Counting Without Scale: Scale-Consistent Error Correction for Crowd Counting

Yi-Kuan Hsieh
Department of Artificial Intelligence
National Yang Ming Chiao Tung University
Tainan, Taiwan (R.O.C.)
yikuan0725@gmail.com

Kuan-Chuan Peng
Mitsubishi Electric Research Labs (MERL)
Cambridge, MA, USA
kpeng@merl.com

Xin Li
Department of Computer Science
State University of New York at Albany
Albany, NY, USA
xin.li@mail.wvu.edu

Ming-Ching Chang
Department of Electrical & Computer Engineering
State University of New York at Albany
Albany, NY, USA
mingching@gmail.com

Chi-Chia Sun
Department of Electronic & Computer Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan (R.O.C.)
chichiasun@ntust.edu.tw

Yu-Chee Tseng
Department of Artificial Intelligence
National Yang Ming Chiao Tung University
Tainan, Taiwan (R.O.C.)
yc0405@gmail.com

Jun-Wei Hsieh*
Department of Artificial Intelligence
National Yang Ming Chiao Tung University
Tainan, Taiwan (R.O.C.)
jwhsieh@nycu.edu.tw

Abstract

Crowd counting is a vital computer vision task with wide-ranging applications, from public surveillance to urban planning. A persistent challenge is achieving scale invariance despite noisy labels, as humans often make counting errors with similar objects. Most existing methods struggle to maintain consistency and robustness when testing images differ in scale from the training data. We propose ScaLe-aware Invariance and Correcting Errors (SLICE) for robust crowd counting, a novel scale-aware loss function that improves scale invariance, enabling consistent and robust predictions under varying scales and annotation noise. The loss function combines two components: a scale-invariance term that applies multiscale supervision to enforce consistency across density maps of the same scene, and an error correction term that addresses inevitable annotation noise by modeling label uncertainty with probabilistic methods such as a mixture of Gaussians. SLICE is model-agnostic and integrates easily into existing crowd counting architectures without extra computational overhead during inference. Extensive experiments on five benchmark datasets show that SLICE greatly improves scale invariance and cross-dataset generalization, achieving the state-of-the-art results in challenging scenarios. Using CrowdDiff, Steerer, and CrowdFormer as baselines, SLICE lowers mean absolute error (MAE) by 5–7% by addressing annotation noise and scale variation.

*Corresponding author: jwhsieh@nycu.edu.tw

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

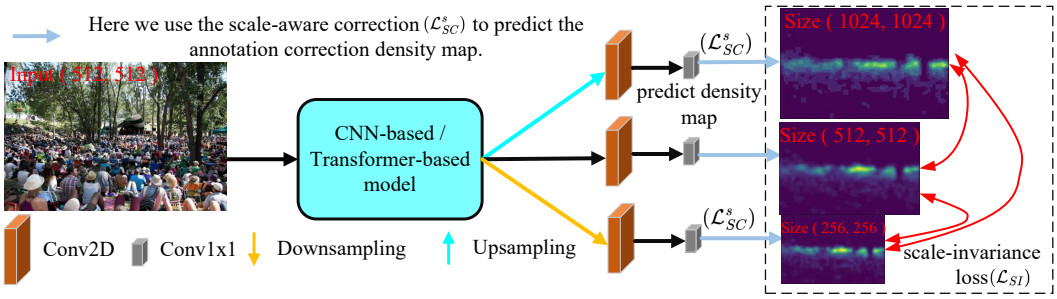


Figure 1: Overview of our proposed crowd counting framework SLICE: A new **scale-invariance loss** promotes consistency across density maps generated at different scales of the same scene. A **scale-aware correction** term models annotation uncertainty using a mixture of Gaussians, ensuring reliable predictions despite resolution changes. The input image of size $W \times H$ is processed by a CNN- or Transformer-based model to generate density maps at multiple scales ($W/2 \times H/2$, $W \times H$, and $2W \times 2H$).

Applying the scale-invariance term across multiple scales further cuts MAE by up to 15%.

1 Introduction

Two fundamental challenges, namely scale variation and annotation noise, limit the generalization and robustness of current crowd counting models. Scale variation occurs when people in the same scene appear at different sizes due to perspective distortion, camera angles, and varying distances [10, 36, 40]. For example, individuals closer to the camera look much larger than those farther away, though each contributes equally to the count. While humans naturally maintain consistent estimates across scales and viewpoints, most AI models struggle to generalize beyond the scales seen in training. When analyzing the same scene at different zoom levels, models often produce inconsistent counts, revealing their sensitivity to scale. This highlights the need for scale-invariant models that perform consistently across diverse real-world settings.

Annotation noise presents another major challenge [33]. Large datasets frequently contain inaccurate or inconsistent labels. This is especially true in dense crowds, where manual annotation is error-prone and automated methods, such as L2-norm [17, 32, 2] or Bayesian Loss [21], can be unreliable. This is partly because humans find counting many similar objects difficult, often losing track of which ones have been counted. Such noisy labels can mislead models during training, reducing accuracy. To address annotation noise, [33] proposed a training framework that models noise probabilistically and adjusts the loss function. While effective, their method focuses on fixed-scale training and does not solve the problem of scale invariance.

To address the aforesaid issues, we propose the ScaLe-aware Invariance and Correcting Errors (SLICE), a scale-aware loss function for robust crowd counting. SLICE is motivated by two key ideas. Inspired by self-supervised crowd counting methods [19, 38, 1], we enforce scale invariance as a self-supervision constraint, requiring consistent counting results across multiscale representations of the same scene. A similar insight appears in AutoScale [40], where the counting module learns to scale dense regions based on camera distance. Unlike AutoScale, we use resampling-based data augmentation to enforce cross-scale consistency. To handle spatial variation in density maps and annotation noise, we propose a scale-aware error correction module. This module uses probabilistic models like mixtures of Gaussians to explicitly capture uncertainty from scale variations and manual annotations for error correction. We hypothesize that scale invariance and error correction can complement each other in a self-supervised way.

SLICE improves crowd counting models' robustness to both scale variation and annotation noise. SLICE encourages scale-invariant counting by using predicted heatmaps at multiple resolutions for each training image. We introduce a **scale-invariance loss** term to promote consistency across density maps generated from different scales of the same scene [10]. Since accurately locating each person's head center is challenging, especially for small or distant subjects, annotation noise is unavoidable. To address this, we introduce a **scale-aware correction term** that models annotation

errors by incorporating uncertainty through probabilistic models like mixtures of Gaussians [9]. With these two terms, SLICE is model-agnostic, needs no architecture change, and integrates easily with existing CNN-, transformer-, and diffusion-based crowd-counting models.

Extensive experiments on multiple benchmark datasets show that SLICE significantly improves performance under varying levels of scale variation and annotation noise. These results validate the efficacy of our scale-invariant training strategy and highlight its potential to make crowd counting more reliable in real-world, unconstrained environments. Our main contributions are summarized below:

- **Scale-aware learning framework**: We propose SLICE, a scale-aware loss function that makes crowd counting models more robust to scale variation and annotation noise. This encourages the model to learn consistent counting patterns across scales, improving generalization without relying on scale-specific assumptions.
- **Generalization across scales without architectural changes**: Unlike previous methods that depend on specific architectures like multi-column networks or scale-aware attention, SLICE is architecture-agnostic and easily integrates with existing CNN-based or generative models. This makes it practical and scalable.
- **Robust supervision via scale-consistent loss**: To handle real-world label noise, we adopt a scale-consistent supervision strategy inspired by prior work on noise modeling. Combining scale invariance with error correction helps reduce the impact of noisy labels by jointly supervising predictions at multiple resolutions.
- **Consistent performance gain across models and datasets**: We conduct extensive experiments using CrowdDiff, Steerer, and CrowdFormer as baselines on five benchmark datasets. Table 2 shows that SLICE reduces mean absolute error (MAE) by 5–7% by addressing annotation errors and scale variation. Applying the scale-invariance term across multiple scales further reduces MAE by up to 15%.

2 Our proposed method – SLICE

2.1 Crowd counting density map

Traditional crowd counting methods often treat the task as a density regression problem [33, 16, 30]. Given an input image $\mathcal{I}$ with N people, let $\mathbf{H}_i$ be the true head location of the i-th person. At any pixel location x in $\mathcal{I}$, the crowd density $\mathbf{D}(x)$ is computed as the sum of Gaussian kernels centered at the annotated head position.

During training, a model learns to predict the density map $\mathbf{D}$ from the input image $\mathcal{I}$ using a regression function $f(\mathcal{I})$. The learning objective is typically defined using the mean squared error (MSE) loss, $\mathcal{L}(\mathbf{D}, f(\mathcal{I})) = |\mathbf{D} - f(\mathcal{I})|^2$, or a Bayesian loss [21]. The final crowd count is estimated by summing the predicted density values over all pixels in the image. However, prior works [16, 30, 33] use a fixed variance β for all instances, disregarding the fact that head sizes vary with perspective. To overcome this, we propose a scale-aware formulation using a mixture of S Gaussian kernels with variances $\{\beta_s\}_{s=1}^{S}$. The revised density is:

$$\mathbf{D}(x) = \frac{1}{S}\sum_{i=1}^{N}\sum_{s=1}^{S}\mathcal{N}(x \mid \mathbf{H}_i, \beta_s\mathbf{I}) \quad = \frac{1}{S}\sum_{s=1}^{S}\mathbf{D}_s(x), \tag{1}$$

where $\mathbf{D}_s(x) = \sum_{i=1}^{N}\mathcal{N}(x \mid \mathbf{H}_i, \beta_s\mathbf{I})$ is the i-th Gaussian kernel at scale s. Given a predicted density map $\mathbf{D}_s(x)$ for image $\mathcal{I}_s$, the total count N_s is computed as $N_s = \sum_{x_j \in \mathcal{I}_s}\mathbf{D}_s(x_j)$.

2.2 Crowd counting scale-invariance loss

Previous crowd counting studies typically assume a fixed input resolution, such as $W \times H$. Consequently, prior models and loss functions are generally designed to operate effectively only at this single scale. When applied to images with other resolutions, their performance often degrades significantly. To adapt to multiple scales, prior works typically require retraining on more images of varying sizes, which substantially increases training time and does not guarantee consistent accuracy.

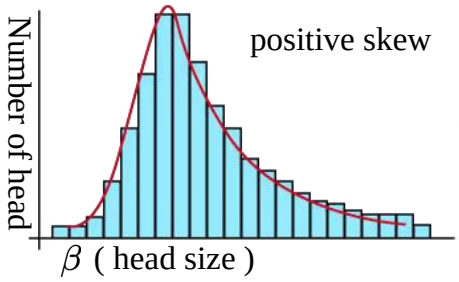
(a)  (b) 

Figure 2: Modeling uncertainty for crowd counting: (a) Inaccurate annotations lead to biased mean (red dots deviating from the human face centers). (b) Varying camera distance yields a positively skewed head-size distribution β, reflecting variance change.

To address this limitation, we propose a novel **Scale-Invariance Loss** $\mathcal{L}_{SI}$ specifically designed to enhance the model's robustness to scale changes without requiring additional multi-scale training. As illustrated in Fig. 1, our method SLICE generates density maps at multiple resolutions ($\frac{W}{2} \times \frac{H}{2}$, $W \times H$, and $2W \times 2H$) from a single input image. The red arrows indicate the mutual supervision mechanism between different scales, ensuring consistency across them. $\mathcal{L}_{SI}$ is defined as:

$$\mathcal{L}_{SI} = \|N_{s_1} - N_{s_2}\| + \|N_{s_2} - N_{s_3}\| + \|N_{s_3} - N_{s_1}\|, \tag{2}$$

where N_{s_i} for $i = 1, 2, 3$ denote the predicted number of people, respectively. The definition of N_s is provided in Eq. (5). From a practical standpoint, no single resolution should be given priority over another. The counting task is inherently invariant to the scale of observation, including arbitrary resolutions (e.g., fractional sampling rates), and each scale provides equally informative signals. Accordingly, the training objective should enforce this principle by assigning equal weight to all terms in Eq. 5. A key advantage of SLICE is that it requires training at only one resolution ($W \times H$) and produces a prediction heatmap of the same resolution during inference. The proposed scale-invariance term naturally enforces cross-scale consistency, enabling the model to achieve high accuracy without relying on additional multi-scale data augmentation or retraining.

2.3 Scale-aware modeling of noisy annotations

We address the uncertainty in manual annotations of human head positions, as illustrated in Fig. 2(a). Such point-wise annotation inaccuracies introduce significant noise into the training data for image-based crowd density estimation. In [9], a scale-aware annotation error model was proposed. It represents head density in single-scale images using a mixture of Gaussian distributions, requiring estimation of the mixing weights w_s. Inspired by [9], we propose a novel annotation error model that forgoes the need to estimate mixing weights. Instead, our method SLICE constructs a scale-aware error correction term directly from each scale image $\mathcal{I}_s$, thereby simplifying the model and reducing parameter dependence. This streamlined formulation improves both correction efficiency and stability across scales. Throughout the remainder of this paper, all symbolic notation aligns with that used in [9].

Let $\tilde{\mathbf{H}}_i$ denote the annotated head position of the i-th person with potential error, and ε_i denote the corresponding noise, such that $\tilde{\mathbf{H}}_i = \mathbf{H}_i + \varepsilon_i$. We assume $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \alpha\mathbf{I})$, where α denotes the annotation noise variance. Given a Gaussian kernel with variance β_s placed at each head location, we define $q_i = x - \mathbf{H}_i$, the offset from pixel x to the true head position. When annotation noise is considered, the density at pixel x is modeled as:

$$\begin{aligned} \mathbb{D}(x) &= \frac{1}{S}\sum_{i=1}^{N}\sum_{s=1}^{S}\mathcal{N}(x \mid \tilde{\mathbf{H}}_i, \beta_s\mathbf{I}) \\ &= \frac{1}{S}\sum_{i=1}^{N}\sum_{s=1}^{S}\mathcal{N}(q_i \mid \varepsilon_i, \beta_s\mathbf{I}). \end{aligned} \tag{3}$$

Let $\phi_i^s(x) = \mathcal{N}(q_i \mid \varepsilon_i, \beta_s \mathbf{I})$ which denotes the i-th Gaussian kernel at the i-th location and at scale s. Let $\mathbb{D}_s(x) = \sum_{i=1}^{N} \phi_i^s(x)$, the full density is:

$$\mathbb{D}(x) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{D}_s(x). \tag{4}$$

Let $\mathcal{I}_s$ be the image domain rescaled for scale s, with J_s pixels. For each pixel $x_j \in \mathcal{I}_s$, the scale-specific density vector is $\mathbb{D}_s = [\mathbb{D}_s(x_1), \ldots, \mathbb{D}_s(x_{J_s})]$. The total count N_s at scale s is computed as:

$$N_s = \sum_{x_j \in \mathcal{I}_s} \mathbb{D}_s(x_j). \tag{5}$$

We then approximate the random variable $\mathbb{D}_s$ using a multivariate Gaussian $\hat{p}(\mathbb{D}_s) \sim \mathcal{N}(\mu_s, \Sigma_s)$ with mean:

$$\begin{aligned} \mu_s = \mathbb{E}[\mathbb{D}_s] &= \sum_{i=1}^{N} \mathbb{E}[\mathcal{N}(q_i \mid \varepsilon_i, \beta_s \mathbf{I})] \\ &= \sum_{i=1}^{N} \mathcal{N}(q_i \mid 0, (\alpha + \beta_s)\mathbf{I}) = \sum_{i=1}^{N} \mu_i^s, \end{aligned} \tag{6}$$

where $\mu_i^s = \mathcal{N}(q_i \mid 0, (\alpha + \beta_s)\mathbf{I})$. The variance is:

$$\begin{aligned} \Sigma_s^2 = \mathrm{Var}(\mathbb{D}_s) &= \mathbb{E}[\mathbb{D}_s^2] - \mathbb{E}[\mathbb{D}_s]^2, \\ &\approx \sum_{i=1}^{N} \left[\frac{1}{4\pi\beta_s} \mathcal{N}(q_i \mid 0, (\beta_s/2 + \alpha)\mathbf{I}) - (\mu_i^s)^2 \right]. \end{aligned} \tag{7}$$

We derive the full covariance matrix Σ_s between pixels x_j and x_k. Since $\mathbb{D}_s(x) = \sum_{i=1}^{N} \phi_i^s(x)$, the diagonal of the scale-aware convairance matrix is:

$$\begin{aligned} \Sigma_{x_j, x_k}^s &= \mathrm{Cov}(\mathbb{D}_s(x_j), \mathbb{D}_s(x_k)) \\ &= \sum_{i=1}^{N} \left[\Omega_i^s(x_j, x_k) - \mu_i^s(x_j) \mu_i^s(x_k) \right], \end{aligned} \tag{8}$$

where $\Omega_i^s(x_j, x_k) = \mathbb{E}[\phi_i^s(x_j)\phi_i^s(x_k)]$. This scale-aware probabilistic framework offers a principled way to model the expected density with uncertainty and spatial correlation arising from annotation noise and object scale.

2.4 Scale-invariant correction loss function

To ensure that the predicted density map near each annotation satisfies the condition of density values summing to 1, we define the regularizer $\mathcal{R}_i^s$ for the i-th annotation point as:

$$\mathcal{R}_i^s = \left| \sum_j \mathbb{D}_s(x_j) \frac{\phi_i^s(x_j)}{\sum_{i=1}^{N} \phi_i^s(x_j)} - 1 \right|, \tag{9}$$

where $\mathbb{D}_s(x_j)$ is the j-th term of $\mathbb{D}_s$. Let $\bar{\mathbb{D}}_s = \mathbb{D}_s - \mu_s$. The Scale-aware Correction loss function ($\mathcal{L}_{SC}$) is:

$$\mathcal{L}_{SC}^s = \bar{\mathbb{D}}_s^T (\hat{\Sigma}^s)^{-1} \bar{\mathbb{D}}_s + \sum_{i=1}^{N} \mathcal{R}_i^s, \tag{10}$$

where $\hat{\Sigma}^s$ is the low-rank approximation of Eq. (8) and the second term is the accumulation of regularization terms from Eq. (9). In the previous section, we explored the design of scale-aware correction under annotation errors; however, no adjustments were made for scale invariance. A straightforward approach to addressing this issue, as illustrated in Fig. 1, is to leverage the fact that scale-aware estimation inherently produces density maps at different scales. By enforcing mutual supervision between these scale-specific density maps, we ensure that their estimated counts remain consistent across different scales. This design effectively mitigates the issue where variations in

Table 1: MAE and MSE (lower is better) increase significantly when tested at different resolutions compared to the original for CrowdDiff, Steerer, and CrowdFormer on five datasets. The red values mark the best results.

Model	Sizes	QNRF		SHHA		SHHB		NWPU		JHU-CROWD++	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CrowdDiff [25]	(H, W)	68.9	125.6	47.4	75.0	5.7	8.2	57.8	221.2	47.3	198.9
	(2H, 2W)	75.8	146.7	59.2	89.3	6.5	9.2	63.1	255.4	58.3	221.9
	(H/2, W/2)	80.4	159.8	56.3	86.1	6.7	9.4	61.5	245.1	57.6	218.7
Steerer [8]	(H, W)	74.3	128.3	54.5	86.9	5.8	8.5	63.7	309.8	54.3	238.3
	(2H, 2W)	87.6	154.4	62.3	98.4	6.9	9.7	70.2	339.1	67.1	266.4
	(H/2, W/2)	85.2	148.9	64.7	99.5	6.4	9.1	74.3	354.8	65.7	260.9
CrowdFormer [42]	(H, W)	78.8	136.1	56.9	97.4	5.7	9.6	67.1	301.6	59.8	274.2
	(2H, 2W)	91.7	189.8	68.4	113.8	6.8	11.3	74.2	340.5	69.1	310.8
	(H/2, W/2)	90.3	180.1	70.3	115.6	6.7	10.8	76.7	355.3	67.8	307.4

resolution lead to discrepancies in estimated crowd counts. At the beginning, N_s is estimated from $\mathcal{I}_s$ based on $\mathcal{L}_{SC}^s$. The final result N is fine-tuned by optimizing the final loss function:

$$\mathcal{L} = \lambda \mathcal{L}_{SI} + \sum_{s=1}^{S} \mathcal{L}_{SC}^s, \tag{11}$$

where $\mathcal{L}_{SI}$ is defined in Eq. (2) and λ controls the tradeoff between scale-invariance constraint and error correction consistency enforcement. The term $\mathcal{L}_{SI}$ allows density maps of different scales to supervise each other and serve as a supervisory and check-and-balance function. Ablation studies on λ are reported in the supplementary material.

3 Experimental results

We evaluated the effectiveness of the proposed scale-invariant loss by applying it to several State-of-the-Art (SoTA) crowd counting models based on CNNs, diffusion models, and Transformers. These three representative models—CrowdDiff [25], Steer [8], and CrowdFormer [42]—corresponding to diffusion-based, CNN-based, and Transformer-based SoTA crowd counting methods, respectively, are selected in this paper for performance evaluation and comparison. These models represent the current SoTA in their respective categories. We did not include comparisons to CLIP-based models such as CLIP-EBC, because they leverage CLIP's text encoder and thus deviate from our image-only setting; our definition of scale invariance is designed for pure visual models rather than language-supervised hybrids. For each model, we conducted a comparative analysis between variants trained with and without our loss function to assess its impact on performance across different input scales. Experiments were performed on six public benchmark datasets: UCF-QNRF [13], UCF CC 50 [12], NWPU-Crowd [37], ShanghaiTech Part A [43], ShanghaiTech Part B [43], and JHU-CROWD++ [31]. The results show that our method SLICE is model-agnostic and that incorporating our loss consistently enhances scale robustness and improves counting accuracy across diverse model architectures. All training stage parameters are listed in Appendix Table 5. All training is performed on a machine with NVIDIA RTX 4090 GPU.

3.1 Model training parameters

SLICE was pre-trained on ImageNet [6] using the Adam optimizer. Given the varying image dimensions in the datasets, we crop patches of a fixed size at random locations and then augment the data by random horizontal flipping with a probability of 0.5. We specify the learning rates used during training for each dataset in Table 5. To stabilize the training loss change, we use batch sizes of 12. The learning rates and batch sizes were selected according to the complexity of each dataset. For more complex datasets, larger updates to the model's weights are needed to ensure convergence during training. In contrast, simpler datasets require smaller learning rates to avoid overshooting the optimal point during gradient descent. For the ShanghaiTech and JHU datasets, higher learning rates

Table 2: Comparison of MAE and MSE (the lower the better) under different input resolutions, with and without the proposed Scale-Invariance (SI) and Scale-aware Correction (SC) terms. The colored values mark the best results in each case.

Models	Rows	SI	SC	Sizes	QNRF		SHHA		SHHB		NWPU		JHU-CROWD++		Improv. Ratio	
					MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CrowdDiff [25]	1	✗	✗	(H, W)	68.9	125.6	47.4	75.0	5.7	8.2	57.8	221.2	47.3	198.9	–	–
	2			(H/2, W/2)	75.8	146.7	59.2	89.3	6.5	9.2	63.1	255.4	58.3	221.9	–	–
	3			(2H, 2W)	80.4	159.8	56.3	86.1	6.7	9.4	61.5	245.1	57.6	218.7	–	–
	4	✗	✓	(H, W)	67.3	106.5	45.7	73.1	5.5	7.6	55.7	204.9	46.1	184.7	+3.1 %	+7.9 %
	5			(H/2, W/2)	73.1	139.6	57.3	86.6	6.3	8.7	61.4	243.8	57.5	218.1	+2.7 %	+3.9 %
	6			(2H, 2W)	73.8	143.6	57.5	88.1	6.2	8.5	59.8	231.2	56.4	215.8	+3.6 %	+4.8 %
	7	✓	✗	(H, W)	68.0	119.3	46.8	74.1	5.4	6.9	55.9	209.7	45.1	176.3	+3.1 %	+7.7 %
	8			(H/2, W/2)	68.9	123.2	46.3	72.9	5.7	7.3	54.9	197.6	45.3	178.1	+15.6 %	+19.7 %
	9			(2H, 2W)	68.7	122.4	46.5	73.8	5.6	7.1	56.7	210.5	45.2	177.1	+15.5 %	+19.6 %
	10	✓	✓	(H, W)	65.7	98.4	43.4	65.3	5.1	6.5	52.9	173.5	43.6	171.8	+7.9 %	+18.1 %
	11			(H/2, W/2)	65.5	97.8	43.9	66.2	5.2	6.6	53.1	177.4	43.1	169.4	+20.2 %	+28.3 %
	12			(2H, 2W)	65.9	99.1	44.2	66.7	5.1	6.6	52.7	170.2	43.4	170.8	+20.4 %	+28.5 %
Steerer [8]	13	✗	✗	(H, W)	74.3	128.3	54.5	86.9	5.8	8.5	63.7	309.8	54.3	238.3	–	–
	14			(H/2, W/2)	87.6	154.4	62.3	98.4	6.9	9.7	70.2	339.1	67.1	266.4	–	–
	15			(2H, 2W)	85.2	148.9	64.7	99.5	6.4	9.1	74.3	354.8	65.7	260.9	–	–
	16	✗	✓	(H, W)	72.5	116.3	52.7	83.3	5.5	8.1	61.2	278.9	52.9	217.7	+3.4 %	+7.3 %
	17			(H/2, W/2)	84.3	142.1	60.3	94.8	6.3	8.8	69.3	308.5	65.9	248.1	+3.7 %	+7.3 %
	18			(2H, 2W)	83.6	137.9	61.5	96.1	6.5	9.1	71.7	319.4	63.3	235.7	+2.4 %	+6.0 %
	19	✓	✗	(H, W)	74.1	125.7	54.0	84.3	5.6	8.0	62.9	301.4	54.2	236.9	+1.2 %	+2.8 %
	20			(H/2, W/2)	74.3	128.6	54.1	84.7	5.5	7.8	63.1	303.7	54.4	238.5	+15.5 %	+14.2 %
	21			(2H, 2W)	74.2	127.4	53.9	84.1	5.7	8.2	63.8	301.6	55.1	244.5	+14.1 %	+12.2 %
	22	✓	✓	(H, W)	71.5	109.7	51.3	81.7	5.3	7.5	59.8	271.3	52.7	213.4	+5.4 %	+11.0 %
	23			(H/2, W/2)	71.3	108.4	51.1	81.2	5.4	7.6	60.3	274.6	53.1	215.3	+18.6 %	+21.4 %
	24			(2H, 2W)	71.6	110.9	51.5	82.1	5.3	7.6	60.7	277.9	53.4	218.6	+18.1 %	+19.4 %
CrowdFormer [42]	25	✗	✗	(H, W)	78.8	136.1	56.9	97.4	5.7	9.6	67.1	301.6	59.8	274.2	–	–
	26			(2H, 2W)	91.7	189.8	68.4	113.8	6.8	11.3	74.2	340.5	69.1	310.8	–	–
	27			(H/2, W/2)	90.3	180.1	70.3	115.6	6.7	10.8	76.7	355.3	67.8	307.4	–	–
	28	✗	✓	(H, W)	75.3	126.4	54.7	91.5	5.4	9.0	64.9	288.3	56.6	261.7	+4.4 %	+5.6 %
	29			(H/2, W/2)	87.4	172.9	65.9	105.7	6.4	10.4	71.3	318.4	65.3	297.6	+4.7 %	+6.9 %
	30			(2H, 2W)	86.2	167.7	67.1	108.3	6.2	10.1	74.2	329.5	62.7	288.4	+5.4 %	+6.6 %
	31	✓	✗	(H, W)	78.1	132.7	56.2	95.9	5.6	9.5	66.7	297.6	59.3	271.4	+1.0 %	+1.4 %
	32			(H/2, W/2)	78.4	135.8	56.9	98.2	5.8	9.7	67.0	299.4	59.7	274.1	+13.8 %	+16.0 %
	33			(2H, 2W)	78.5	136.4	56.6	97.5	5.7	9.6	67.2	301.3	60.1	278.7	+14.2 %	+15.1 %
	34	✓	✓	(H, W)	74.2	117.7	53.5	82.8	5.2	7.7	62.3	276.9	54.4	250.8	+7.3 %	+13.0 %
	35			(H/2, W/2)	74.4	119.1	53.4	82.4	5.4	8.0	62.6	279.2	54.3	249.9	+19.6 %	+26.3 %
	36			(2H, 2W)	74.7	122.1	53.8	83.1	5.2	7.8	62.4	278.1	54.7	253.6	+20.2 %	+25.4 %

were chosen due to their greater complexities and diversities. Similar to other SoTA methods [8, 25], we use mean absolute error (MAE) and mean squared error (MSE) as evaluation metrics.

3.2 Analyses of scale-invariance and error correction terms in crowd counting models

To investigate the impact of input resolution on crowd counting performance, we evaluate several models under three different input scales: the original resolution (H, W), an up-sampled resolution $(2H, 2W)$, and a down-sampled resolution $(H/2, W/2)$. Table 1 shows that most existing crowd counting models and loss functions are developed and evaluated solely under a fixed input size, typically 512×512. However, we argue that an ideal AI model should exhibit *scale-invariance*—that is, it should maintain consistent counting performance regardless of input scale—mimicking the human visual system, which can perceive quantities consistently across different levels of zoom. Our experimental results confirm this hypothesis: current models, including CrowdDiff [25], Steerer [8], and CrowdFormer [42], show significant performance degradation when evaluated on either upscale or downscale versions of the input image, compared to the original resolution. For example, for the QNRF dataset, CrowdDiff increases MSE from 125.6 to 146.7 and 159.8 for the up-sampled resolution $(2H, 2W)$ and the down-sampled resolution $(H/2, W/2)$, respectively. Similarly, for the same dataset, Steerer increases MSE from 128.3 to 154.4 and 148.9 for the up-sampled resolution $(2H, 2W)$ and the down-sampled resolution $(H/2, W/2)$, respectively.

To further validate the efficacy of the proposed scale-invariance term, we conduct a comprehensive analysis across multiple crowd counting models under different input resolutions, as shown in Table 2. Specifically, we compare the baseline models (CrowdDiff [25], Steerer [8], CrowdFormer[42]) with their corresponding variants trained with our scale-invariance loss. Each model is evaluated at three input resolutions: the original resolution (H, W), a down-sampled version $(H/2, W/2)$, and an

Table 3: Evaluation of the proposed Scale-Invariance and Scale-aware Correction terms added to different loss functions, tested at various input resolutions on five benchmark datasets using the Steerer [8] model. The results show how each term improves MAE and MSE across scales, highlighting the robustness and generalization of our method.

Loss	SI	SC	Input Res.	QNRF		SHHA		SHHB		NWPU		JHU-CROWD++	
				MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
BL[21]	✗	✗	(H,W)	85.8	150.6	61.9	99.6	8.3	13.1	66.3	321.8	67.7	258.5
			(H/2,W/2)	99.6	198.4	74.3	114.5	9.6	15.4	76.5	348.9	74.6	289.4
			(2H,2W)	98.2	192.7	73.7	109.5	9.8	16.9	75.2	341.4	76.2	293.1
	✗	✓	(H,W)	82.4	136.9	57.8	89.4	7.8	12.2	64.7	302.7	63.6	233.8
			(H/2,W/2)	95.1	161.5	68.5	104.5	8.9	13.7	74.6	325.1	72.4	258.7
			(2H,2W)	93.7	159.8	69.9	106.4	9.2	14.6	73.3	317.6	74.8	266.1
	✓	✗	(H,W)	84.7	145.9	57.8	92.4	7.8	11.2	64.9	314.5	65.6	233.8
			(H/2,W/2)	85.1	151.5	58.1	94.5	7.9	11.7	65.3	317.7	66.4	248.7
			(2H,2W)	83.5	143.8	59.9	96.4	8.1	12.1	65.1	321.6	65.8	236.1
	✓	✓	(H,W)	80.7	132.4	56.3	86.6	6.2	10.3	62.6	187.9	57.2	224.5
			(H/2,W/2)	81.1	133.7	56.6	87.4	6.4	10.5	62.9	192.2	57.7	229.6
			(2H,2W)	81.2	134.1	56.6	87.0	6.6	10.7	63.2	197.8	57.5	227.8
NoiseCC[33]	✗	✗	(H,W)	78.6	154.8	59.1	96.9	6.4	9.1	69.2	330.7	56.4	253.8
			(H/2,W/2)	89.2	184.7	71.3	134.5	7.3	10.9	83.5	369.1	68.3	297.6
			(2H,2W)	86.2	176.7	70.7	128.5	7.1	9.9	84.3	378.4	66.5	280.7
	✗	✓	(H,W)	74.7	135.1	57.9	90.4	6.2	8.8	67.3	318.1	54.7	238.3
			(H/2,W/2)	85.1	167.5	68.8	124.6	6.9	9.4	78.4	352.6	65.3	279.1
			(2H,2W)	86.4	169.5	67.5	121.8	6.8	9.3	77.4	347.5	64.1	265.8
	✓	✗	(H,W)	76.3	150.7	57.5	94.2	6.2	8.8	67.7	321.4	55.3	242.9
			(H/2,W/2)	77.2	155.4	58.1	94.9	6.4	9.1	67.9	323.7	55.9	245.4
			(2H,2W)	76.8	153.1	57.3	93.8	6.3	8.9	68.1	323.9	56.2	249.6
	✓	✓	(H,W)	73.1	128.8	55.3	84.4	5.9	8.5	65.2	305.7	53.2	224.6
			(H/2,W/2)	73.4	130.5	55.6	84.9	6.1	8.7	65.7	308.4	53.9	228.1
			(2H,2W)	73.8	131.4	55.7	85.1	6.2	8.7	65.4	306.2	54.6	223.5

BL = Bayesian loss SI = Scale-Invariance SC = Scale-aware Correction

Table 4: Comparison of MAE and MSE (the lower the better) between our proposed **Scale-invariant Correction Loss (SCL)** in our method SLICE against the **Scale-Aware Loss (SAL)** from [9] across various input resolutions on five benchmark datasets.

Model	Input Resolution	QNRF		SHHA		SHHB		NWPU		JHU-CROWD++	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Scale-Aware Loss (SAL) [9]	(H, W)	73.9	121.7	52.1	76.6	6.1	9.7	70.0	211.4	53.6	201.5
	(2H, 2W)	83.8	146.7	59.2	89.3	6.7	10.2	80.1	255.4	58.3	221.9
	(H/2, W/2)	84.4	159.8	58.3	86.1	6.7	10.4	79.5	245.1	57.6	218.7
Scale-invariant Correction Loss (SCL)	(H, W)	71.5	109.7	51.3	71.7	5.3	7.5	59.8	181.3	52.7	190.4
	(2H, 2W)	71.3	108.4	51.1	71.2	5.4	7.6	60.3	185.4	52.6	189.6
	(H/2, W/2)	71.6	110.9	51.5	72.1	5.3	7.6	60.7	189.7	52.8	191.1

up-sampled version $(2H, 2W)$. As shown in Rows 1-3, Rows 13-15, and Rows 25-27 of Table 2, CrowdDiff [25], Steerer [8], and CrowdFormer [42] exhibit noticeable degradation in MAE and MSE when evaluated on inputs with doubled or halved spatial dimensions. After introducing our scale-invariance (SI) term into the training process, we observe significant improvements in MAE and MSE across different datasets (see Rows 7-9, Rows 19-21, and Rows 31-33). For example, the CrowdDiff model optimized with the SI loss function achieves highly consistent results across all resolutions (Rows 7-9) with minimal variance in both MAE and MSE across datasets. Similarly, for the Steerer [8] and CrowdFormer [42] models, the scale-invariance term also demonstrates stable performance improvements.

In real-world scenarios, most training datasets contain annotation errors in head locations. To address this inherent annotation noise, we propose an error correction term in addition to the SI term. For the CrowdDiff method [25], comparing Rows 4–6 to Rows 1–3 reveals substantial improvements in both MAE and MSE across various datasets and resolutions when the Scale-aware Correction (SC) term is

applied. Similarly, Rows 16–18 show significant performance gains for the Steerer method [8] under the same conditions. The CrowdFormer model [42] also shows comparable improvements when the SC term is incorporated.

Further accuracy enhancements can be achieved by combining both the SI and SC terms. For the CrowdDiff model, individual application of the SI and SC terms yields a 3.1% improvement in MAE (see Rows 4 and 7). However, when both terms are jointly optimized, the MAE improvement increases to 7.9% (see Row 10). The gains are even more pronounced in terms of MSE, where joint integration leads to an 18.1% improvement compared to the individual contributions (Rows 4 and 7 vs. Row 10). Similar trends are observed for the Steerer and CrowdFormer models, with consistent MAE and MSE improvements across different datasets and resolutions when both terms are integrated.

These findings strongly support our hypothesis that explicitly incorporating scale consistency and error correction during training significantly enhances the robustness of crowd counting models. By aligning predictions across varying resolutions, the models become less sensitive to changes in input size—an essential characteristic for real-world deployment, where image resolutions can be highly variable. Furthermore, the experiments confirm that SLICE is model-agnostic and can be seamlessly integrated into existing crowd counting frameworks without introducing any additional inference-time computational overhead.

3.3 Parameter setup for the annotation variance β_s

As suggested in [9], given an initial β_s, subsequent values can be computed recursively using $\beta_{s+1} = \beta_s/2$, based on the pooling operation that halves the feature map size and thus effectively reduces the head size in the feature space. To better understand the influence of β_s on model performance, we treat β_1 as a tunable hyperparameter. In Table 2, CrowdDiff [25] achieves the lowest MAE and MSE, making it the baseline for evaluating the effect of β_1. Our supplement contains the results of the ablation study on β_1. As β_1 increases from 2 to 10, both the MAE and MSE metrics decrease, reaching their minimum at $\beta_1 = 8$ in all data sets (see Table 1 in the supplement). This suggests that $\beta_1 = 8$ yields optimal performance. However, further increases beyond 8 result in rising MAE and MSE values, indicating that excessively large β_1 values may degrade accuracy.

3.4 Impact of Scale-Invariance Term on Different Loss Functions.

To evaluate the generalization of the proposed Scale-Invariance (SI) and Scale-aware Correction (SC) terms, we integrate them into two widely used loss functions: Bayesian loss [21] and NoiseCC loss [33], and assess their performance across multiple input resolutions and benchmark datasets using the Steerer [8] model. As shown in Table 3, adding the SI term consistently improves MAE and MSE compared to the original losses across different scales. In particular, incorporating both SI and SC into the Bayesian loss significantly boosts accuracy on both downsampled and upsampled inputs, showing strong robustness to scale variation. Similarly, for NoiseCC, the SI and SC terms reduce the performance drop typically seen at non-original resolutions. These results confirm the effectiveness and generalization of our scale-aware design across different loss functions in crowd counting.

3.5 Ablation study on annotation-error loss functions

In [9], a Gaussian Mixture Model (GMM) is employed to model annotation noise, leading to the development of a scale-aware loss function for crowd counting. In contrast, SLICE uses a scale-invariant correction loss that enhances crowd counting performance without relying on GMM.

Let N denote the estimated number of heads in the original image $\mathcal{I}$, and N_s represent the estimated count in the downscaled version of $\mathcal{I}$ at scale s. The fundamental distinction between the scale-aware loss (SAL) function in [9] and SLICE lies in their respective constraints: the former enforces $N = \sum_{s=1}^{S} w_s N_s$, where w_s are weights satisfying $\sum_{s=1}^{S} w_s = 1$, whereas SLICE requires $N = N_s$ for all s. Table 4 presents a comparative analysis of our proposed **Scale-invariant Correction Loss (SCL)** against the **Scale-Aware Loss (SAL)** from [9] across various input resolutions on five benchmark datasets. The evaluation metrics include Mean Absolute Error (MAE) and Mean Squared Error (MSE), with lower values indicating superior counting accuracy. SLICE consistently achieves the lowest MAE at the original resolution (H, W) and significantly reduces errors across other

resolutions. These findings underscore the limitations of existing methods like SAL in handling scale variations, while SLICE effectively addresses this challenge.

4 Conclusion

We propose ScaLe-aware Invariance and Correcting Errors (SLICE) for robust crowd counting, a novel scale-invariant loss function that addresses the challenge of scale variation without needing fixed input resolutions or multiple models. SLICE uses a scale-consistency term, enabling networks to learn robust, scale-adaptive representations. This loss integrates seamlessly with various architectures and existing objectives, consistently improving performance on standard benchmarks. Extensive experiments across diverse datasets confirm that SLICE mitigates scale-induced performance drops with minimal computational overhead. These results underscore the value of scale-invariant training for enhancing generalization and robustness, paving the way for future advances in scale-adaptive dense prediction.

Limitations SLICE improves scale robustness without adding inference-time overhead, but it requires multi-scale predictions during training and therefore increases training cost. Its performance also depends on scale-related hyperparameters, such as the annotation noise variance and Gaussian kernel variance, which may require tuning under extreme perspective changes or very small head sizes. Moreover, our experiments focus on image-based crowd counting; extending the framework to video, generic object counting, and multimodal settings is left for future work.

Future works include exploring automatic parameter selection for α and β_s through data-driven learning. A truly scale-invariant crowd counting framework will be agostic to the backbone architecture (e.g., via transformer or diffusion models). It is also worth investigating the extension of this work from crowd counting to automatic counting of generic objects (e.g., fruits and vehicles) and from still image to video-based dynamic counting (e.g., cell growth rate and traffic pattern estimation).

Acknowledgments and Disclosure of Funding

Kuan-Chuan Peng was exclusively supported by MERL. We thank Xin Li and Ming-Ching Chang for their collaboration. The other authors gratefully acknowledge the assistance provided by the Advanced Technology Diversified Integration Development Project for Unmanned Vehicles at NCSIST.

References

[1] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A Sindagi, R Venkatesh Babu, and Vishal M Patel. Completely self-supervised crowd counting via distribution matching. In *European Conference on Computer Vision*, pages 186–204. Springer, 2022.

[2] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[3] Jian Cheng, Haipeng Xiong, Zhiguo Cao, and Hao Lu. Decoupled two-stage crowd counting and beyond. *IEEE Transactions on Image Processing*, 30:2862–2875, 2021.

[4] Jing-an Cheng, Wenzhe Zhai, Qilei Li, and Mingliang Gao. Crowd counting via lightweight neural networks: a literature review. *Measurement Science and Technology*, 36(7):072001, jul 2025.

[5] 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.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Junyu Gao, Qi Wang, and Yuan Yuan. Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing*, 363:1–8, 2019.

[8] 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.

[9] Yi-Kuan Hsieh, Jun-Wei Hsieh, Xin Li, Yu-Ming Zhang, Yu-Chee Tseng, and Ming-Ching Chang. Scale-aware crowd counting network with annotation error modeling. *IEEE Transactions on Image Processing*, 34:2750–2764, 2025.

[10] Siyu Huang, Xi Li, Zhi-Qi Cheng, Zhongfei Zhang, and Alexander Hauptmann. Stacked pooling: Improving crowd counting by boosting scale invariance. *arXiv preprint arXiv:1808.07456*, 2018.

[11] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Shenghua Gao, Rongrong Ji, and Junwei Han. Body structure aware deep crowd counting. *IEEE Transactions on Image Processing*, 27(3):1049–1059, 2017.

[12] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

[13] 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.

[14] Muhammad Asif Khan, Hamid Menouar, and Ridha Hamila. Lcdnet: a lightweight crowd density estimation model for real-time video surveillance. *Journal of Real-Time Image Processing*, 20(2):29, 2023.

[15] Muhammad Asif Khan, Hamid Menouar, and Ridha Hamila. Revisiting crowd counting: State-of-the-art, trends, and future perspectives. *Image and Vision Computing*, 129:104597, 2023.

[16] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in neural information processing systems*, 23:1324–1332, 2010.

[17] 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.

[18] Wei Lin, Chenyang Zhao, and Antoni B. Chan. Point-to-region loss for semi-supervised point-based crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29363–29373, June 2025.

[19] Weizhe Liu, Nikita Durasov, and Pascal Fua. Leveraging self-supervision for cross-domain crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5341–5352, 2022.

[20] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7661–7669, 2018.

[21] 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.

[22] Baoyang Mu, Feng Shao, Zhengxuan Xie, Hangwei Chen, Zhongjie Zhu, and Qiuping Jiang. Misf-net: Modality-invariant and -specific fusion network for rgb-t crowd counting. *IEEE Transactions on Multimedia*, 27:2593–2607, 2025.

[23] Baoyang Mu, Feng Shao, Zhengxuan Xie, Long Xu, and Qiuping Jiang. Rgbt-booster: Detail-boosted fusion network for rgb-thermal crowd counting with local contrastive learning. *IEEE Internet of Things Journal*, 12(11):18331–18349, 2025.

[24] Akshita Patwal, Manoj Diwakar, Vikas Tripathi, and Prabhishek Singh. Crowd counting analysis using deep learning: A critical review. *Procedia Computer Science*, 218:2448–2458, 2023.

[25] Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Crowddiff: Multi-hypothesis crowd density estimation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12809–12819, 2024.

[26] Viresh Ranjan, Boyu Wang, Mubarak Shah, and Minh Hoai. Uncertainty estimation and sample selection for crowd counting. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[27] Avinash Rohra, Baoqun Yin, Hazrat Bilal, Aakash Kumar, Munawar Ali, and Yang Li. Msffnet: multi-scale feature fusion network with semantic optimization for crowd counting. *Pattern Analysis and Applications*, 28(1):21, 2025.

[28] David Ryan, Simon Denman, Sridha Sridharan, and Clinton Fookes. Scene invariant crowd counting. In *2011 International Conference on Digital Image Computing: Techniques and Applications*, pages 237–242. IEEE, 2011.

[29] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5245–5254, 2018.

[30] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1221–1231, 2019.

[32] 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.

[33] Jia Wan and Antoni Chan. Modeling noisy annotations for crowd counting. *Advances in Neural Information Processing Systems*, 33:3386–3396, 2020.

[34] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. *arXiv preprint arXiv:2009.13077*, 2020.

[35] Mingjie Wang, Hao Cai, Yong Dai, and Minglun Gong. Dynamic mixture of counter network for location-agnostic crowd counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 167–177, 2023.

[36] Mingjie Wang, Hao Cai, Jun Zhou, and Minglun Gong. Interlayer and intralayer scale aggregation for scale-invariant crowd counting. *Neurocomputing*, 441:128–137, 2021.

[37] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 43(6):2141–2149, 2021.

[38] Rui Wang, Yixue Hao, Long Hu, Jincai Chen, Min Chen, and Di Wu. Self-supervised learning with data-efficient supervised fine-tuning for crowd counting. *IEEE Transactions on Multimedia*, 25:1538–1546, 2023.

[39] Zhen Wang, Yuelei Li, Jia Wan, and Nuno Vasconcelos. Diffusion-based data augmentation for object counting problems. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[40] Chenfeng Xu, Dingkang Liang, Yongchao Xu, Song Bai, Wei Zhan, Xiang Bai, and Masayoshi Tomizuka. Autoscale: Learning to scale for crowd counting. *International Journal of Computer Vision*, 130(2):405–434, 2022.

[41] Maochen Yang, Zekun Li, Jian Zhang, Lei Qi, and Yinghuan Shi. Taste more, taste better: Diverse data and strong model boost semi-supervised crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24440–24451, June 2025.

[42] Shaopeng Yang, Weiyu Guo, and Yuheng Ren. Crowdformer: An overlap patching vision transformer for top-down crowd counting. In *IJCAI*, volume 1, page 2, 2022.

[43] 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.

A Related work on crowd counting

Crowd counting, scale robustness, and generative modeling. Crowd counting aims to estimate the number of people in an image or video, and has broad applications in urban planning, public safety, traffic monitoring, and event management [15, 24, 4]. Early approaches mainly relied on detection- or regression-based pipelines, while recent methods commonly formulate crowd counting as density map estimation using deep neural networks [26, 34]. Despite substantial progress, crowd counting remains challenging because of severe occlusion, large perspective-induced scale variations, complex backgrounds, and non-uniform crowd distributions [14]. Recent surveys further indicate that modern crowd counting models increasingly need to balance accuracy with computational efficiency, especially for real-time or edge-device deployment [4].

Scale variation is one of the central difficulties in crowd counting. To address this issue, [10] introduced multi-kernel and stacked pooling operations to improve scale robustness without adding extra parameters, while [7] proposed a single-column architecture with inter-layer and intra-layer scale aggregation to extract scale-invariant features. Scene-invariant counting has also been studied from the perspective of camera calibration and local features to improve transferability across environments [28]. Beyond feature aggregation, [29] proposed Adversarial Cross-Scale Consistency Pursuit (ACSCP), which combines adversarial learning with a cross-scale consistency constraint so that the count estimated from local patches remains coherent with the count estimated from their larger region. More recently, MSFFNet uses grouped VGG features, contrast-based multi-scale fusion, and semantic optimization to enhance head localization while suppressing background noise [27]. These methods show that scale robustness is not only a matter of using larger receptive fields, but also requires consistent feature fusion and reliable density-map supervision across scales.

Diffusion and generative models have recently been explored for crowd counting and data augmentation. CrowdDiff adopts conditional diffusion to predict density maps and alleviate limitations of conventional regression-based methods [25]. Diffusion-based augmentation has also been used to synthesize crowd images for improving model robustness [39]. In semi-supervised crowd counting, [41] further showed that generic augmentations such as Mixup may damage the spatial structure of crowd scenes and distort density maps. To better exploit unlabeled data, they proposed Taste More Taste Better (TMTB), which performs background inpainting augmentation while preserving foreground crowd regions, and combines a Visual State Space Model backbone with an anti-noise classification head to provide more robust supervision [41]. These works suggest that generative models are useful for crowd counting only when the spatial correspondence between images, point annotations, and density maps is carefully preserved.

Multi-modal crowd counting. RGB-only crowd counting often suffers under low illumination, heavy occlusion, and cluttered backgrounds. To improve robustness in such unconstrained scenarios, recent studies have explored RGB-D and RGB-T crowd counting, where depth or thermal images provide complementary spatial or radiation cues. For RGB-T crowd counting, MISF-Net decomposes RGB and thermal features into modality-invariant and modality-specific components through similarity and difference supervision, and further uses reconstruction supervision to avoid biased feature decomposition [22]. It also introduces a modality-invariant fusion module with different receptive

Table 5: Detailed parameters used for training on the six datasets were fixed and kept consistent across different based models.

Dataset	learning rate	batch size	crop size
UCF-QRNF	1e-5	12	512×512
UCF CC 50	1e-5	12	512×512
NWPU	1e-5	12	512×512
ShanghaiTech A	1e-4	12	512×512
ShanghaiTech B	1e-4	12	512×512
JHU	1e-4	12	512×512

fields to better adapt to pedestrian head-size variations [22]. RGBT-Booster focuses on another limitation of RGB-T counting: insufficient extraction of low-level head details and incomplete reduction of modality gaps. It introduces detail auxiliary branches for RGB and thermal streams, an enhanced detail fusion module, and local contrastive learning to bridge modality differences at local feature locations [23]. These RGB-T methods demonstrate that multi-modal crowd counting requires more than direct feature concatenation; effective fusion should consider both modality complementarity and modality discrepancy.

Semi-supervised and point-based crowd counting. Because dense crowd images require hundreds or thousands of point annotations, reducing annotation cost has become an important research direction. Learning-to-rank methods exploit unlabeled images by using relative count constraints between image crops and full images [20]. Recent semi-supervised methods often rely on teacher-student learning and pseudo labels, but pseudo-label quality remains a key bottleneck. For density-based semi-supervised counting, [41] improves unlabeled-data utilization through inpainting augmentation and anti-noise classification. For point-based crowd counting, [18] observed that directly applying point-to-point supervision to pseudo labels is ill-posed because pseudo-label confidence is not properly propagated to background pixels. They therefore proposed a Point-to-Region (P2R) loss, where each pseudo point supervises a local region rather than a single pixel, allowing foreground and nearby background pixels to share a more reliable confidence assignment [18]. This line of work indicates that semi-supervised crowd counting depends not only on generating pseudo labels, but also on designing supervision schemes that prevent local over-activation and background misclassification.

Annotation deviation and loss function. Crowd-counting datasets typically use *point-wise* or *dotted* annotations to mark each person in an image. Unlike bounding boxes, point annotations do not provide explicit size information, which makes it difficult to define the spatial extent of each person and to measure annotation deviation. To address this problem, [43] estimates head size by computing the average distance from each head point to its three nearest neighbors and using it to determine the Gaussian kernel width. The impact of body structure on counting accuracy was further studied in [11]. Locally connected Gaussian kernels were introduced in [5] to relax strict pixel-level invariance, and a decoupled two-stage framework was proposed in [3] to alleviate annotation-related issues. DM-Count formulates density-map supervision as distribution matching and uses Optimal Transport to measure the discrepancy between predicted and ground-truth density maps [34]. Location-agnostic crowd counting has also been explored to reduce the dependence on exact point locations [35]. More recently, a scale-aware loss was proposed to explicitly model annotation errors and generate finer density maps under scale variations [9]. Together with recent semi-supervised losses such as P2R [18], these studies show that loss design is crucial for handling the inherent uncertainty of point-level crowd annotations.

B Experimental setting

Overview

In this appendix, we provide additional ablation studies and experiments to complement the main paper. § C presents the ablation study on initial Gaussian variance. § D describes the ablation study on our scale-invariance term. § E evaluates the generalization of the scale invariance term integrated into SAL.

Table 6: Ablation study on the effects of the Gaussian annotation variance β_1 on MAE and MSE metrics using the CrowdDiff [25] method on various datasets with the same 512×512 input resolution. The best result is highlighted in red, and the second-best is in blue.

Datasets	β_1	2	4	6	8	10	12	14
QNRF	MAE	71.3	69.4	67.7	65.7	67.1	68.4	70.3
	MSE	114.6	109.3	104.1	98.4	102.2	106.8	111.6
SHHA	MAE	46.6	45.2	44.7	43.4	44.5	45.9	46.8
	MSE	69.9	68.5	67.2	65.3	66.9	68.9	70.4
SHHB	MAE	5.7	5.6	5.3	5.1	5.2	5.4	5.4
	MSE	7.3	7.0	6.9	6.5	6.6	6.8	6.9
NWPU	MAE	57.9	56.8	54.5	52.9	54.1	54.9	55.7
	MSE	207.2	196.8	183.3	173.5	181.7	184.4	191.6
JHU-CROWD++	MAE	48.3	47.5	46.2	43.6	45.8	46.5	47.9
	MSE	202.8	196.6	187.1	171.8	178.7	189.2	198.9

Table 7: Evaluation of alternative scale-invariance by comparing full-image counting with *Non-overlapping Patches* and *Overlapping Patches* strategies, with and without our proposed loss, under different resolutions across five benchmark datasets.

Method	**Input Resolution**	**QNRF**		**SHHA**		**SHHB**		**NWPU**		**JHU-CROWD++**	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Full image	(H,W)	71.5	109.7	51.3	81.7	5.3	7.5	59.8	271.3	52.7	213.4
	(H/2,W/2)	71.3	108.4	51.1	81.2	5.4	7.6	60.3	274.6	53.1	215.3
	(2H,2W)	71.6	110.9	51.5	82.1	5.3	7.6	60.7	277.9	53.4	218.6
Non-overlapping Patches	(H,W)	76.3	128.3	56.5	93.9	5.8	8.5	67.7	309.8	57.3	248.3
	(H/2,W/2)	89.6	147.4	66.3	108.4	6.9	9.7	78.2	339.1	76.3	281.6
	(2H,2W)	90.4	151.8	65.7	107.1	6.7	9.4	81.5	345.1	75.6	278.7
Non-overlapping Patches + Ours	(H,W)	74.1	115.7	54.0	84.3	5.6	8.0	62.9	281.4	54.2	229.9
	(H/2,W/2)	74.3	118.6	54.1	84.7	5.5	7.8	63.1	293.7	54.4	231.5
	(2H,2W)	74.2	117.4	53.9	84.1	5.7	8.2	63.8	296.6	54.5	232.1
Overlapping Patches	(H,W)	83.6	144.1	59.9	102.4	6.3	9.0	73.3	324.1	68.7	264.8
	(H/2,W/2)	99.3	165.6	69.5	114.8	7.8	11.7	85.1	350.4	80.3	299.3
	(2H,2W)	98.4	162.7	71.3	121.5	7.2	10.8	87.2	358.7	81.5	305.7
Overlapping Patches + Ours	(H,W)	78.6	125.1	57.9	92.4	6.2	8.8	68.3	308.1	64.7	248.3
	(H/2,W/2)	79.1	131.5	58.8	94.6	6.3	9.0	68.5	311.6	65.1	253.4
	(2H,2W)	78.5	134.8	59.1	95.4	6.3	9.0	69.8	313.7	65.3	254.7

C Ablation Study on Initial Gaussian Variance

To validate our choice of setting $\beta_1 = 8$ in the main method, we conduct an ablation study using the CrowdDiff framework [25], as summarized in Table 6. Following the formulation in [9], the Gaussian variances β_s at each scales is derived recursively $\beta_{s+1} = \beta_s/2$. We vary β_1 from 2 to 14 while keeping all other settings fixed, and report the resulting MAE and MSE on five benchmark datasets.

Results in Table 6 show that increasing β_1 from 2 to 8 consistently improves performance, achieving the lowest MAE and MSE when $\beta_1 = 8$ across all datasets. Further increasing β_1 degrades performance, likely due to over-smoothing from excessively wide Gaussian kernels. These findings validate the choice of $\beta_1 = 8$ for accurate density estimation in our main experiments.

D Ablation study on the scale-invariance term

To further assess the effectiveness of our loss under scale transformations, we design an alternative evaluation protocol that tests scale-invariance via partial image views. Specifically, we compare the predicted count from a full image with counts aggregated from two patch-based inference strategies:

- *Non-overlapping Patches*: the image is split into two disjoint regions (e.g., two $(H, W/2)$ patches).
- *Overlapping Patches*: the image is split into two partially overlapping regions (e.g., two $(H, 2W/3)$ patches), and the overlapping area is adjusted to avoid double-counting.

In both cases, the sum of patch-level predictions should match the full-image count.

Table 8: Generalization evaluation of the Scale-Invariance term integrated into SCALE-AWARE LOSS (SAL). Color values mark the best results across different input resolutions.

Loss	SI	Input Res.	QNRF		SHHA		SHHB		NWPU		JHU-CROWD++	
			MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
SAL [9]	×	(H, W)	73.9	121.7	52.1	76.6	6.1	9.7	70.0	211.4	53.6	201.5
		(H/2, W/2)	84.4	159.8	58.3	86.1	6.7	10.4	79.5	245.1	57.6	218.7
		(2H, 2W)	83.8	146.7	59.2	89.3	6.7	10.2	80.1	255.4	58.3	221.9
	✓	(H, W)	72.3	114.7	51.5	75.4	5.7	8.8	68.5	204.3	52.9	196.4
		(H/2,W/2)	72.7	118.3	51.6	75.9	5.6	8.6	69.3	207.6	53.1	198.3
		(2H,2W)	72.6	117.6	51.5	75.6	5.9	8.9	69.1	206.5	53.0	197.6

SI = Scale-Invariance

Table 7 shows that models trained without our loss suffer notable accuracy drops, especially with resolution changes and patch-based inference. This inconsistency is due to duplicated or missed heads near patch boundaries, particularly in dense scenes like NWPU. Sparse datasets like SHHB are less affected. In contrast, models trained with our proposed loss demonstrate strong robustness across all settings. Even under aggressive cropping and scaling, the prediction errors remain stable, and accuracy is preserved. This validates the ability of our loss to regularize the model toward scale-consistent behavior and handle local-region inconsistencies during inference.

Clarification: In the *Non-overlapping Patches* setting, two disjoint regions are used to estimate local counts, which are then summed. In the *Overlapping Patches* setting, two partially overlapping regions are used, and their predictions are summed and adjusted by subtracting the predicted count in the overlapping area. This approximates a more realistic multi-view aggregation process but still introduces potential redundancy or loss, which our method helps mitigate.

E Evaluating the Generalization of the Scale-Invariance Term Integrated into SAL

To assess the generalizability of our proposed Scale-Invariance term ($\mathcal{L}_{\text{SI}}$), we integrate it into the SCALE-AWARE LOSS (SAL) framework [9]. As shown in Table 8, at the original resolution (H, W), the SAL loss achieves relatively low MAE since it matches the training scale. However, without the inclusion of our $\mathcal{L}_{\text{SI}}$, SAL struggles to maintain performance at other resolutions such as $(2H, 2W)$ and $(H/2, W/2)$, showing its limitation in handling scale variation.

When combined with our $\mathcal{L}_{\text{SI}}$, MAE significantly improves across these resolutions, proving that our method boosts robustness to resolution changes. This confirms that our approach generalizes well and maintains consistent performance across scales without retraining for each resolution.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see limitations.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see methodology and experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see section 3.1, 3.2, 3.3, 3.4 and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We use the common result report on ImageNet, COCO datasets and ADE20K datasets, which does not include error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our paper did not deviate from the NeurIPS Code of Ethics

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM has little to do with our research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.