This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Scale-Invariant Implicit Neural Representations For Object Counting

Siyuan Xu[1] Yucheng Wang[1] Xihaier Luo[2] Byung-Jun Yoon[1,2] Xiaoning Qian[1,2]
[1]Texas A&M University, College Station, TX, USA
[2]Computational Science Initiative (CSI), Brookhaven National Laboratory, Upton, NY, USA
{siyuanxu, wangyucheng, bjyoon, xqian}@tamu.edu, xluo@bnl.gov

Abstract

Existing object counting methods relying on Density Map Estimation (DME) struggle with large variations in object size or input image resolution due to different imaging conditions and perspective effects. Especially, discrete grid representations of density maps result in information loss with blurred or vanished details for low-resolution inputs. To overcome these limitations, we design new Scale-Invariant Implicit Neural Representations (SI-INR) for counting to map arbitrary-scale input signals into a continuous function space, where function values over continuous spatial coordinates indicate probabilities observing objects of interest. Extensive experiments on diverse benchmark datasets have validated that SI-INR achieves robust counting performances with respect to changing input sizes, leading to better or comparable object counting accuracy compared to state-of-the-art methods. Our code is available at https://github.com/SiyuanXu-tamu/SI-INR.

1. Introduction

Understanding the distribution and abundance of people as well as geographic entities such as buildings and cars becomes crucial for various "smart city" applications, such as urban planning, traffic management and beyond. Object counting holds promising potential for such tasks and has also been studied in other fields, including crowd counting for security [25, 37], animal crowd estimation in agriculture [36], and cell counting in biomedicine [41]. Successful counting methods have been developed by introducing deep learning [10, 54] and self-attention [11, 28]. In recent years, the best-performing methods are mostly based on training Convolutional Neural Networks (CNNs) to generate Density Map Estimation (DME) over discrete grid image domains [12, 28, 37, 51],

However, several challenges persist in applying current deep learning methods for reliable counting: 1) *Scale-Dependence*: CNNs [22] lack intrinsic scale equivariance, leading to degraded performance when input sizes deviate from those seen during training. This issue is particularly pronounced for inputs at resolutions differing from the training set, as CNNs rely on fixed receptive fields that cannot dynamically adapt to scale variations; 2) *Expressiveness-Bottleneck*: Traditional grid-based density maps approximate object distributions with Gaussian kernels, imposing a fixed spatial structure that misaligns with irregular object arrangements [50]. Gaussian smoothing reduces noise but blurs local details, degrading fidelity in dense and sparse regions, and limiting counting accuracy in complex scenes.

To address these issues, we design a new object counting framework named Scale-Invariant Implicit Neural Representations (SI-INR) mapping arbitrary-scale discrete images into 2D continuous functions which are invariant to the object or structure scales. This allows the model to preserve the fine details and reduce potential information loss for better counting accuracy and generalizability. Moreover, the scale-invariance, an important property for the mapping between input images and output density maps, is explicitly introduced as the inductive bias of model itself to potentially improve data efficiency and model robustness. Our main contributions can be summarized as follows:

1. We propose Scale-Invariant Implicit Neural Representations (SI-INR), an object counting framework mapping discrete grid signals into continuous 2D functions which are invariant to image scaling.
2. SI-INR adopts existing transfer learning algorithm and a novel deep neural operator based INR module to achieve scale-invariance. A sampling-based optimization objective is then derived for efficient model training. SI-INR can be easily integrated into existing methods, introducing scale-invariant properties to them.
3. We conduct extensive experiments to evaluate the effectiveness of our SI-INR on object counting, demonstrating notable performance improvements over state-of-the-art methods, especially on the remote sensing counting dataset.

2. Related Work

Object counting: Object counting, such as well-studied crowd counting [29], has been developed by detecting or segmenting individual objects in the scene. End-to-end learning to directly map image features to object counts has been the most successful counting strategy with rapid advancements in deep learning Krizhevsky et al. [21], Wang et al. [52], especially for object counting in densely populated scenes [4]. Counting based on Density Map Estimation (DME) using convolutional neural networks (CNNs) [10, 24, 42, 44] to preserve translation-invariant multi-scale image features has shown superior performance over conventional object counting techniques. More recent ASPDNet [11] and PSGCNet [12] have integrated attention, deformable convolution, pyramidal scale modules (PSM) to address challenges in counting such as complex cluttered backgrounds, viewing perspective, object appearance, and size variability. Besides, Huang et al. [15] proposed an optimized global regression model EfreeNet, which is more annotation-efficient. Yi et al. [57] introduced a lightweight multiscale context fusion module (LMCFM) and a lightweight counting scale pooling module (LCSPM) to reduce the model complexity and computing cost. These models have achieved state-of-the-art (SOTA) counting performance on the RSOC (Remote Sensing Object Counting) dataset [11].

Scale-equivariance and invariance: The concept of scale-equivariance and invariance was first proposed in image processing and computer vision [33, 34]. To handle variations in scale effectively, multi-scale features can be learned by applying the convolutions to several rescaled versions of the images or feature maps in every layer [17, 38] or by rescaling trainable filters [55]. Cai et al. [3] proposed a pyramidal structure to learn scale-dependent features, which is widely used in object detection. Later, Gaussian scale-space theory [30] and group theory [7] have been used for achieving scale-equivariance and invariance. Lindeberg [31], Yang et al. [56] parameterized convolutional filters as a linear combination of Gaussian derivative filters with different scales, and achieved scale-equivariance in image classification and segmentation tasks. Unlike models rooted in Gaussian scale-space theory, Sosnovik et al. [47] proposed a Scale-Equivariant Steerable Network (SESN), which utilizes steerable filters parameterized by a trainable linear combination of pre-calculated Hermite basis functions. These models all first build a scale-equivariant model, and use a simple pooling layer or rescale the outputs to convert the model into a scale-invariant one [47]. However, such methods have significant demands on memory and computational resources and can lose information in the equivariance to invariance conversion. In 2023, Basu et al. [2] proposed Equivariant Finetuning, which achieves equivariance by finetuning existing methods without the need to modify the architectures.

In SI-INR, we propose a scale-invariant framework by utilizing a scale-equivariant model to learn deep representations that adapt to input scale variations. Additionally, we introduce a scaling operator to transform the equivariant mapping into an invariant one.

Implicit neural representations: Implicit Neural Representations (INRs) allow for continuous flexible representations of complex objects and scenes [1, 39, 40, 53]. Together with positional encoding strategies [45, 49] and end-to-end hypernetwork-based learning [9, 18, 23] that help better capture high-frequency details, efficient model training has been developed for different computer vision tasks with complex natural signals. More recent Hierarchical Neural Operator Transformer (HiNOTE) [35] integrates neural operators in implicit neural representations, which can preserve more local information and improve the generalizability of INR models.

However, there is no scale-equivariant INR models to our knowledge. Existing INR models typically require the fixed-size inputs and their performances suffer when handling significant scale variations in inputs. In SI-INR, we adopt a lightweight INR implementation and replace the traditional coordinate input by continuous latent variables, which helps capture continuous representations of targets. This can be viewed as a deep neural operator [20] for object counting to map each object to a scale-invariant Gaussian distribution, offering greater flexibility when incorporating images of varying sizes during training.

3. Method

We develop a novel object counting framework, Scale-Invariant Implicit Neural Representation (SI-INR), that adopts continuous INR for robust object counting with scale variations. We start the discussion by first presenting the problems of existing methods in Section 3.1. Next, we describe the background of scale-invariance and equivariance in Section 3.2. We then present the detailed construct of SI-INR in Section 3.3 with the corresponding analysis. Lastly in Section 3.4, we provide our sampling based model training for SI-INR on object counting.

3.1. Problem Statement

In many real-world object detection and counting problems, input images can vary significantly in dimension and size. This variation necessitates models that can adapt to different input dimensions while maintaining accurate detection and counting performance.

Recent works, such as Gao et al. [12] and Li et al. [26], integrate multi-scale features to enhance information extraction. However, no existing approach directly utilizes scale-invariant features that remain consistent despite

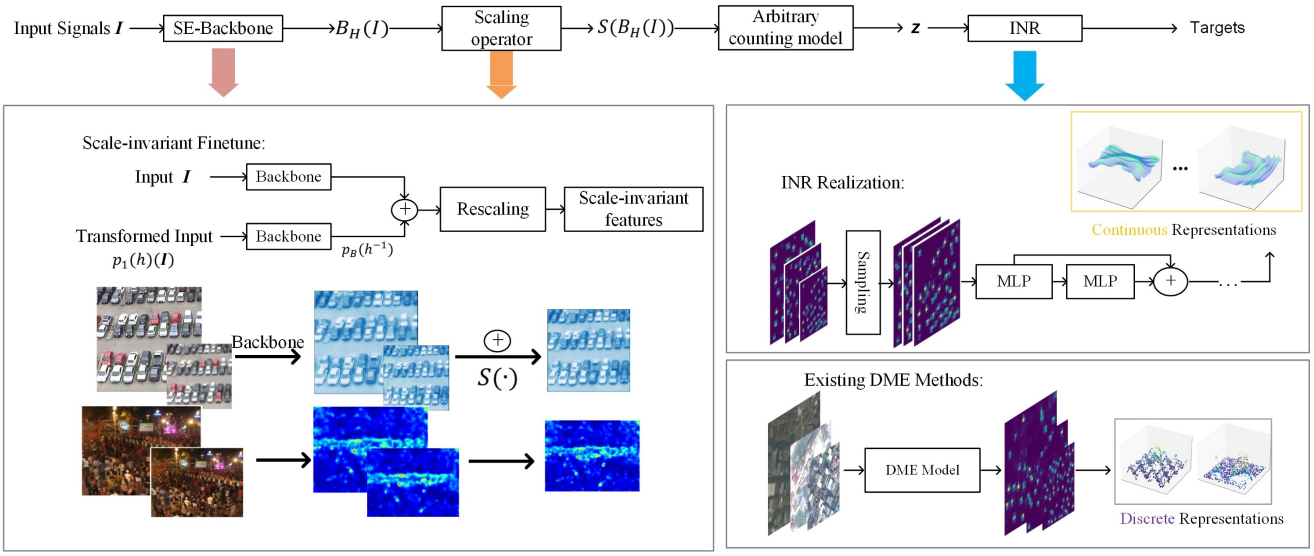

Figure 1. Schematic diagram of Scale-Invariant Implicit Neural Representation (SI-INR) compared to existing Density Map Estimation (DME) methods. SI-INR learns scale-invariant continuous representations in two steps: first, a scale-invariant backbone is finetuned to extract deterministic scale-equivariant features; then, an INR decoder converts extracted features into an invariant output, a continuous representation of task targets. Visualization of continuous and discrete representations demonstrates that continuous representations preserve more information, leading to better reconstruction of the continuous output.

changes in input resolutions. In this paper, we propose SI-INR, a novel object counting framework that maps inputs with varying resolutions into a consistent continuous representation of density maps. Instead of forcing the model to adapt to objects at different scales, SI-INR directly transforms inputs of different resolutions into a unified scale-invariant continuous feature space. This approach enhances feature consistency, providing a more robust foundation for detection while enabling models to handle varying input resolutions more effectively. By mapping inputs into a unified continuous space, SI-INR also facilitates the extraction of richer and more informative features, improving the model's ability to capture fine-grained details and structural patterns across different resolutions. Moreover, by leveraging efficient transfer learning techniques, we demonstrate that SI-INR can be seamlessly integrated into existing models.

3.2. Scale Invariance and Equivariance

Consider a Scale-Translation Group [47] consisting of a Scaling Group G_S and a Translation Group G_T, $H = \{h = (s,t) \,|\, s \in G_S, t \in G_T\}$, where h denotes an element of H and represents one scale-translation operator, G_S denotes the Scaling Group, which accounts for transformations that scale an object or function, and G_T denotes the Translation Group, which handles shifting the object or function within its domain. Besides, s is the scaling parameter, indicating how the input is stretched or compressed; t is the translation parameter, specifying the shifting in the domain.

From the group theory, given an image $\mathbf{I}_a \in V_1$, a mapping $\Phi : V_1 \to V_2$ is *scale-equivariant* if

$$\Phi(p_1(h)(\mathbf{I}_a)) = p_2(h)(\Phi(\mathbf{I}_a)),\ \forall h \in H \tag{1}$$

where V_2 denotes the output domain, $p_1(\cdot)$ and $p_2(\cdot)$ denote the corresponding group actions of h acting on V_1, V_2. If $p_2(h)$ is the identity mapping, the mapping Φ is *scale-invariant*.

3.3. SI-INR

For a given image $\mathbf{I}$, existing methods aim to establish a mapping $f : \mathbb{R}^{d_\mathbf{I}} \to \mathbb{R}^{d_{\mathbf{D}^{gt}}}$ from input image $\mathbf{I} \in \mathbb{R}^{d_\mathbf{I}}$ to corresponding output $\mathbf{D}^{gt} \in \mathbb{R}^{d_{\mathbf{D}^{gt}}}$.

However, traditional convolution models are not scale-invariant as they will generate different features for inputs of different scales, where $f(p_1(h)(\mathbf{I}_a)) \neq f(\mathbf{I}_a)$. This forces existing methods to build larger and more complex models to detect targets in different scales [8, 12, 26]. To address this limitation, we propose SI-INR to model the mapping for scale-invariant continuous signal representations. SI-INR learns a mapping $\Psi : \mathcal{I} \to \mathcal{F}$ from input image space $\mathcal{I}$ to the continuous function space $\mathcal{F}$, given by

$$\Psi(p_1(h)(\mathbf{I}_a))(\mathbf{x}) = \Psi(\mathbf{I}_a)(\mathbf{x}) = \mathbf{D}^{gt}(\mathbf{x}), \tag{2}$$

where we emphasize that the input image space $\mathcal{I}$ here is more flexible considering arbitrary normalized spatial coordinates, $\mathbf{x} \in [0,1]^2$, sampling from continuous image domain. $\Psi(\mathbf{I}_a)$ denotes the predicted continuous representation of density maps for $\mathbf{I}_a$. $\mathbf{D}^{gt}$ denotes a continuous

ground truth which contains more information than discrete density maps.

Following the above formulation, to achieve a scale-invariant mapping from the image space $\mathcal{I}$ to the function space $\mathcal{F}$, we propose the SI-INR modular framework consisting of two components: a scale-invariant encoder and an INR decoder. The encoder is designed to extract deterministic scale-invariant features resilient to different resolutions of inputs; then the INR decoder converts extracted features into a scale-invariant output, which is a continuous representation of density maps.

3.3.1. Scale-Invariant Encoder

Considering the fact that backbones pretrained on large datasets offer higher-quality features and are less prone to local optimization pitfalls compared to training from scratch. Instead of directly designing a scale-invariant model, SI-INR achieves scale-invariance through a two-step process as shown in Figure 1: First, transfer a pretrained backbone to a scale-equivariant one. Then, rescale scale-equivariant features to a scale-invariant representation.

Inspired by Equi-Tuning [2], SI-INR applies transfer learning to achieve scale-equivariance. Consider a non-equivariant pretrained backbone $\mathbf{B}(\cdot)$, we want to transfer $\mathbf{B}(\cdot)$ to a scale-equivariant one $\mathbf{B}_H(\cdot)$ by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{B}_H(\mathbf{I}_a)} \sum_{h\in H} \left\| p_B(h^{-1})\mathbf{B}(p_1(h)\mathbf{I}_a) - \mathbf{B}_H(\mathbf{I}_a) \right\|_2^2 \\ &\text{s.t. } \mathbf{B}_H(p_1(h)\mathbf{I}_a) = p_B(h)\mathbf{B}_H(\mathbf{I}_a) \text{ for all } h \in H, \end{aligned} \tag{3}$$

where $p_B(\cdot)$ denotes the corresponding group actions of h acting on the feature domain. One solution to this convex programming formulation is shown below in (4), via Reynold's operator [48] to $\mathbf{B}(\cdot)$:

$$\mathbf{B}_H(\mathbf{I}_a) = \frac{1}{|H|} \sum_{h\in H} h^{-1}\mathbf{B}(p_1(h)\mathbf{I}_a). \tag{4}$$

As illustrated in Fig. 1, this approach modifies the training process by allowing the pretrained backbone to process inputs at arbitrary scales and compute the average features. This effectively enables a pretrained model to achieve scale-equivariance without requiring architectural changes. Note that within SI-INR's scale-invariance framework, any scale-equivariant module can be incorporated to ensure scale-invariant feature extraction.

After getting a scale-equivariant backbone, SI-INR further rescales the features into a fixed resolution by a scaling operator $S(\cdot)$ to transform the scale-equivariant encoder to a scale-invariant one. Since SI-INR maps any scale features into a fixed resolution, we have

$$S(\mathbf{B}_H(p_1(h_1)\mathbf{I}_a)) = S(\mathbf{B}_H(p_1(h_2)\mathbf{I}_a)),\ \forall\, h_1, h_2 \in H. \tag{5}$$

In summary, our scale-invariant encoder $E(\cdot)$ satisfies $E(p_1(h)\mathbf{I}_a) = S(\mathbf{B}_H(p_1(h)\mathbf{I}_a)) = E(\mathbf{I}_a)$. We now prove the scale-invariance of our SI-INR for any scale-translation action on $\mathbf{I}_a$ in Theorem 1.

Theorem 1 *Given a scale-translation operation h and an input image $\boldsymbol{I}_a$, SI-INR is scale-invariant:*

Proof:

$$\begin{aligned} \Psi(p_1(h)(\mathbf{I}_a))(\mathbf{x}) &= \mathcal{H}(E(p_1(h)(\mathbf{I}_a)))(\mathbf{x}) \\ &= \mathcal{H}(E(\mathbf{I}_a))(\mathbf{x}) = \Psi(\mathbf{I}_a)(\mathbf{x}). \end{aligned} \tag{6}$$

where $\mathcal{H}(\cdot)$ denotes our INR-based decoder. Based on the transfer learning algorithm and scaling operator, any non-invariant encoder can be transformed into a scale-invariant one, enhancing its capability to process inputs of varying scales.

3.3.2. INR-based Decoder

In counting, traditional DME-based methods tend to predict a density map $\mathbf{D}$ to capture spatial distribution, handling scale variations and cluttered scenes while enabling pixel-level supervision for improved counting accuracy.

Given an image $\mathbf{I}_a$, let the counting label (annotation map) $\mathcal{D}_{\mathbf{I}} = \{(\mathbf{m}_n, y_n)\}_1^N$ where N is the number of labeled objects, $\mathbf{m}_n$ denotes the normalized image-coordinate-based position of the n-th object (typically its center or head), $\mathbf{m}_n \in [0,1]^2$ and $y_n = n$ denotes the corresponding object label. Then, the density map $\mathbf{D}^{gt}$ is modeled as 2D stochastic processes in the continuous spatial domain:

$$\begin{aligned} \mathbf{D}^{gt}(\mathbf{x}) &\overset{\text{def}}{=} \sum_{n=1}^{N} \mathcal{N}\left(\mathbf{x}; \mathbf{m}_n, \sigma^2 \mathbf{1}_{2\times 2}\right) \\ &= \sum_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|\mathbf{x}-\mathbf{m}_n\|_2^2}{2\sigma^2}\right), \end{aligned} \tag{7}$$

where $\mathbf{x}$ denotes the normalized spatial coordinates, $\mathbf{x} \in [0,1]^2$, $\mathcal{N}\left(\mathbf{x}; \mathbf{m}_n, \sigma^2 \mathbf{1}_{2\times 2}\right)$ denotes the 2D Gaussian distribution with the mean $\mathbf{m}_{\text{n}}$ and isotropic covariance matrix $\sigma^2 \mathbf{1}_{2\times 2}$.

Typically, annotation maps are converted into density maps by convolving with Gaussian kernels. However, traditional methods produce only a discrete representation, leading to information loss and inconsistent results across different scale inputs.

To address this problem, we introduce a scale-invariant continuous representation mapping $\mathcal{H}$. For a query image $\mathbf{I}_a$, $\mathcal{H}$ maps the output of encoder into a continuous function $u_a : [0,1]^2 \to \mathbb{R}_{\geq 0}$, which means the corresponding density value for arbitrary normalized query position $\mathbf{x}$ can be predicted by evaluating the mapped continues function at $\mathbf{x}$: $u_a(\mathbf{x})$. Given an image $\mathbf{I}_a$,

$$\mathcal{H}(S(\mathbf{B}_H(\mathbf{I}_a)), \boldsymbol{\theta}_{INR})(\mathbf{x}) = u_a(\mathbf{x}|\mathbf{z}^a, \boldsymbol{\theta}_{INR}), \tag{8}$$

where $\boldsymbol{\theta}_{INR}$ denotes the trainable parameters and the INR model consists of L_{INR} linear layers, $\boldsymbol{\theta}_{INR} = [\boldsymbol{W}_1, \boldsymbol{b}_1 \ldots, \boldsymbol{W}_{L_{INR}}, \boldsymbol{b}_{L_{INR}}]$. $\mathbf{z}^a$ denotes the scale-invariance continuous features.

We assume that $\mathbf{z}^a$ encodes continuous information, treating it as a continuous scale-invariant feature instead of 2D discrete arrays, which can fully utilize local details in downstream analyses [35]. As shown in Fig. 1, our SI-INR framework allows any arbitrary counting model to generate scale-invariant continuous features, $\mathbf{z}^a$, provide the flexibility to varing models that handling different targets.

For any query position $\mathbf{x}$, we obtain $\mathbf{z}^a_{\mathbf{x}}$ by sampling from $\mathbf{z}^a$ and use it as the input to the INR model. The INR model then analyzes the spatial structure and content to predict the density value $\mathbf{D}(\mathbf{x})$. We call this continuous-to-continuous INR module a deep neural operator based INR.

With all these, the predicted value at position $\mathbf{x}$ can be estimated as:

$$u_a(\mathbf{x}) = \boldsymbol{\phi}^{INR}_{L_{INR}}\big(\boldsymbol{W}^T_{L_{INR}} \ldots \boldsymbol{\phi}^{INR}_1(\boldsymbol{W}^T_1(\mathbf{z}^a_{\mathbf{x}}) + \boldsymbol{b}_1) \ldots + \boldsymbol{b}_{L_{INR}}\big). \tag{9}$$

where $\boldsymbol{\phi}^{INR}(\cdot)$ denotes the activation function.

In this way, SI-INR effectively models outputs as 2D stochastic processes. This continuous representation not only provides a unified space for inputs of different scales but also enables the model to restore more detailed information in density maps. Furthermore, SI-INR supports flexible training sample-based algorithms, enhancing the training process as discussed in Section 3.4.

Unlike traditional methods that rely on a fixed downsampling ratio, SI-INR leverages its continuous property to generate outputs at arbitrary resolutions during inference. By using a larger sampling grid, $S_{INR} \times S_{INR}$, SI-INR can produce more detailed density maps. Our experiments in 4.3 show that higher-resolution density maps lead to improved counting performance.

3.4. Training with Regional Sampling

To train a continuous representation of the density map, a continuous ground truth is needed. We achieve this by directly constructing the likelihood function of any position $\mathbf{x}$ given label y_n as $p(\mathbf{x} \mid y_n) = \mathcal{N}\left(\mathbf{x}; \mathbf{m}_n, \sigma^2 \mathbf{1}_{2\times 2}\right)$, which models the density map $\mathbf{D}^{gt}$ as 2D stochastic processes in the continuous spatial domain.

During the training, we randomly sample N_S positions $\mathbf{x}$ from the image domain and compute the corresponding $\mathbf{D}^{gt}(\mathbf{x})$ through interpolation from the discrete density maps. The counting loss function in SI-INR is:

$$\mathcal{L} = \frac{1}{A}\sum_{a=1}^{A}\Big[\mathbb{E}_{\mathbf{x}\sim p_s(\mathbf{x})}(\big\|\mathbf{D}^{gt}(\mathbf{x}) - \mathbf{D}(\mathbf{x})\big\|_2^2)\Big], \tag{10}$$

where $p_s(\mathbf{x})$ is any probability distribution of $\mathbf{x}$ which enables our model to be trained using any existing stochastic optimization algorithm.

The count predictions can be hereby obtained by sampling uniformly over the normalized image domain and computing the summation of $\mathbf{D}(\mathbf{x})$.

Building on these findings, we observe that adapting an existing model to the SI-INR framework requires only the integration of our scale-invariant fine-tuning and the addition of our lightweight INR module at the end of the model. This approach is both efficient and easily applicable to any existing method.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the model's performance on three challenging datasets: (1) the Remote Sensing Object Counting (RSOC) dataset [11]; (2) the Car Parking Lot Dataset (CARPK) [14]; and (3) the Pontifical Catholic University of Parana+ Dataset (PUCPR+) [14]. Details on the datasets and the train-test split are provided in Appendix 6.1.

Baselines. We compare SI-INR with four baseline methods: ASPDNet [11], an attention-based network with scale pyramid and deformable convolutions; PSGCNet [12], which integrates pyramidal scale and global context modules; eFreeNet [15], an ensemble of first-rank-then-estimate networks; and STEERER [13], which selects the most suitable scale for patch objects to boost feature extraction.

Implementation. Our SI-INR can be seamlessly integrated into any method requiring scale-invariance. To demonstrate its effectiveness in handling inputs of varying scales, we applied the SI-INR modification to two baseline models—PSGCNet [12] and STEERER [13]. We will call them SI-INR(PSGCNet) and SI-INR(STEERER) later. Experimental results confirm that SI-INR enhances these models' ability to process scale-varying inputs.

To implement the scale-equivariant backbone, we fine-tuned the backbone of the baseline models. For computational efficiency, our rescaling operator $S(\cdot)$ utilizes bilinear interpolation. The INR network comprises four fully connected layers with residual connections, followed by an additional fully connected layer with learnable parameters to generate raw density maps. We use the Adam [19] optimizer for both our SI-INR and baseline models, and set the learning rate to be $1\mathrm{e}-4$. We initialize parameters in SI-INR by random sampling from a Gaussian distribution $\mathcal{N}(0, 0.01^2)$. We follow the baselines' setup when generating density maps ground truth, $\sigma = 8$ for PSGCNet and ASPDNet, $\sigma = 15$ for STEERER. We evaluate our SI-INR with baseline models on the RSOC dataset, CARPK dataset, and PUCPR+ dataset. Data augmentation

has been implemented during model training by randomly flipping the input images horizontally. We select the models with the lowest RMSE and proper density maps in the first 300 training epochs and report the results. We run all our experiments with the fixed random seed 64 on a workstation with a NVIDIA V100 32GB GPU. We adopt two widely used metrics in object counting tasks following previous work [11, 37] to evaluate baselines and our model: the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). We additionally compare SI-INR with SOTA crowd-counting methods on the UCF-QNRF dataset in Appendix 6.3.

4.2. Main Results

Quantitative Results. Our performance evaluation on different benchmark datasets with the reported experimental results on RSOC in Table 1, and CARPK and PUCPR+ in Table 2, respectively. It shows that SI-INR improves the MAE on all the datasets compared with our baseline models. Note that the RSOC small-vehicle and RSOC ship datasets exhibit the largest scale variations and the smallest target objects within the RSOC dataset as we show in Appendix 6.1, the significant improvement on these datasets shows that SI-INR makes PSGCNet and STEERER more robust to scale changes of inputs. Compared with other methods with fixed downsample ratio and can only generate fixed scale density maps, SI-INR improves the counting performance by directly generating larger and clearer density maps as we showed in Table 4. These results highlight the significant advantages of SI-INR in handling targets across varying scales.

Visualization. Note that PSGCNet and STEERER follow different setup when generate ground truth, we visualize the predicted density maps generated by SI-INR(PSGCNet) and SI-INR(STEERER) separately in Figure 2 and Figure 3, excluding eFreeNet [15], as it is a regression-based counting method.

In Figure 2, all four images are randomly sampled from the RSOC dataset, with the first three from the RSOC ship dataset and the last one from the RSOC small-vehicle dataset. As shown, SI-INR(PSGCNet) delivers more accurate counting performance, particularly when the objects' appearance and scales are complex. In the first three images, SI-INR(PSGCNet) generates clearer density maps. In the last image from the RSOC small-vehicle dataset, where the cars are too small for the other SOTA methods to detect, SI-INR(PSGCNet)'s scale-invariant property enables it to produce higher-quality density maps and achieve better counting accuracy. In Figure 3, SI-INR (STEERER) produces nearly identical Gaussian distributions for different objects, a more stable output, whereas STEERER exhibits more fluctuating brightness variations

Inference Efficiency. Thanks to the transfer learning algorithm and the lightweight INR-based decoder, which consists of only four linear layers, integrating SI-INR into existing methods does not significantly impact inference speed. On the RSOC building dataset, ASPD-Net requires approximately 15.13 seconds for inference, PSGC-Net takes around 2.47 seconds, eFreeNet runs in about 3.84 seconds, and our SI-INR(PSGCNet) model completes inference in approximately 2.67 seconds. This slight increase in inference time results in a more robust model that remains invariant to scale and resolution variations.

Generalization Results. In this section, we evaluate the robustness of SI-INR(PSGCNet) and baseline methods to the scale variation in testing data by testing models using images with resolutions different from training images. To further increase the scale variance inside the RSOC dataset, we limit the training images in a fixed scale 512×512, rescale test images into 5 different resolutions: 205×205, 307×307, 410×410, 512×512, 614×614, and test our SI-INR(PSGCNet) as well as baselines on the rescaled test images. The varying scales in the test set better reflect real-world conditions, where input images appear at different sizes due to changes in altitude, camera settings, or image cropping. SI-INR(PSGCNet) significantly outperforms the baseline models when the variation of resolution in test image is presented. The performance advantages over baseline models illustrate that our SI-INR(PSGCNet) is not only more robust compared to other baselines when processing images with unseen resolutions, but also more data efficient. Especially on PUCPR+ dataset, SI-INR(PSGCNet) reduces MAE by 70.9% and RMSE by 71.64% compared with efREENet. Furthermore, as illustrated in Figure 4, SI-INR(PSGCNet) achieves superior counting accuracy and produces higher-quality density maps when facing different resolution inputs even with extremely low resolution like 104×104, demonstrating its enhanced ability to handle scale variance compared to traditional methods. Although the performances of all the models degrade with small-scale inputs, our SI-INR(PSGCNet) can still produce density maps with the objects well separated. Moreover, the underestimation of object counts is much less severe in SI-INR(PSGCNet) compared to the baselines, which demonstrates the robustness of SI-INR(PSGCNet) under scale variation.

4.3. Ablation studies

In this section, we test the sensitivity of the counting performance of our SI-INR(PSGCNet) model with respect to Sampling Rate S_{INR}.

We report the prediction accuracy and include the predicted density maps by our SI-INR trained with the loss estimated by sampling from the grids of different size $S_{INR} \times S_{INR}$ in Table 4. In this section, we evaluate the counting performance of SI-INR(PSGCNet) on the RSOC ship

Table 1. Comparison of counting performances on the RSOC datasets.

Method	Loss		Building		Small-vehicle		Large-vehicle		Ship	
	MSE.	Bayes.	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN	✓		13.65	16.56	488.65	1317.44	36.56	56.55	263.91	412.30
eFreeNet	✓		6.99	**9.61**	195.86	463.62	14.55	19.77	65.34	85.45
PSGCNet	✓		7.33	11.02	346.78	952.64	21.54	32.75	75.27	94.79
PSGCNet		✓	7.18	10.98	196.25	360.15	14.47	26.19	72.07	98.06
ASPDNet	✓		7.40	11.06	378.23	978.93	18.76	31.06	63.32	84.85
ASPDNet		✓	7.59	11.07	365.69	1101.25	16.61	29.26	64.82	89.24
STEERER	✓		6.60	9.95	84.18	197.65	9.96	15.76	38.38	55.13
SI-INR(PSGC)		✓	**6.54**	9.80	157.18	306.43	12.61	21.78	59.76	81.79
SI-INR(STEERER)	✓		6.81	10.31	**78.06**	**192.83**	**8.56**	**15.33**	**33.40**	**46.39**

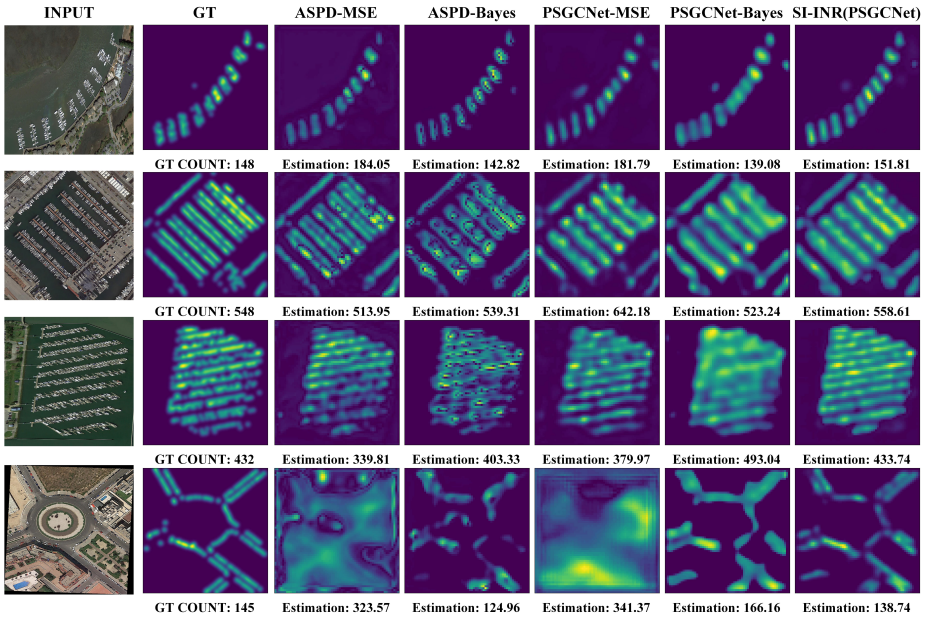

Figure 2. Predicted density maps by SI-INR(PSGCNet) and its baselines for four test images from RSOC. The test images (Test Images) and corresponding density maps (GT) are randomly sampled. The illustrated density maps are predicted by PSGCNet with MSE loss (PSGCNet+MSE), PSGCNet with Bayesian counting loss (PSGCNet+Bayes), ASPDNet with MSE loss (ASPD+MSE), ASPDNet with Bayesian counting loss (ASPD+Bayes), and SI-INR. Warmer colors denote higher values while cooler colors denote lower values.

Table 2. Comparison of counting performances on the CARPK and PUCPR+ datasets.

Method	Loss		CARPK		PUCPR+	
	MSE.	Bayes.	MAE	RMSE	MAE	RMSE
MCNN [59]	✓		24.95	39.63	21.86	29.53
eFreeNet [15]	✓		46.42	52.34	18.98	23.03
PSGCNet [12]		✓	11.07	14.55	3.87	4.86
PSGCNet [12]		✓	7.71	10.28	3.17	5.27
ASPDNet [11]	✓		10.01	12.84	4.21	5.02
ASPDNet [11]		✓	9.98	13.19	4.48	5.93
SAFECount [58]	✓		5.33	7.04	2.24	3.44
STEERER [13]	✓		3.89	5.73	2.84	3.64
SI-INR(PSGCNet)		✓	5.54	7.43	**2.09**	**2.70**
SI-INR(STEERER)	✓		**3.38**	**4.33**	2.55	3.36

and small-vehicle datasets with $S_{INR} = 8, 16, 32, 64, 128$. The results show that SI-INR(PSGCNet) achieves better

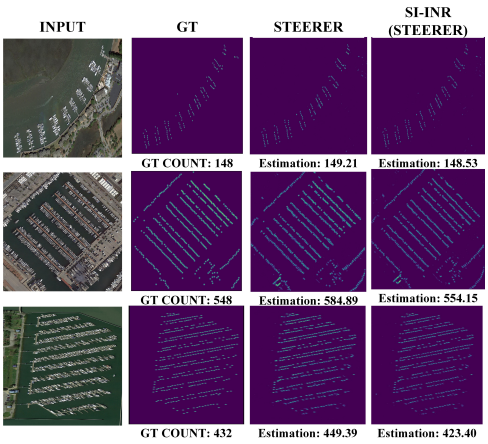

Figure 3. Predicted density maps by SI-INR(STEERER) and its baseline for test images from RSOC.

Table 3. Counting performance of handling unseen scales images on the CARPK, PUCPR+, RSOC building datasets and RSOC large-vehicle datasets.

Method	Loss		CARPK		PUCPR+		Building		Large-vehicle	
	MSE.	Bayes.	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
eFreeNet	✓		50.37	56.83	30.45	35.59	16.97	19.72	49.39	57.26
PSGCNet	✓		39.36	54.13	49.51	77.05	11.56	16.43	28.74	46.22
PSGCNet		✓	37.63	52.46	32.58	52.38	12.09	16.96	22.47	39.99
ASPDNet	✓		41.28	53.62	46.31	75.18	11.31	15.60	26.86	40.17
ASPDNet		✓	37.25	52.26	35.02	63.90	11.37	16.11	22.11	39.37
SI-INR(PSGCNet)		✓	**24.30**	**28.09**	**8.85**	**11.91**	**7.96**	**11.29**	**21.89**	**30.47**

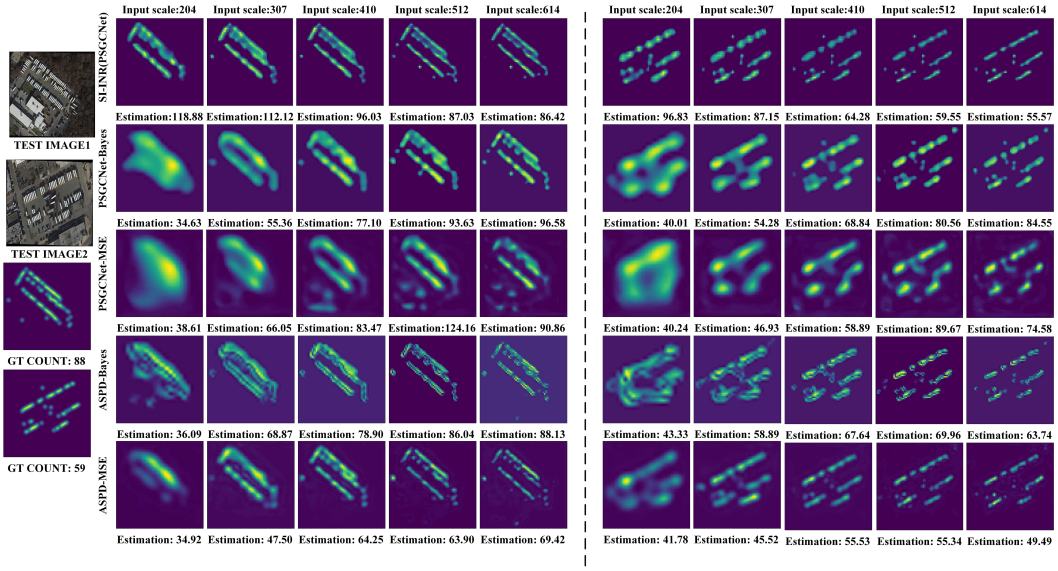

Figure 4. Predicted density maps by SI-INR(PSGCNet) and other baselines for two test images from RSOC. Two test images are rescaled to 205×205, 307×307, 410×410, 512×512, 614×614 before fed into the models.

counting performance as S_{INR} increases from 8 to 128. This effect is particularly pronounced for the RSOC small-vehicle dataset, where S_{INR} significantly impacts counting accuracy. Since the targets are very small, training with a higher S_{INR} helps the model more accurately locate the vehicles. Besides, this continuous property helps make it easy to balance the computation costs and the counting accuracy requirement. We further visualize several results in Appendix 6.4.

Table 4. Effect of sampling rate S_{INR} on object counting in SI-INR(PSGCNet).

Method	RSOC-ship		RSOC-small-vehicle	
	MAE	RMSE	MAE	RMSE
S_{INR}=8	127.14	173.21	277.50	1017.56
S_{INR}=16	65.49	93.03	273.78	1016.11
S_{INR}=32	62.97	86.81	255.12	821.21
S_{INR}=64	62.26	83.98	243.66	731.51
S_{INR}=128	**59.76**	**81.79**	**157.18**	**306.43**

5. Conclusions

In this paper, we introduce SI-INR, a novel scale-invariant INR implementation that maps discrete grid image signals into continuous 2D function space, maintaining invariance to scaling variation of the input signals. For object counting, SI-INR achieves SOTA performance, our experiments demonstrate that SI-INR makes existing methods exceptionally more robust and flexible, capable of processing images of unseen resolutions during testing and effectively handling images of various scales during training. This flexibility allows SI-INR to learn and capture more detailed features from different input training images. SI-INR can be easily applied to other image analysis tasks to achieve arbitrary-scale SOTA performance robustly with respect to input image size/resolution. Future work will focus on applying SI-INR to multi-task scenarios.

Acknowledgements. This work was supported in part by the Department of Energy (DOE) Award DE-SC0012704, and the National Science Foundation (NSF) Award IIS-2212419.

References

[1] Oliver JD Barrowclough, Georg Muntingh, Varatharajan Nainamalai, and Ivar Stangeby. Binary segmentation of medical images using implicit spline representations and deep learning. *Computer Aided Geometric Design*, 85:101972, 2021.

[2] Sourya Basu, Prasanna Sattigeri, Karthikeyan Natesan Ramamurthy, Vijil Chenthamarakshan, Kush R Varshney, Lav R Varshney, and Payel Das. Equi-tuning: Group equivariant fine-tuning of pretrained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6788–6796, 2023.

[3] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 354–370. Springer, 2016.

[4] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[5] I Chen, Wei-Ting Chen, Yu-Wei Liu, Ming-Hsuan Yang, Sy-Yen Kuo, et al. Improving point-based crowd counting and localization based on auxiliary point guidance. In *European Conference on Computer Vision*, pages 428–444. Springer, 2025.

[6] Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, and Alexander G Hauptmann. Rethinking spatial invariance of convolutional networks for object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19638–19648, 2022.

[7] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[8] Li Dong, Haijun Zhang, Yuzhu Ji, and Yuxin Ding. Crowd counting by using multi-level density-based spatial information: A multi-scale cnn framework. *Information Sciences*, 528:79–91, 2020.

[9] Emilien Dupont, Yee Whye Teh, and Arnaud Doucet. Generative models as distributions of functions. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 2989–3015. PMLR, 2022.

[10] Min Fu, Pei Xu, Xudong Li, Qihe Liu, Mao Ye, and Ce Zhu. Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 43:81–88, 2015.

[11] Guangshuai Gao, Qingjie Liu, and Yunhong Wang. Counting from sky: A large-scale data set for remote sensing object counting and a benchmark method. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):3642–3655, 2020.

[12] Guangshuai Gao, Qingjie Liu, Zhenghui Hu, Lu Li, Qi Wen, and Yunhong Wang. PSGCNet: A pyramidal scale and global context guided network for dense object counting in remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

[13] Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang. Steerer: Resolving scale variations for counting and localization via selective inheritance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21848–21859, 2023.

[14] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017.

[15] Yongbo Huang, Yuanpei Jin, Liqiang Zhang, and Yishu Liu. Remote sensing object counting through regression ensembles and learning to rank. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023.

[16] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018.

[17] Angjoo Kanazawa, Abhishek Sharma, and David Jacobs. Locally scale-invariant convolutional neural networks. *arXiv preprint arXiv:1412.5104*, 2014.

[18] Chiheon Kim, Doyup Lee, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Generalizable implicit neural representations via instance pattern composers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11808–11817, 2023.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Doyup Lee, Chiheon Kim, Minsu Cho, and WOOK SHIN HAN. Locality-aware generalizable implicit neural representation. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in Neural Information Processing Systems*, 23, 2010.

[25] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018.

[26] Zhaoxin Li, Shuhua Lu, Yishan Dong, and Jingyuan Guo. Msffa: a multi-scale feature fusion and attention mechanism network for crowd counting. *The Visual Computer*, 39(3): 1045–1056, 2023.

[27] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022.

[28] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19628–19637, 2022.

[29] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6): 645–654, 2001.

[30] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.

[31] Tony Lindeberg. Scale-covariant and scale-invariant gaussian derivative networks. *Journal of Mathematical Imaging and Vision*, 64(3):223–242, 2022.

[32] Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. Point-query quadtree for crowd counting, localization, and more. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1676–1685, 2023.

[33] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999.

[34] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[35] Xihaier Luo, Xiaoning Qian, and Byung-Jun Yoon. Hierarchical neural operator transformer with learnable frequency-aware loss prior for arbitrary-scale super-resolution. *arXiv preprint arXiv:2405.12202*, 2024.

[36] Zheng Ma, Lei Yu, and Antoni B Chan. Small instance detection by integer programming on object density maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3689–3697, 2015.

[37] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6142–6151, 2019.

[38] Diego Marcos, Benjamin Kellenberger, Sylvain Lobry, and Devis Tuia. Scale equivariance in cnns with vector fields. *arXiv preprint arXiv:1807.11783*, 2018.

[39] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[40] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019.

[41] Joseph Paul Cohen, Genevieve Boucher, Craig A Glastonbury, Henry Z Lo, and Yoshua Bengio. Count-ception: Counting by fully convolutional redundant counting. In *Proceedings of the IEEE International Conference on Computer Vision workshops*, pages 18–26, 2017.

[42] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–285, 2018.

[43] Jihye Ryu and Kwangho Song. Crowd counting and individual localization using pseudo square label. *IEEE Access*, 2024.

[44] Vishwanath A Sindagi and Vishal M Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2019.

[45] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[46] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021.

[47] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*, 2019.

[48] Bernd Sturmfels. *Algorithms in invariant theory*. Springer Science & Business Media, 2008.

[49] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[50] Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1130–1139, 2019.

[51] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1974–1983, 2021.

[52] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1299–1302, 2015.

[53] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 431–441. Springer, 2022.

[54] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2021.

[55] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369*, 2014.

[56] Yilong Yang, Srinandan Dasmahapatra, and Sasan Mahmoodi. Scale-equivariant unet for histopathology image segmentation. *arXiv preprint arXiv:2304.04595*, 2023.

[57] Jun Yi, Zhilong Shen, Fan Chen, Yiheng Zhao, Shan Xiao, and Wei Zhou. A lightweight multiscale feature fusion network for remote sensing object counting. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[58] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6315–6324, 2023.

[59] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016.

[60] Fushun Zhu, Hua Yan, Xinyue Chen, Tong Li, and Zhengyu Zhang. A multi-scale and multi-level feature aggregation network for crowd counting. *Neurocomputing*, 423:46–56, 2021.