Taste More, Taste Better: Diverse Data and Strong Model Boost Semi-Supervised Crowd Counting

Supplementary Material

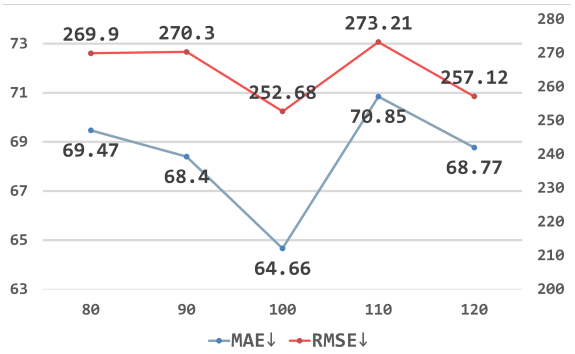


Figure 5. The impact of the weight decay speed T^{inpw} on the JHU-Crowd++ dataset with 5% labeled data.

6. Mathematical Definitions of MAE and RMSE Metrics

The MAE and RMSE are defined as follows:

$$\begin{aligned} \text{MAE} &= \frac{1}{N_i}\sum_{i=1}^{N_i}|C_i^{gt} - \hat{C}_i|, \\ \text{RMSE} &= \sqrt{\frac{1}{N_i}\sum_{i=1}^{N_i}(C_i^{gt} - \hat{C}_i)^2}. \end{aligned} \tag{9}$$

where N_i is the number of images, C_i^{gt} is the ground truth count, and $\hat{C}_i$ is the estimated count.

7. Detailed study of the weight decay speed

We study the impact of the weight decay speed T^{inpw} on the JHU-Crowd++ dataset with 5% labeled data. The influence of T^{inpw} on the MAE and RMSE is shown in Figure 5. We train and test TMTB with different T^{inpw} values ranging from 80 to 120. The results show that both MAE and RMSE decrease as T^{inpw} increases from 80 to 100. However, when T^{inpw} is greater than 100, the performance begins to degrade. Therefore, we set T^{inpw} to 100 in experiments, which achieves the best performance.

8. Detailed study of the number of warm-up epochs

We calculate λ_w based on training epoch t and pre-defined warm-up epoch T^w as $\lambda_w = e^{-5.0*(1.0-t/T^w)^2}$, when $t < T^w$. λ_w is set to 1.0 when $t \geq T^w$. Table 6 presents the experiments on 10% ShanghaiTech A to find optimal T^w.

Table 6. The impact of the number of warm-up epochs T^w on the ShanghaiTech A dataset with 10% labeled data.

T^w	MAE	RMSE
0	70.0	121.1
5	70.6	116.2
10	66.0	113.6
20	**65.7**	**110.4**
30	68.5	113.1
40	67.4	114.2
60	70.6	118.4
80	70.2	120.5

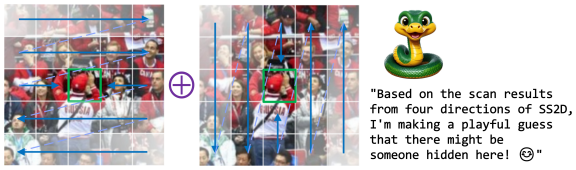


Figure 6. 2D-Selective-Scan (SS2D) helps the establishment of global receptive fields. Green boxes indicate the query image patch, with patch opacity representing the degree of information loss.

9. Text Prompts in Inpainting

We generate some positive text prompts to inpaint the images with diverse backgrounds and scenarios. The prompts are generated by large language models (LLMs), GPT-4o [35].

All inpainting processes share the same negative prompt. For the balance of diversity and quality, we do not forbid drawing extra people. The negtive text prompt we employed is:

> disfigured face, broken limbs, deformed body parts

And we design the positive prompts with diverse scenarios containing plants or animals. All used postive prompts are listed in Table 7. Note that, updating the prompts while inpainting is also available, and we employ the fixed database for simplicity.

10. Further Discussion of Architectures

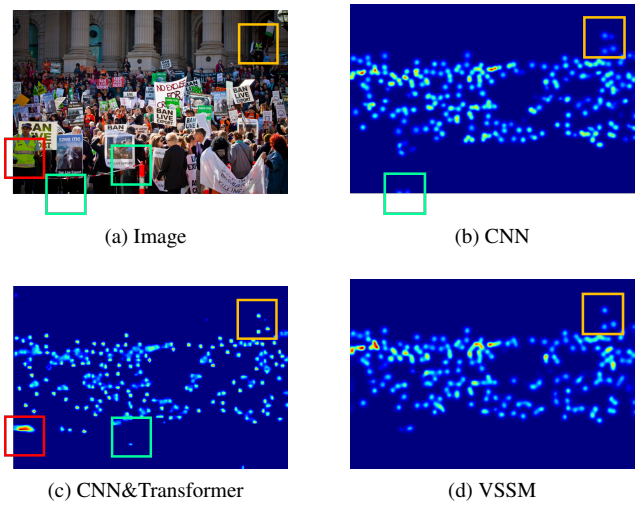
In Figure 6, we visualize the 2D-Selective-Scan mechanism of VMamba [29], which benefits the modeling of global context. The scanning operation in S6 fits well with NLP

Table 7. Positive prompts that we used in inpainting.

Positive Prompts
sunset over mountains, a lone eagle soaring, vibrant colors
ancient forest, misty atmosphere, deer grazing among trees
futuristic city skyline, neon-lit drones flying, cyberpunk style
serene beach, seashells scattered on the sand, gentle waves
snowy village, a fox prowling near cozy cabins, northern lights above
bustling marketplace, exotic fruits and spices, colorful fabrics swaying
abandoned castle, ivy-covered walls, crows perched on towers
desert landscape, cacti scattered, a lone lizard basking in the sun
underwater world, coral reefs teeming with fish, jellyfish drifting
enchanted garden, blooming roses, butterflies fluttering around
rainy city street, puddles reflecting streetlights, stray cat in the alley
starry night sky, a full moon shining, an owl perched on a tree
autumn forest, falling leaves in warm tones, a squirrel gathering acorns
medieval town square, horses tied to a post, pigeons pecking on cobblestones
tropical jungle, dense foliage, a parrot perched on a branch
twilight in the mountains, calm lake with lily pads, fireflies glowing
bustling urban park, tall trees with squirrels, children flying kites
ancient ruins, crumbling stone with moss, a snake slithering through the grass
futuristic lab, clean and sterile with robotic arms, plants growing in glass chambers
rustic farmhouse, golden wheat fields swaying, a scarecrow standing tall

tasks, while it faces a challenge when applied to vision data. The 2D-Selective-Scan is proposed to adapt S6 in Mamba [12] to vision data. SS2D consists of three steps: cross-scan, selective scanning with S6, and cross-merge. SS2D unfolds input patches of an image into sequences along four traversal paths (*i.e.*, Cross-Scan), and processes each sequence with a separate S6 block. After processing, the four sequences are reshaped and the resultant sequences are merged to form the output map (*i.e.*, Cross-Merge). As shown in Figure 6, SS2D enables every pixel in the image to integrate information from other pixels in four directions, facilitating the establishment of global receptive fields, which is necessary for counting in adverse scenarios.

We visualize the density maps predicted by different architectures in Figure 7. When facing low-light and adverse scenarios, VSSM still shows great capacity of capturing context features.



(a) Image (b) CNN (c) CNN&Transformer (d) VSSM

Figure 7. Predicted density maps. For the green bounding box, both CNNs and CNN with Transformer overfit to local texture information, generating nonexistent prediction. For the red bounding box, CNN with Transformer makes extremely erroneous predictions on regions with a strong contrast to the scene. For the yellow bounding box, CNN and CNN with Transformer make inaccurate predictions in low-light regions.