passed to the discriminator. The discriminator is comprised of 5 convolutional layers which downsample the density map. Then, these features are passed through a final fully connected layer. Each model for each experiment is trained for 200 epochs using the Adam optimizer with a batch size of 32. Wet set the learning rates to 5×10^{-5}, 3×10^{-5} and 7×10^{-5} for the Trancos, Mall and Penguins datasets respectively. Due to the instability of GANs, selecting the best model is difficult. To mitigate this, we perform early stopping using the small validation set of 15 examples annotated with global counts.

D. Results

Ablation of Model Components.: In Table I, we explore the contribution of each model component to the test error: (i) $\mathcal{L}_{\texttt{rank}}$ (eq. 2); (ii) $\mathcal{L}_{\texttt{gan}}$ (eq. 5 & 4); and (iii) the ranking loss with simulated Weber-Fechner annotator noise. We evaluate each experiment using MAE and R^2. MAE is the mean absolute error and measures the difference between the predicted count and ground truth count, with a lower score

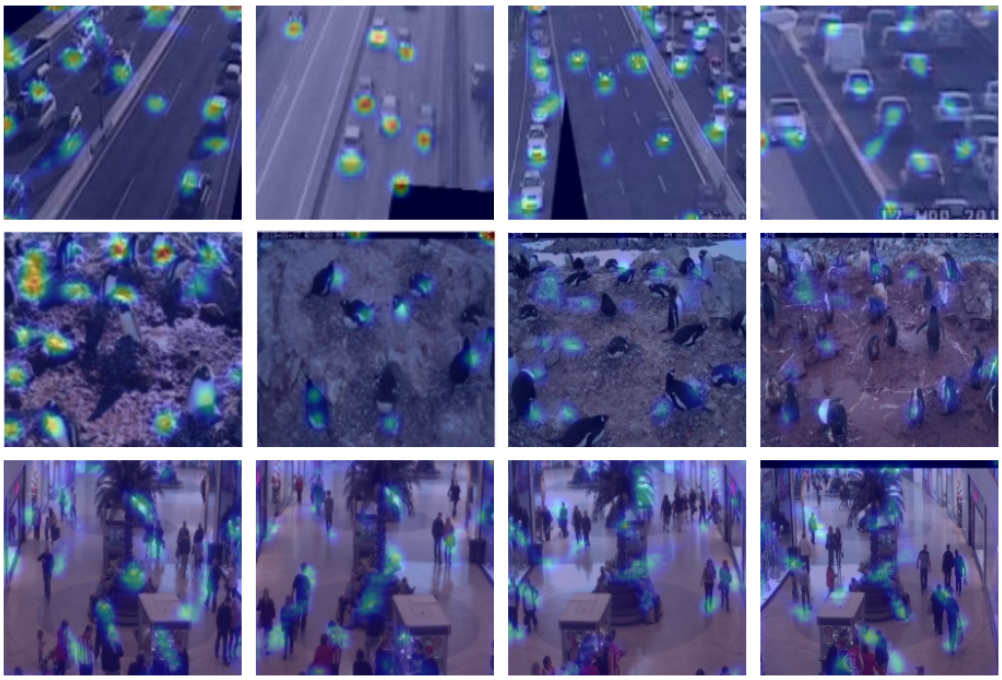

Figure 4. Qualitative examples of density maps predicted by the baseline model at inference time. Top: Trancos. Middle: Penguins. Bottom: MALL.

corresponds to a better performing model. R^2 is the coefficient of determination, which describes how well the model fits the data with a higher score corresponding to a better fit. When only $\mathcal{L}_{\texttt{gan}}$ is included during model training, we find that the model fails to detect and count objects. This result is intuitive, as the weak quantity signal is only provided by $\mathcal{L}_{\texttt{rank}}$. However, when only $\mathcal{L}_{\texttt{rank}}$ is included during model training, we find that the model under-performs when compared to the model trained with both $\mathcal{L}_{\texttt{rank}}$ and $\mathcal{L}_{\texttt{gan}}$. This result demonstrates that $\mathcal{L}_{\texttt{gan}}$ contributes an important training signal when solving the object counting problem using pairwise image rankings. Interestingly, when noisy ranking data is used as the source of weak quantity signal, the model provides a nearly equivalent performance when compared to the clean ranking annotations. Given that this noise primarily affects examples that are close in their underlying count, and given that the noise quantity is small, we argue that our model is robust to annotator noise. Figure 4 shows qualitative examples of the density map proposals learned by the network. The network learns to detect the relevant objects in many cases.

Evaluating the Annotation Burden vs. Error Trade-Off.: We compare our results with state-of-the-art object counting methods and we provide an estimate of the annotation burden for each method. We estimate the annotation burden for dot-maps and global object counts calculated over the dataset and compare it to the estimated annotation burden of our method. To estimate the annotation burden for dot-maps, we use the per-object annotation time of 1.1 s established by Cholakkal et al. [15] and multiply this by the number of objects in the dataset. To estimate the annotation time for global object counts, we use a slightly more complex formula, which includes the human ability to rapidly count objects within the range of 1 to 4, often referred to as the subitizing range. Saltzman et al. [41] established a counting speed of 0.1 s for each object within the subitizing range and 0.35 s for each additional object outside of the range. However, the participants in this experiment were only asked to count simple shapes. Cholakkal et al. [15] evaluated human counting in complex scenes and established a counting speed of 0.5 s within the subitizing range and 1.0 s for objects outside of this range. We use these two measures to create a range for our estimate of object counting speed

Table I
ABLATION STUDY OF MODEL MODIFIED BY REMOVING DIFFERENT LOSS COMPONENTS. RANKING NOISE REFERS TO THE INCLUSION OF WEBER-FECHNER BASED SIMULATED ANNOTATOR NOISE.

Method	Trancos		Mall		Penguins	
	MAE	R^2	MAE	R^2	MAE	R^2
$\mathcal{L}_{\texttt{rank}} + \mathcal{L}_{\texttt{gan}}$	5.22	0.81	2.60	0.68	7.22	0.58
$\mathcal{L}_{\texttt{gan}}$	13.19	-0.15	4.80	0.00	13.70	-0.08
$\mathcal{L}_{\texttt{rank}}$	9.47	0.51	5.46	-0.14	7.72	0.58
$\mathcal{L}_{\texttt{rank}} + \mathcal{L}_{\texttt{gan}}$ +ranking noise	5.42	0.80	2.62	0.69	7.33	0.57

Table II
TEST ERROR AND ANNOTATION TIME FOR SOTA COUNTING METHODS ON THE TRANCOS CAR COUNTING DATASET. WHEN METHODS USE THE DOT-MAPS, THEIR ANNOTATION TIMES ARE EQUIVALENT.

Method	Supervision	Est. Annotation Time	MAE
Hydra CCNN [34]	Dot-map	551 min	10.99
FCN-MT [35]	Dot-map	551 min	5.31
FCN-HA [36]	Dot-map	551 min	4.20
LC-PSPNet [37]	Dot-map	551 min	3.57
CSRNet [38]	Dot-map	551 min	3.56
SPN [39]	Dot-map	551 min	3.35
ADSCNet [14]	Dot-map	551 min	2.60
Glance [11]	Counts	161 min - 474 min	7.00
Adv. Dmap (Ours)	Pairwise Rank	25 min	5.42

Table III
TEST ERROR AND ANNOTATION TIME FOR SOTA COUNTING METHODS ON THE MALL CROWD COUNTING DATASET.

Method	Supervision	Est. Annotation Time	MAE
CNN-Boosting [40]	Dot-map	433 min	2.01
LC-PSPNet [37]	Dot-map	433 min	2.01
AL-AC [20]	10% Dot-map	43 min	3.80
Adv. Dmap (Ours)	Pairwise Rank	25 min	2.62

and assign a lower and upper bounds for the counting time of $[0.1, 0.5]$ seconds for each object within the subitizing range and $[0.35, 1.0]$ seconds for objects outside of the range. To calculate pairwise image ranking speed, we use the per image-pair ranking time of 0.75 s established by [26], [27] and multiply it by the number of image-pairs in our ranking dataset. This calculation assumes that annotators noise is permitted, and annotator are encouraged to prioritize speed over accuracy within the Weber-Fechner ratio.

Table II compares our method to previous state-of-the-art counting methods evaluated on the TRANCOS dataset, where we find that our method performs similarly to the method proposed by [35], despite their method being supervised by dot maps requiring $\times 22$ the annotation time. More recent methods, such as the method proposed by [14], outperform ours by a mean error of 2.82 vehicles per images, where each image contains an average of 38 vehicles. However, our method requires 4.54% of the annotation time as the best performing fully supervised methods. We also find that our method outperforms *Glance* [11], which learns from image-level object counts, while also requiring a smaller annotation burden (by a factor of 15.5% to 5.27%). Likewise, Table III presents the same comparison evaluated on the MALL dataset. We find that our method performs comparably to current state-of-the-art counting methods, while requiring a fraction (5.77% to 5.81%) of the annotation time. The best performing methods outperform ours by a mean absolute error of 0.61 pedestrians. Further, our method outperforms the method proposed by [20], which was specifically developed to deal with the annotation burden, while only requiring a fraction (5.81%) of the annotation time. These results demonstrate the value of pairwise image-ranking as a weak object counting signal and the value of our method for extracting counts while minimizing the annotation burden.

V. CONCLUSION

We present a solution to the weakly supervised object counting problem using, for the first time, an approach to extract object counts and locations from inter-image ranking annotations. These pairwise image ranking annotations can be rapidly collected by annotators, requiring only a single click. We develop a GAN based strategy for regularizing the network's intermediate representation such that it proposes valid density maps. We demonstrate that our method performs well on benchmarks, and approaches the performance of fully supervised counting methods, while requiring a fraction of the annotation cost. We further show that our method is robust even in presence of simulated annotator noise. This work demonstrates the value of exploring novel weak counting annotation formulations and suggests a direction forward for solving counting problems in domains where annotation collection would otherwise be impermissible.

REFERENCES

[1] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," in *ECCV*, 2016.

[2] C. Change Loy, S. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *IEEE ICCV*, 2013, pp. 2256–2263.

[3] K. Chen, S. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in *IEEE CVPR*, 2013, pp. 2467–2474.

[4] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, simulation and visual analysis of crowds*. Springer, 2013, pp. 347–382.

[5] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting." in *Bmvc*, vol. 1, no. 2, 2012, p. 3.

[6] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *IEEE CVPR*, 2016, pp. 589–597.

[7] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *IEEE CVPR*, 2015, pp. 833–841.

[8] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Onoro-Rubio, "Extremely overlapping vehicle counting," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2015, pp. 423–431.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *NeurIPS*, vol. 28, pp. 91–99, 2015.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE CVPR*, 2016, pp. 779–788.

[11] P. Chattopadhyay, R. Vedantam, R. R. Selvaraju, D. Batra, and D. Parikh, "Counting everyday objects in everyday scenes," in *IEEE CVPR*, 2017, pp. 1135–1144.

[12] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *IEEE CVPR*, 2018, pp. 7661–7669.

[13] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *NeurIPS*, vol. 23, pp. 1324–1332, 2010.

[14] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *IEEE/CVF CVPR*, 2020, pp. 4594–4603.

[15] H. Cholakkal, G. Sun, S. Khan, F. S. Khan, L. Shao, and L. Van Gool, "Towards partial supervision for generic object counting in natural scenes," *IEEE TPAMI*, 2020.

[16] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in CNNs by self-supervised learning to rank," *IEEE TPAMI*, vol. 41, no. 8, pp. 1862–1878, 2019.

[17] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost unsupervised learning for dense crowd counting," in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8868–8875.

[18] V. Ranjan, U. Sharma, T. Nguyen, and M. Hoai, "Learning to count everything," in *IEEE/CVF CVPR*, 2021, pp. 3394–3403.

[19] V. Ranjan, B. Wang, M. Shah, and M. Hoai, "Uncertainty estimation and sample selection for crowd counting," in *ACCV*, 2020.

[20] Z. Zhao, M. Shi, X. Zhao, and L. Li, "Active crowd counting with limited supervision," in *ECCV*. Springer, 2020, pp. 565–581.

[21] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang, and N. Sebe, "Weakly-supervised crowd counting learns from sorting rather than locations," in *ECCV 2020*. Springer, 2020, pp. 1–17.

[22] Y. Lei, Y. Liu, P. Zhang, and L. Liu, "Towards using count-level weak supervision for crowd counting," *Pattern Recognition*, vol. 109, p. 107616, 2021.

[23] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *22nd ICML*, 2005, pp. 89–96.

[24] S. Chen, C. Zhang, and M. Dong, "Deep age estimation: From classification to ranking," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2209–2222, 2018.

[25] K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *ECCV*. Springer, 2016, pp. 753–769.

[26] P. Pica, C. Lemer, V. Izard, and S. Dehaene, "Exact and approximate arithmetic in an amazonian indigene group," *Science*, vol. 306, no. 5695, pp. 499–503, 2004.

[27] J. Halberda and L. Feigenson, "Developmental change in the acuity of the" number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults." *Developmental psychology*, vol. 44, no. 5, p. 1457, 2008.

[28] S. Dehaene, "The neural basis of the Weber-Fechner law: a logarithmic mental number line," *Trends in cognitive sciences*, vol. 7, no. 4, pp. 145–147, 2003.

[29] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," *arXiv preprint arXiv:2103.14749*, 2021.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[31] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *IEEE ICCV*, 2017, pp. 2794–2802.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015, pp. 3431–3440.

[34] D. Oñoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 615–629.

[35] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura, "Understanding traffic density from large-scale web camera data," in *IEEE CVPR*, 2017, pp. 5898–5907.

[36] ——, "FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras," in *IEEE ICCV*, 2017, pp. 3667–3676.

[37] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Where are the blobs: Counting by localization with point supervision," in *ECCV*, 2018, pp. 547–562.

[38] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *IEEE CVPR*, 2018, pp. 1091–1100.

[39] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *2019 IEEE WACV*. IEEE, 2019, pp. 1941–1950.

[40] E. Walach and L. Wolf, "Learning to count with CNN boosting," in *ECCV*. Springer, 2016, pp. 660–676.

[41] I. Saltzman and W. Garner, "Reaction time as a measure of span of attention," *The Journal of psychology*, vol. 25, no. 2, pp. 227–241, 1948.