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

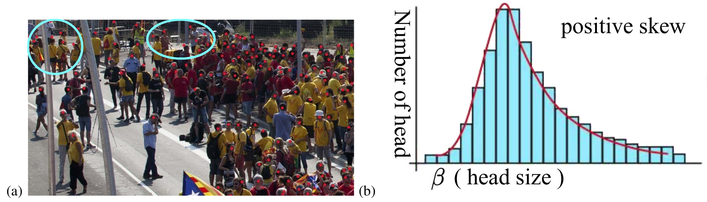
Yi-Kuan Hsieh[1], Kuan-Chuan Peng[2], Xin Li[3], Ming-Ching Chang[3], Chi-Chia Sun[4], Yu-Chee Tseng[1], Jun-Wei Hsieh[1]

國立陽明交通大學 NATIONAL YANG MING CHIAO TUNG UNIVERSITY
UNIVERSITY AT ALBANY STATE UNIVERSITY OF NEW YORK
MITSUBISHI ELECTRIC RESEARCH LABORATORIES

NEURAL INFORMATION PROCESSING SYSTEMS 2026

Why Scale-Consistent Counting?

- Zoom changes apparent head sizes, but the number of people should remain unchanged.
- Fixed-scale models produce inconsistent counts when the same scene changes resolution.
- Noisy head-point annotations can bias density regression, especially in dense crowds.



Scale variation + annotation uncertainty (Fig. 2)

One person contributes one count at every scale

SLICE jointly enforces cross-scale count consistency and corrects uncertainty in point annotations.

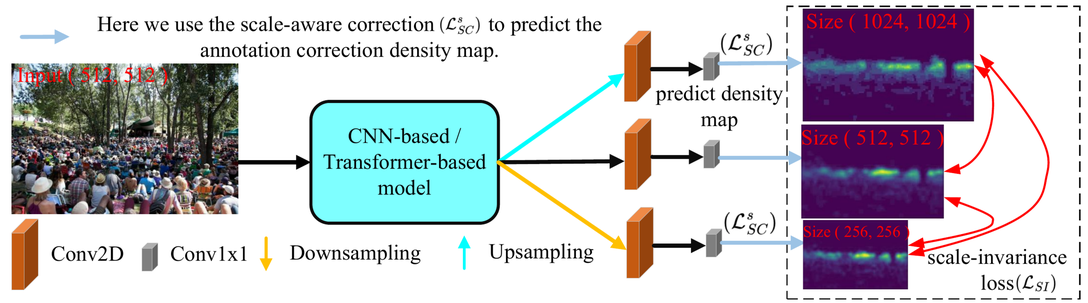
Probabilistic Annotation Model

Annotated position = true position + ε

ε ~ N(0, αI). α: annotation noise variance.
βs: Gaussian kernel variance at scale s.

SLICE: Method Overview

ScaLe-aware Invariance and Correcting Errors: a training objective for existing CNN-, Transformer- and diffusion-based crowd counting models.



Joint Training Objective

Scale consistency + error correction

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

λ balances cross-scale consistency with uncertainty-aware supervision.

Scale-Invariance Loss (SI)

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

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

Predict density maps at 0.5×, 1× and 2×.
Sum each map to obtain a crowd count.
Match counts across all pairs of scales.

Scale-aware Correction (SC)

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

Residual: predicted density − expected density.
Covariance models spatial uncertainty.
R encourages unit mass per head annotation.

Scale Robustness

(A) Transfer across loss functions

Steerer, QNRF MAE ↓ (paper Table 3)

Loss	0.5×	1×	2×
Bayesian loss	99.6	85.8	98.2
BL + SLICE	**81.1**	**80.7**	**81.2**
NoiseCC	89.2	78.6	86.2
NoiseCC + SLICE	**73.4**	**73.1**	**73.8**

Both SI and SC improve accuracy and keep errors stable when input resolution changes.

(B) Complementary loss components

CrowdDiff on QNRF

Training objective	MAE ↓	MSE ↓
Baseline	68.9	125.6
SC only	67.3	106.5
SI only	68.0	119.3
SI + SC (SLICE)	**65.7**	**98.4**

Joint optimization gives the best result

SI stabilizes counts across scales.
SC accounts for uncertain annotations.

Gaussian Variance Ablation

CrowdDiff, fixed input resolution (paper Table 6)

β1	2	4	6	8	10	12	14
QNRF MAE	71.3	69.4	67.7	65.7	67.1	68.4	70.3
QNRF 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

β1 = 8 gives the lowest errors among tested values.

Scale-wise variance follows β(s + 1) = βs / 2.

Accuracy Across Three Model Families

MAE ↓ at original input resolution, five benchmarks (paper Table 2)

Model	QNRF	SHHA	SHHB	NWPU	JHU
CrowdDiff (diffusion)	68.9	47.4	5.7	57.8	47.3
CrowdDiff + SLICE	**65.7**	**43.4**	**5.1**	**52.9**	**43.6**
Steerer (CNN)	74.3	54.5	5.8	63.7	54.3
Steerer + SLICE	**71.5**	**51.3**	**5.3**	**59.8**	**52.7**
CrowdFormer (Transformer)	78.8	56.9	5.7	67.1	59.8
CrowdFormer + SLICE	**74.2**	**53.5**	**5.2**	**62.3**	**54.4**

SHHA/B: ShanghaiTech A/B. JHU: JHU-CROWD++. Lower values are better.

Training & Deployment

Architecture-agnostic
Add SLICE to existing density estimators.

Single-scale inference
No additional inference-time modules.

Training setup
512 × 512 crops, batch 12, RTX 4090.

Practical limitation
Higher training cost. α/β may require tuning.

Contact: yikuan0725@gmail.com Corresponding author: jwhsieh@nycu.edu.tw