We thank you for the valuable comments on our submitted manuscript. Comments from the reviewers have been addressed point-by-point as described below. **Statement:** Firs of all, we want to apologize for some mistakes in the paper, which is because of the incompatible document format and negligence of self-review. We want to thank all reviewers' serious and professional review and apologize for inconvenience in review. It mainly includes: (1) The figure 1 does not show completely. (2) some mistakes in table 1. Because a revised paper cannot be submitted at this stage, our revised version is available at https://github.com/simonchance/ICASSP2023 for reviewers' better examination. Revised parts are marked by red color. Moreover, related codes and pre-trained model can be obtained at mentioned website too, which is for better validation of our work.

Reviewer #3

6. Technical Correctness: Has major problems

Response: we already modified some mistakes in our paper as shown in mentioned revised paper.

7. Justification of Technical Correctness: (1)Figure 1 has some minor errors (2) The results of sota method has major errors:

Response: (1) Just like we respond in statement, we already revised it. (2) Thank you for your careful review, the result of CAN [1] is correct but we tagged the wrong paper. And the result of MFANet is like you pointed, it should be 7.2, 11.6 on ShanghaiTech Part B. We made these two mistakes carelessly when wrote and already modified them.

15. Justification of Overall evaluation of this paper: I seriously doubt the authenticity of the experimental results in this paper, because this paper reports incorrect results of other methods.

Response: Because of urgency of submission, some incorrect results exist in our paper, we apologize for our careless review. Authenticity of the experimental results can be validated in codes and pre-trained model that we provide.

19. Detailed assessment of the paper **Response:** Like we respond before in 7.

Reviewer #4

11. Justification of Clarity of Presentation Score: figure1 does not show completely.

Response: Thank you for your careful review. It is because of the incompatible document format and we already revised the figure 1 in our paper.

14. Overall evaluation of this paper: Marginal accept

Response: Thank for your careful and professional review. Revising according to your comments really makes our paper better.

15.Justification of Overall evaluation of this paper: Compared with the existing work, this paper is innovative. From the experimental results, the proposed method is effective. But I have some questions about the design of the network structure.

Response: Thank for your careful review and recognition of our work.

19. Detailed assessment of the paper

Response: Due to the good performance of VGG in image classification and other fields, VGG pre-trained on ImageNet is widely adopted in crowd counting, which is because it has less parameters and makes model faster to convergence. But we just use a part of VGG16 (first 10 convolutions and 3 max-pooling layers), which contains less max-pooling layers. In this way, we can remain mentioned advantages but not make features extracted to smaller scale. VGG extracts small-scale information certainly which is because of its three max-pooling layers. However, crowd scenes in public datasets and real world almost consist of small-scale and medium-scale people. The features what VGG16 extracts contain more information in channels while they are in small-scale. Although the size of feature maps are reduced, we obtain bigger receptive field. Thus, these features help us obtain abstract crowd distribution information. Meanwhile, this affect is unavoidable because the max-pooling layers of VGG can reduce the parameters and computational complexity, which is necessary. And we conduct up-sampling at the back-end network through transposed convolution, which can recover the density map to a large-scale degree. Other backbones may bring more parameters while accuracy is not necessarily better [2][3] ([2] utilize VGG19 to extract features while achieve 67.0 MAE, and [3] utilize ResNet101 to obtain 64.8 MAE on SHA dataset). On the other hand, most of previous works first utilize VGG to extract the features of input images, we follow them to achieve fair comparison.

Reviewer #5

2. Importance/Relevance: Irrelevant or out of scope

Response: Thank you for your careful review but our work is definitely of sufficient interest to ICASSP where accept dozens of relevant papers in recent years.

5. Justification of Novelty/Originality: Why do we use the density map as ground truth? There is a basic error in the field of crowd counting. Density map is not accurate enough!

Response: At the beginning of crowd counting study, researchers utilize the number of crowd as the ground truth to train counting model. However, this manner cannot reflect the spatial distribution information of crowd. Thus, Lempitsky proposed to use density map as ground truth and the integral of it gives the number of people. Of course, at the beginning of this study, there are other manners to train counting model such as bounding box [4][5] which is mainly utilized in object detection and point super supervision [6]. Bounding box may be accurate in objection, however, bounding box is not suitable for crowd counting. In one aspect, people in crowd counting scenes is very small, which makes it much harder to classify and regress people via bounding box. Tiny people are usually not detected. Thus, the counting accuracy of detection-based methods is much worse than that of density map. What's worse, the number of people in crowd counting task is massive, which means annotating the bounding boxes of each person will cost a great many of labor force. Recently, point supervision crowd counting has presented by few researchers, but it is difficult to converge and this manner usually spend a lot of training time which makes it hard to train. Density map based methods actually have certain errors, like we illustrated in figure 4, which is overestimation or underestimation people number in some regions. So, our work is devoted to designing network for obtaining rich multi-scale information to achieve more accurate crowd counting. In conclusion, presently, utilizing the density map as ground truth still is the mainstream supervision to train counting model and we also adopt this manner to achieve stable and fair comparison.

6. Technical Correctness: Contains minor errors **Response:** we apologize for our careless mistakes and we already revised it.

15. Justification of Overall evaluation of this paper: Why do we use the density map as ground truth? There is a basic error in the field of crowd counting. Density map is not accurate enough! 2. Do not use "*" to represent multiply operation.

Response: (1) Like we respond in 5. (2) Thank you for your professional suggestion, we already modified it in our revised paper.

19. Detailed assessment of the paper **Response:** Like we respond in 15.

[1] Weizhe Liu, Mathieu Salzmann, and Pascal Fua, "Context-aware crowd counting," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5099–5108.

[2] Chen X, Yan H, Li T, et al. Adversarial scale-adaptive neural network for crowd counting[J]. Neurocomputing, 2021, 450: 14-24.

[3] Wang Q, Gao J, Lin W, et al. Learning from synthetic data for crowd counting in the wild[C]//CVPR. 2019: 8198-8207.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn:Towards real-time object detection with region proposal networks," TPAMI., vol.39, no. 6, pp. 1137–1149, 2017.

[5] Peiyun Hu and Deva Ramanan, "Finding tiny faces," in CVPR, 2017.

[6] Y. Wang, X. Hou and L. -P. Chau, "Dense Point Prediction: A Simple Baseline for Crowd Counting and Localization," *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Shenzhen, China, 2021, pp.