Supplementary Materials for "Shifted Autoencoders for Point Annotation Restoration in Object Counting"

Yuda Zou*, Xin Xiao*, Peilin Zhou, Zhichao Sun, Bo Du, and Yongchao Xu ✉

National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Hubei Luojia Laboratory, Wuhan University, Wuhan, China
{zouyuda,xinxiao,peilinzhou,zhichaosun,dubo,yongchao.xu}@whu.edu.cn

In this supplementary submission, we first conduct a toy experiment to quantify the enhanced consistency of the point annotations restored by SAE in Sec. 1. Next, we validate the effectiveness of further using the trained SAE as a pre-trained model for the counting task (similar to Masked Autoencoders (MAE) [1] for downstream tasks) in Sec. 2. We also analyze the efficiency of our method in Sec. 3. The ablation studies of the network and shifting mode are detailed in Sec. 4 and Sec. 5, respectively. Finally, some additional visualization results of the point annotations restored by SAE are present in Sec. 6.

1 Toy Experiment

In the main paper, we demonstrate that the point annotations revised by SAE are more consistent than the initial manual point annotations via some qualitative visualizations and effectiveness for final counting on eleven widely used datasets. Here we design a toy experiment to further validate the effectiveness of SAE by quantifying the consistency of the point annotations restored by SAE.

Specifically, we generate many images of identical stick figure heads (a circle shape with radius r) as training images. Since the relative spatial distribution of manual point annotations *w.r.t.* corresponding objects often roughly obeys a Gaussian function [7], we simulate the manual point annotations with random offset noise for each stick figure head by randomly sampling from a two-dimensional Gaussian distribution, which is centered at the center of the head circle. The distribution of simulated point annotations for the stick figure heads is illustrated in the left part of Fig. 1. Some annotations may even locate outside of the head region. This phenomenon also occurs in actual object counting datasets, which is shown by the green point within the white boxed area in Fig.Fig. 3. With simulated training images and corresponding point annotations, we apply our SAE to restore the simulated point annotations. As shown in the right part of Fig. 1, the point annotations restored by SAE are closer to the distribution center, thereby exhibiting improved consistency compared with the

* Equal contribution

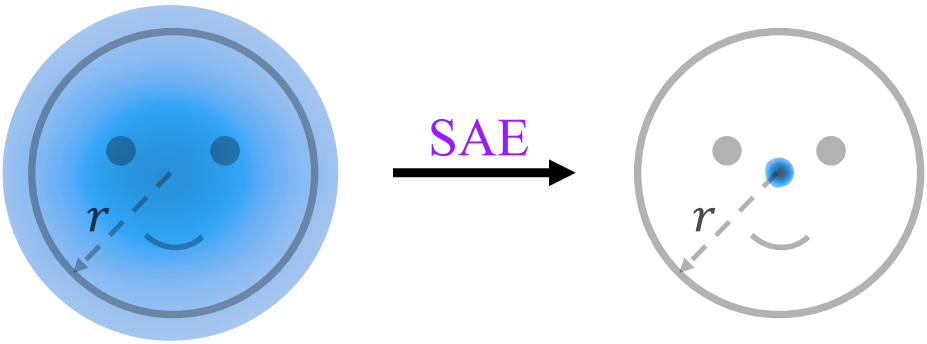

Fig. 1: The distributions of point annotations relative to the stick figure's heads before and after SAE restoration. Darker blue hues signify higher distribution density.

original simulated point annotations. To quantify the consistency of the point annotations, we compute the 2D position variances for the original simulated point annotations and the point annotations restored by SAE, respectively. Through the SAE restoration, the position variance of point annotations is reduced from $0.223 \times r^2$ to $0.016 \times r^2$, quantitatively demonstrating the effectiveness of SAE in enhancing the consistency of point annotations.

2 Pretraining of SAE

The Masked Autoencoders (MAE) [1] is trained by masking a portion of an image and reconstructing the masked image to the original unmasked one. This trained MAE backbone can be utilized to enhance the models of downstream tasks. Similarly, the proposed SAE is trained by restoring the shifted point annotations to their original positions. This process could be considered as a form of representation learning for object counting like MAE [1]. Therefore, the trained SAE also provides a pretrained backbone for counting models. We validate its effectiveness on two counting methods (one density-map-based and one localization-map-based) that employ the same VGG16 [4] backbone with the SAE network.

As shown in Tab. 1, utilizing the pretrained backbone of the SAE alone consistently enhances the counting performance of CSRNet [2] and P2PNet [6] on the SH PartA dataset [10]. While utilizing both the annotations restored by SAE and the pretrained backbone of SAE, further improvements in counting accuracies are achieved. Notably, P2PNet using both the SAE-revised annotations and the SAE pretrained backbone, decreases the Mean Absolute Error (MAE) and Mean Squared Error (MSE) from 52.8 and 85.8 to 47.2 ($\downarrow$ 5.6) and 75.7 ($\downarrow$ 10.1), respectively. This study demonstrates the effectiveness of the SAE pretrained backbone and the versatility of the SAE methodology for object counting.

Table 1: Experimental results of using the pretrained backbone of SAE and/or point restoration by SAE. Restoration: train with the point annotations restored by SAE; Pretraining: adopt the pretrained backbone of SAE.

Method	CSRNet [2]		P2PNet [6]	
	MAE↓	MSE↓	MAE↓	MSE↓
Baseline	64.0	110.3	52.8	85.8
& Restoration	57.2 (↓6.8)	96.8 (↓13.5)	48.2 (↓4.6)	76.1 (↓9.7)
& Pretraining	59.5 (↓4.5)	100.4 (↓9.9)	49.7 (↓3.1)	81.6 (↓4.2)
& Restoration & Pretraining	56.9 (↓7.1)	92.1 (↓18.2)	47.2 (↓5.6)	75.7 (↓10.1)

3 Efficiency Analysis

We analyze the efficiency of our SAE on the largest dataset we used, JHU-Crowd++ [5] containing 2272 training images with an average resolution of 1450×920. As stated in line 271-277, the training and testing were performed on a single NVIDIA 3090 GPU (24GB). The training process (with a batch size of 8 using 512×512 patches) takes about 4.8 hours and requires approximately 19.1GB of VRAM. The restoration phase on all 2272 training images takes about 5.4 minutes. Since our SAE only needs to perform once per dataset and provides refined point annotations that can then be repeatedly utilized by other counting methods without extra cost, this computation cost is acceptable and time could be reduced with more computational resources.

4 Network

We adopt UNet mainly because of its simplicity. As suggested, we test two other well-known segmentation networks (PSPNet and UPerNet) and achieve similar performance to UNet (see Tab. 2). Besides, we also try an object detector YOLOv7 using its regression branch. Though still being effective, the improvement declines (likely due to the smaller spatial size of output).

Table 2: Results of using various networks for point restoration.

Network	Dataset	MAE↓	MSE↓
P2PNet [36]	SH PartA	52.8	85.8
& SAE (UNet [3])	SH PartA	48.2 (↓4.6)	76.1 (↓9.7)
& SAE (PSPNet [11])	SH PartA	48.3 (↓ 4.5)	75.5 (↓ 10.3)
& SAE (UPerNet [9])	SH PartA	48.1 (↓ 4.7)	77.2(↓ 8.6)
& SAE (YOLOv7 [8])	SH PartA	50.7 (↓ 2.1)	83.7 (↓ 2.1)

5 Shifting Mode

We now directly shift (x, y) coordinates on SH PartA dataset, resulting in a rectangular sampling region. As depicted in Tab. 3, it is also quite effective. The

slightly inferior performance is mainly because the circular sampling region by angle and magnitude shift better matches the shape of human heads compared to the rectangular one.

Table 3: Results of different shift modes on SH PartA dataset.

Method	MAE↓	MSE↓
P2PNet[36]	52.8	85.8
& SAE (angle and magnitude shift)	48.2 (↓4.6)	76.1 (↓9.7)
& SAE (shifting x and y directly)	48.4 (↓4.4)	77.4 (↓ 8.4)

6 Visualization

We illustrate some more visualization results of restored point annotations given by the proposed SAE in Fig. 2 Fig. 3 Fig. 4. Green points: initial point annotations; Red points: revised point annotations; Yellow points: initial point annotation coincides with the corresponding revised point annotation.

References

1. He, K., et al.: Masked autoencoders are scalable vision learners. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16000–16009 (2022)
2. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1091–1100 (2018)
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. pp. 234–241 (2015)
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Int. Conf. Learn. Represent. pp. 1–14 (2015)
5. Sindagi, V.A., Yasarla, R., Patel, V.M.: Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. IEEE Trans. Pattern Anal. Mach. Intell. **44**(5), 2594–2609 (2022)
6. Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y.: Rethinking counting and localization in crowds: A purely point-based framework. In: Int. Conf. Comput. Vis. pp. 3365–3374 (2021)
7. Wan, J., Wu, Q., Chan, A.B.: Modeling noisy annotations for point-wise supervision. IEEE Trans. Pattern Anal. Mach. Intell. **45**(12), 15065–15080 (2023)
8. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7464–7475 (2023)
9. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Eur. Conf. Comput. Vis. pp. 418–434 (2018)
10. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 589–597 (2016)

11. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2881–2890 (2017)

Fig. 2: Visualization of restored point annotations provided by the SAE for crowd counting datasets. Zooming in is recommended for better readability in digital format.

Fig. 3: Visualization of restored point annotations provided by the SAE for crowd counting datasets. The white boxed area in the middle left image presents some heads with point annotations outside the head regions. Zooming in is recommended for better readability in digital format.

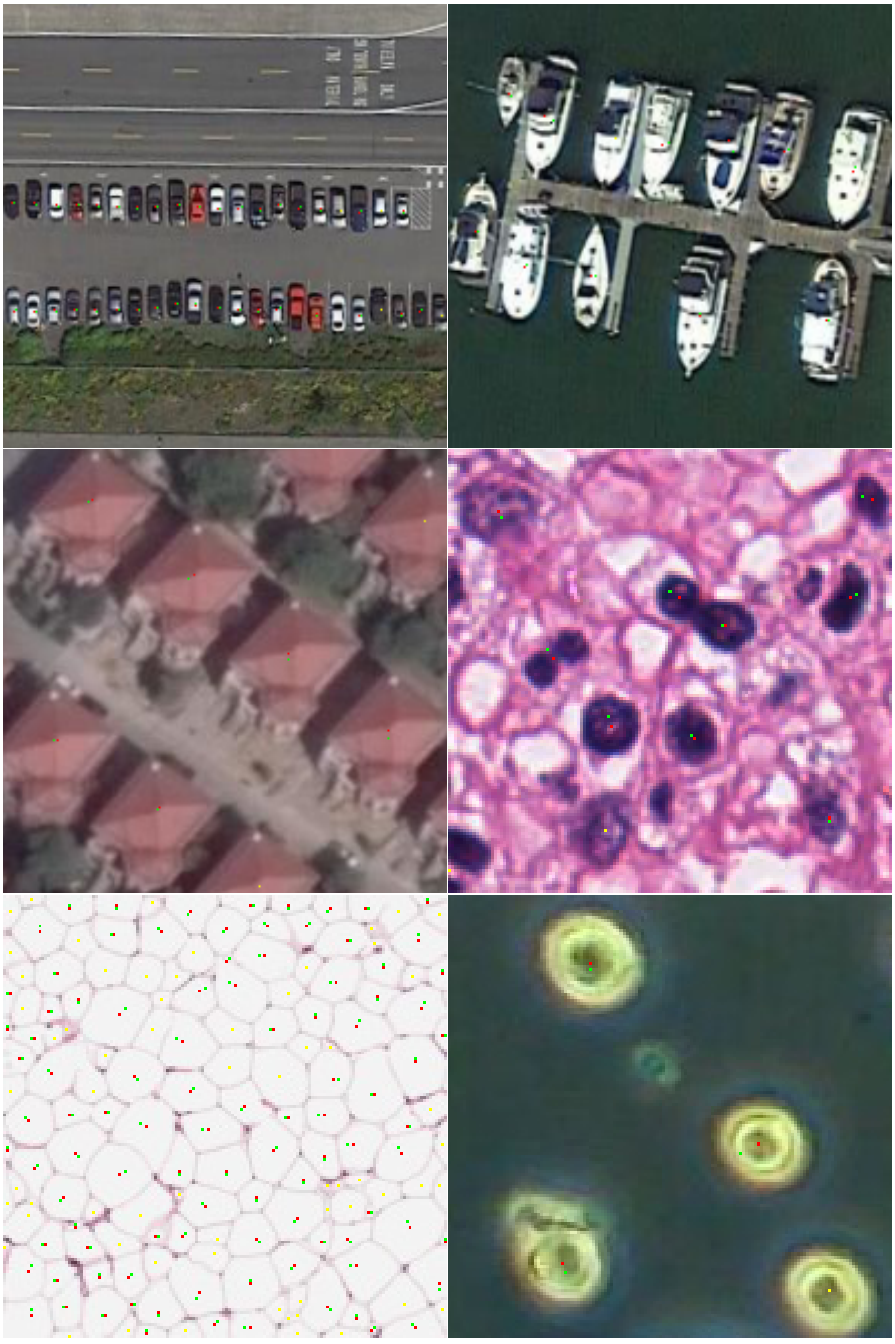

Fig. 4: Visualization of restored point annotations provided by the SAE for remote sensing object counting datasets and cell counting datasets. Zooming in is recommended for better readability in digital format.