Scale-Invariant Implicit Neural Representations For Object Counting

Supplementary Material

6. Supplementary information of experiments

6.1. Data

RSOC [11]: The Remote Sensing Object Counting (RSOC) dataset is a large-scale benchmark specifically designed for counting objects in satellite imagery. It includes a total of $3,057$ images with $286,539$ annotated object instances. The dataset is divided into four distinct subgroups, each focused on a different object type: Buildings, Small Vehicles, Large Vehicles, and Ships. The RSOC Buildings dataset contains 1,205 training images and 1,263 test images, where the image resolution is 512×512. The RSOC Small Vehicles dataset has 222 training images and 58 test images. The image resolution ranges from 421×799 to 12029×5014. Large Vehicles consists of 108 training images and 64 test images, The image resolution ranges from 731×596 to 6327×5662. And the Ships subset has 97 training images and 60 test images. The image resolution ranges from 606×1065 to 6335×3591.

CARPK [14]: The Car Parking Lot Dataset (CARPK) is a benchmark for car counting tasks, consisting of 1,148 images taken from drone perspectives over four parking lots, containing 89,777 annotated cars. These images capture real-world scenarios with dense vehicle arrangements, making the dataset challenging for object detection and counting tasks. The average resolution of the images is 1280×720 pixels, providing detailed aerial views. Each image is annotated with bounding boxes around individual cars, making the dataset suitable for both object counting and detection. The dataset is split into 989 training images and 459 testing images.

PUCPR+ [14]: The Pontifical Catholic University of Parana+ Dataset (PUCPR+) is a specialized car counting resource where all images are captured from the 10th floor of a building. PUCOR+ contains 125 images with $16,456$ cars, where 100 images are set for training, while the remaining images are utilized for testing the models.

Visualization We further provide several exemplar images from RSOC datasets in Figure 5. It can be found that the objects within the same image naturally appear with similar size. However, remote sensing datasets, including RSOC, encompass images with a wide range of resolutions. As a result, object sizes vary significantly across different images, even if they appear uniform within a single image. Furthermore, we resize images to various resolutions to evaluate robustness to scale variability which further increases the range of scale differences across different images in our experiments. In figure 5, the top-left two images are both from the RSOC large-vehicle dataset, clearly showing that the cars in the second image are three times larger than those in the first image. Similarly, the bottom-left two images, from the RSOC small-vehicle dataset, highlight the differences in visibility: cars are clearly seen in the first image but are almost invisible in the second.

6.2. Baselines

In this section, we delve into the training specifics for all baseline models utilized in our experiments.

ASPDNet [11]: ASPDNet is an advanced attention-based network that integrates scale pyramids and deformable convolutions to effectively utilize attention mechanisms. This architecture captures extensive contextual and high-level semantic information, which aids in reducing the impact of cluttered backgrounds while emphasizing the regions of interest. In our study, we follow its original network design, set the batch size as 16, replace the original Stochastic Gradient Descent (SGD) optimizer with ADAM [19] optimizer, and set the learning rate as $1e-4$ to enhance the training results. Our training spans 200 epochs. ASPDNet is trained under MSE counting and Bayes counting [37] respectively to get the best counting performance.

PSGCNet [12]: PSGCNet integrates pyramidal scale and global context modules to handle scale variations of remote sensing images. We follow the network setup, setting the learning rate as $1e-4$, and using a batch size of 16. We trained PSGCNet with original Bayesian-based counting loss and MSE respectively. Our training spans 200 epochs.

eFreeNet [15]: The eFreeNet is an ensemble of first-rank-then-estimate networks that tailors a ranking metric optimization scheme to fit object counting. The study employs the default network architecture. In the optimization setup, we set the backbone's learning rate as $1e-5$ while $1e-5$ for other components following the original setup in Huang et al. [15]. The ensemble number is set as 8 to get the best counting performance. Our training spans 3000 epochs with a batch size of 8.

STEERER [13]: STEERER (Resolving Scale Variations for Counting and Localization via Selective Inheritance Learning) is designed to tackle scale variations in object counting by leveraging selective inheritance learning. We adopt the default network architecture and follow the original training protocol. The model is trained with a batch size of 8 for 300 epochs.

6.3. Comparison on UCF-QNRF dataset

We compare our SI-INR with state-of-the-art methods as well as our baselines on the UCF-QNRF (University of Central Florida - Qatar National Research Fund) dataset [16], which is a highly diverse dataset consisting of 1,535 images with over 1.2 million annotated individuals, spanning a wide range of crowd densities and changing object sizes. We report the results in Table 5.

The results demonstrate that SI-INR(PSGCNet) and SI-INR(STEERER) achieve competitive performance on crowd counting datasets. Notably, integrating SI-INR with PSGCNet leads to improved counting accuracy. For STEERER, SI-INR(STEERER) maintains comparable counting performance. As

Figure 5. Example images from the RSOC datasets.

Table 5. Performance Comparison on the UCF-QNRF Dataset

Model	**MAE**	**RMSE**
MMNet [8]	104.00	178.00
MSFFA [26]	94.60	170.60
MFANet [60]	97.7	166.00
CLTR [27]	85.80	141.30
Bayesian+ [37]	88.70	154.80
P2PNet [46]	85.32	154.50
GauNet [6]	81.60	153.71
APGCC [5]	80.10	136 .60
PSL-Net [43]	85.50	144.40
PET [32]	79.53	144.32
STEERER [13]	**74.30**	128.30
PSGCNet [12]	86.30	149.50
SI-INR (PSGCNet)	80.89	134.73
SI-INR (SETTRER)	74.81	**125.45**

discussed earlier, SI-INR enhances the robustness of baseline models to input scale variations. These experiments confirm that our method preserves counting accuracy while improving generalization across different input scales.

6.4. Visualization the influence of S_{INR} to SI-INR

To further demonstrate the effect of S_{INR} of SI-INR, we visualize SI-INR(PSGCNet)'s outputs when setting the S_{INR} from 8 to 128 in the Figure 6. In this ablation experiment, we let the well-trained SI-INR model directly generate 5 different resolution density maps, we can find that SI-INR can generate high-quality density maps when S_{INR} increases.

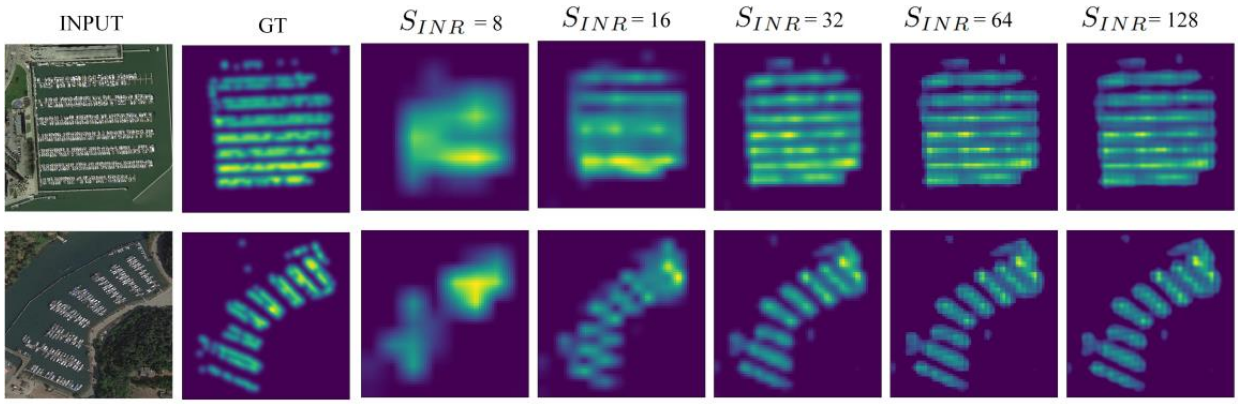

Figure 6. Predicted density maps by SI-INR(PSGCNet) with different S_{INR}.