# Towards Discriminative Semantic Relationship for Fine-grained Crowd Counting

Shiqi Ren, Chao Zhu*, Mengyin Liu, Xu-Cheng Yin
*School of Computer and Communication Engineering*
*University of Science and Technology Beijing*
Beijing, China
shiqiren@xs.ustb.edu.cn; chaozhu@ustb.edu.cn; blean@live.cn; xuchengyin@ustb.edu.cn

***Abstract*—As an extended task of crowd counting, fine-grained crowd counting aims to estimate the number of people in each semantic category instead of the whole in an image, and faces challenges including 1) inter-category crowd appearance similarity, 2) intra-category crowd appearance variations, and 3) frequent scene changes. In this paper, we propose a new fine-grained crowd counting approach named DSR to tackle these challenges by modeling Discriminative Semantic Relationship, which consists of two key components: Word Vector Module (WVM) and Adaptive Kernel Module (AKM). The WVM introduces more explicit semantic relationship information to better distinguish people of different semantic groups with similar appearance. The AKM dynamically adjusts kernel weights according to the features from different crowd appearance and scenes. The proposed DSR achieves superior results over state-of-the-art on the standard dataset. Our approach can serve as a new solid baseline and facilitate future research for the task of fine-grained crowd counting.**

***Index Terms*—Crowd counting, fine-grained counting**

## I. Introduction

The goal of crowd counting [1] is to estimate the number of people in an image, especially in crowded scenes. The crowd density and quantity information provided by this task can help people make further decisions. It is widely used in public safety, traffic planning, business management, etc. and has attracted more and more attention. Common crowd counting methods do not provide fine-grained semantic categorial information which is more useful than merely the whole number under certain circumstances and thus can bring wider application potential. Considering this deficiency, [2] propose a novel fine-grained crowd counting task including four subtasks, Standing/Sitting, Waiting/Not waiting, Towards/Away and Violent/Non-violent. It aims to divide the crowd into categories and then estimate the number of people belonging to each category. They also construct a dataset for these four subsets and a baseline approach which achieves good performance on the dataset.

The two-branch architecture of this baseline approach comprises a density estimation branch and a semantic segmentation branch. The former performs traditional crowd counting and the latter segments different categorial groups as well as the background. Then fine-grained density maps are the product of density map and segmentation maps. Finally, the number of people in each group is obtained by the summing integral of its corresponding fine-grained density map.

*Corresponding author. This work was supported by National Key Research and Development Program of China (2020AAA0109701), National Natural Science Foundation of China (62072032, 62076024), and National Science Fund for Distinguished Young Scholars (62125601).

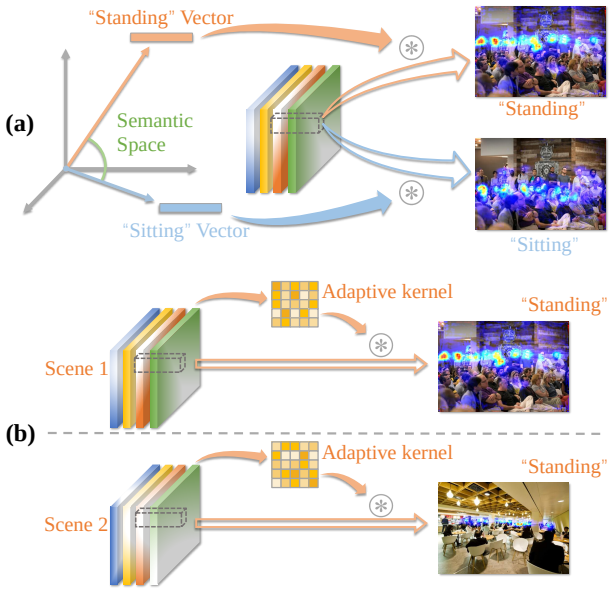


Fig. 1. Illustrations of our proposed method. (a) The word vectors in the semantic space can represent both the similarity among people (they are not orthogonal), as well as the inter-category difference (they have the angle measured by cosine distance). (b) Adaptive kernels are generated for specific scenes to better model the intra-category variations than static kernels. The hollow arrows show the pipeline of previous methods and the solid ones show the pipeline of ours.

This new task is more challenging than the traditional counting task due to the following main problems: 1) high similarity of inter-category crowd appearance, 2) intra-category crowd appearance variations, and 3) frequent foreground and background scene changes. Considering the crowded scene in Fig. 1(a), the upper image shows the density map of the "Standing" group and the lower one shows that of the "Sitting" group. These two groups have similar appearance and are hard for the model to differentiate. In Fig. 1(b), two images on the right shows that appearance of the same group varies in different scenes. The two "Standing" groups have distinct scales, density and context. Therefore, facing the

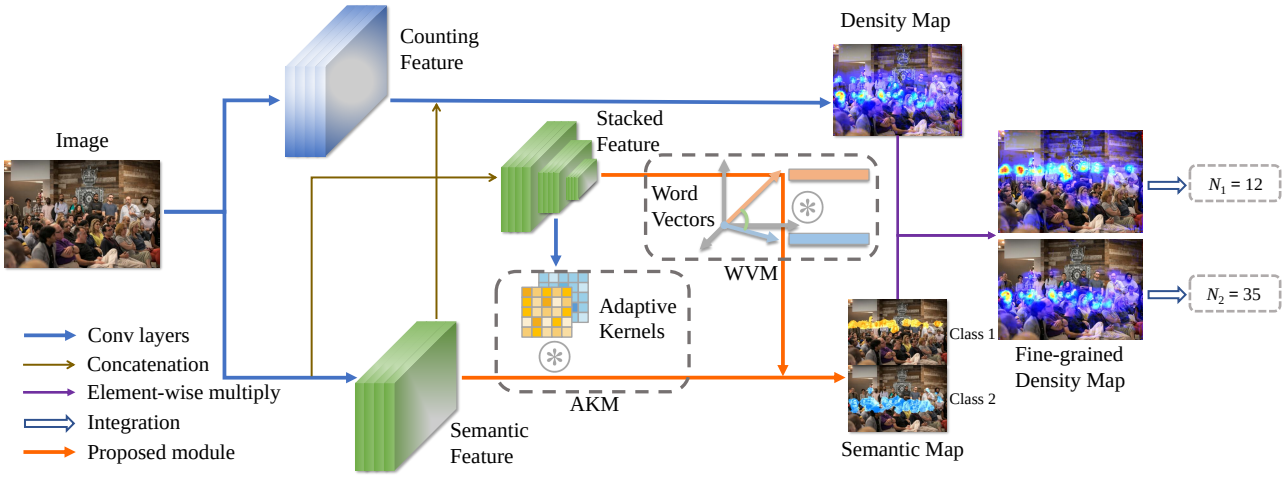


Fig. 2. The overall architecture of our DSR framework. For an input image, the counting feature, semantic feature and stacked feature are first obtained. Then the proposed WVM and AKM are adopted to generate the semantic maps. The fine-grained density maps are calculated by multiplying the obtained semantic maps with the density map generated from the counting feature. The final count number of $i$th class $N_i$ is the summing integral of the corresponding fine-grained density map.

above challenges, how to improve categorial counting accuracy by effectively modeling more discriminative relationship is essential to achieve better performance on fine-grained crowd counting task.

Due to the similarity of crowd appearance, it's challenging to distinguish different groups of people only from their visual features. By modeling the semantic relationship among different groups, features are mapped into a semantic space where both the similarity and the difference between categories are considered, just similar as what human beings do. Inspired by this idea, we propose the Word Vector Module (WVM) in the task of fine-grained crowd counting to boost its performance. As illustrated in Fig. 1(a), noticing that natural angle between the word vectors are neither orthogonal nor overlap to each other and properly measure the similarity and difference between categories, we propose to adopt different word vectors (e.g., "Standing" and "Sitting") to perform pixel-by-pixel convolution and to gain better segmentation result of different categories.

As for the challenge of intra-category variations and scene changes, previous methods are mainly based on static network structures, which are insufficient to handle the variations of crowd appearance, crowd density and scene context changes, as illustrated in hollow arrows in Fig. 1(b). To address this challenge, we propose a novel dynamic structure, Adaptive Kernel Module (AKM), to better model intra-category semantic relationship. As shown in Fig. 1(b), kernels are dynamically generated in a joint-trained network according to different input features and the "Standing" groups with diverse appearance in different scenes can be properly recognized.

Moreover, we introduce stacked feature (concatenation of multi-level features) to effectively broaden visions for the two proposed modules, so that our proposed modules can receive more global information and contextual information to further improve their effectiveness.

In conclusion, we have observed that: inter-category similarity among crowd appearance, as well as the variations among intra-category crowd appearance and different scenes are main challenges to suppress the performance improvement of fine-grained crowd counting, which remain to be comprehensively investigated. Therefore, we propose a novel fine-grained crowd counting approach to more semantically and dynamically generate fine-grained density map. Our main contributions can be summarized as follows:

- We propose the Word Vector Module (WVM) to better distinguish crowd in different categories by modeling both their similarity and difference in a more explicit semantic space, which overcomes the deficiency of classifying only from visual features of their similar appearance.
- We propose the Adaptive Kernel Module (AKM) to enhance the flexibility and handle the intra-category variations of crowd appearance, crowd density and scene context more easily than previous static structures.
- Extensive experiments and ablation studies are conducted to highlight the benefits of the proposed approach. The results show that our approach achieves superior performance over state-of-the-art on fine-grained crowd counting dataset.

## II. Related Work

### A. Crowd Counting

Generally, crowd counting methods can be divided into detection based methods [3], regression based methods [4] and density map estimation methods [5]. Density map estimation methods have been dominant since spatial distribution information of the crowd is utilized. In the past decade, with the emergence of deep learning, feature representation has changed from manual design to a learning-based one. CSRNet [6] uses the split convolution to achieve the goal of expanding the receptive field and uses the first 10 layers of VGG-16 [7]

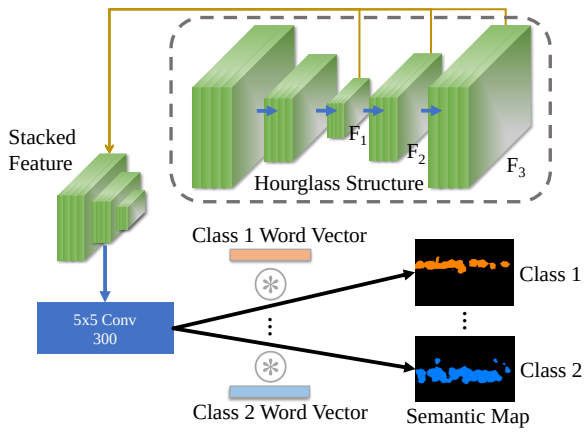


Fig. 3. Illustration of WVM. The stacked feature is extracted from the last three layers of Hourglass structure and then is convolved with the word vectors of each category respectively to obtain the semantic maps for each category.

with strong feature extraction ability as the backbone network to extract the image features. These two advantages improve the performance of the model and make it a widely used baseline. Recently,

### B. Fine-grained Counting

Besides the human-centric approach [2], some other fine-grained counting works are reported. [8] propose KR-GRUIDAE dataset containing images of 5 kinds of birds and a simple architecture including one encoder and two decoders. Equivalent to two-branch architecture, the two decoders respectively generate total density map and class probability mask. [9] propose a crowd-sourced dataset named Seal Watch, which contains 8 classes of seals. A density map generation method for crowd-sourced dataset is proposed and the fine-grained density map estimation method follows the previous two-branch strategy in [2]. These two works do not take extra effective means to improve the classification accuracy.

## III. Proposed Method

The overall architecture of our proposed DSR network is illustrated in Fig. 2. For an input image, the counting feature and semantic feature are firstly obtained respectively. Stacked feature is the concatenation of several layers' output during semantic feature generation for representing more context information. Then the proposed Word Vector Module (WVM) is adopted to classify each pixel more semantically and the proposed Adaptive Kernel Module (AKM) is adopted to dynamically generate the semantic maps which will be fused and then multiplied with the density map generated from the counting feature. The fine-grained density map can be obtained from these structures and the final counting results can be calculated through integration.

### A. Baseline Framework

We adopt CSRNet [6] as the backbone to generate initial counting feature $F_{co}$ and initial semantic feature $F_{so}$. Then with two coupled Hourglass [10] structure, we get the counting feature $F_c$ and the semantic feature $F_s$. The density map for the whole crowd $D$ is obtained after a 5x5 convolution layer. In baseline, the semantic map $S$ is the output of another 5x5 convolution layer. All the above settings and loss functions follow [2].

### B. Word Vector Module

Since the crowd of different categories may have similar appearance and are hard to be differentiated only from the visual features, we aim to better model their features in a more explicit semantic space by the proposed Word Vector Module (WVM).

The structure of WVM is shown in Fig. 3. We adopt a 300-dimensional fastText [11] word vector as the semantic representation for each category. For example, for Standing/Sitting subtask, we adopt the word vectors of "standing" and "sitting" respectively to represent these two categories. And then we calculate the similarity between the semantic feature of each pixel and the word vector of each category to get the classification result at each pixel. The similarity calculation can be simply done by a convolution operation. Denote the word vector for category $i$ as $V_i$, the semantic map generated by WVM $S_{Wi}$ is represented as:

$$S_{Wi} = \delta(F_s) \circledast V_i, \tag{1}$$

where $\circledast$ means the convolution operation and $\delta(\cdot)$ represents the 5x5 convolution layer. During training, word vectors can be set as trainable parameters since they may not perfectly describe the semantic distance for all cases but still guide the model to map feature to the semantic space. Related experiments are reported in the supplementary file.

### C. Adaptive Kernel Module

Besides the WVM, we further propose the Adaptive Kernel Module (AKM) to model the intra-category semantic relationship. The details of AKM are illustrated in Fig. 4. We get kernel feature through four convolution layers, and then generate $k$ kernels for segmenting background, $k$ kernels for segmenting foreground and others for segmenting each category from kernel feature. Kernel feature $F_{kernel}$ can be described as:

$$F_{kernel} = \eta(F_s), \tag{2}$$

where $\eta$ is the four-layer convolution structure.

As a hierarchical structure, the AKM generate semantic maps in two stages. At the first stage, the semantic feature is transferred into background semantic map and foreground feature by respectively convolving with the corresponding $k$ kernels. At the second stage, foreground feature is then transferred into semantic map of each category. The $k$ is a hyperparameter whose proper value will be explored by experiments.

The semantic maps for the background $S_{bg}$ and for the $i$th crowd category $S_{Ai}$ are:

$$S_{bg} = \sum_{j=1}^{k} F_s \circledast \psi_{bg}(F_{kernel})_j, \tag{3}$$

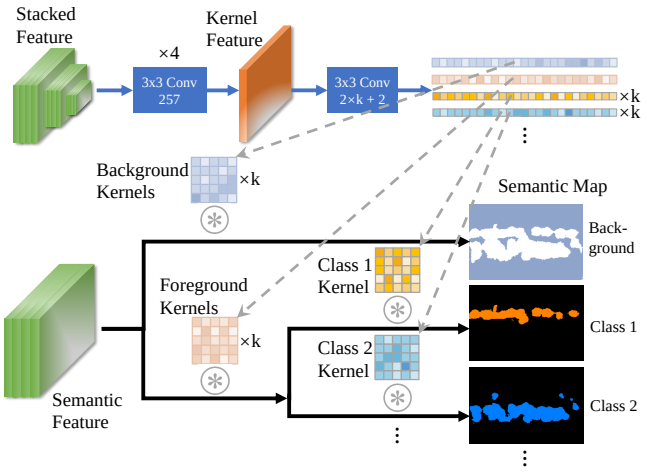


Fig. 4. Illustration of AKM. We first get kernel feature from stacked feature, and then generate $k$ kernels for segmenting background, $k$ kernels for segmenting foreground and others for segmenting each category from kernel feature.

$$S_{Ai} = \left( \sum_{j=1}^{k} F_s \circledast \psi_{fg}(F_{kernel})_j \right) \circledast \psi_i(F_{kernel}), \quad (4)$$

where $\psi$ is the convolution layer and $\psi_{bg}(\cdot)_j$, $\psi_{fg}(\cdot)_j$, $\psi_i(\cdot)$ respectively stand for the $j$th kernel for segmenting background, the $j$th kernel for segmenting foreground and the kernel corresponding to category $i$.

### D. Stacked Feature Extraction

To leverage more global information and contextual information to further improve the effectiveness of the two proposed modules, we concatenate the output of last three layers of Hourglass structure as stacked feature, which will take the place of semantic feature when convolving with word vectors and generating adaptive kernels. Due to the scale difference of these features, they should first be upsampled into the same scale. As illustrated in Fig. 3, we annotate the output of last three layers as $F_1$, $F_2$ and $F_3$, then the stacked feature $F_{stacked}$ is represented as:

$$F_{stacked} = \phi(\rho(F_1, 4), \rho(F_2, 2), F_3), \quad (5)$$

where $\phi(\cdot)$ represents concatenation and $\rho(x, t)$ represents upsampling $x$ with $t$ times.

### E. Semantic Map Fusion

The final semantic map $S$ is obtained by fusing the outputs of WVM and AKM. We set a hyperparameter $\lambda$ as the fusion ratio whose proper value is obtained via experimental trials. Then the final semantic map for $i$th category $S_i$ can be derived from:

$$S_i = \lambda S_{Wi} + (1 - \lambda) S_{Ai}. \quad (6)$$

### F. Fine-grained Density Map and Counting Results

For both baseline and our proposed approach, the fine-grained density map $G_i$ for $i$th category can uniformly be calculated by:

$$G_i = S_i \odot D, \quad (7)$$

where $\odot$ is element-wise multiply and $D$ is density map for the whole crowd. And counting number $N_i$ for $i$th category can be calculated as the integration of $G_i$.

## IV. Experiments

In this section, extensive experiments are conducted on fine-grained crowd counting dataset to evaluate our proposed method. Ablation study is performed to validate the effectiveness of key components in the proposed framework. Furthermore, we also report the state-of-the-art comparison.

### A. Dataset

The fine-grained crowd counting dataset [2] contains four fine-grained counting tasks: 1) Standing/Sitting, 2) Waiting/Not waiting, 3) Towards/Away, and 4) Violent/Non-violent. For each subtask, 100 images are used for testing except that test set for Towards/Away subtask has 800 images. The images are labeled with dot annotations indicating each person's category and location in the image. Following its evaluation settings, we report MAE (Mean Absolute Error) for each category, CMAE (Category MAE) and OMAE (Overall MAE) in the following experiments. The closer these errors are to zero, the better performance they indicate. Specifically, CMAE is the category-wise average value of MAEs, thus it's the most important evaluation criterion for fine-grained counting.

### B. Implementation Details

One Nvidia 1080Ti GPU is utilized for training, with batchsize as 1 and learning rate as $1 \times 10^{-4}$.

Similar to crowd counting, the ground truth of fine-grained density maps are generated from dot maps by convolving with a Gaussian kernel with bandwidth $\sigma$ which can be fixed or geometry-adaptive [5] (changed with crowd density at different pixel location). [2] use fixed-bandwidth Gaussian kernel to generate density map ground truth.

However, we find that the geometry-adaptive method outperforms the fixed one during reimplementation. And we fail to reproduce the results on Towards/Away subtask trying both fixed and geometry-adaptive methods possibly due to low resolution and single channel of images for this subtask. Some special settings may be required. Given all of these results, we adopt the geometry-adaptive method. In following experiments, "FCC" indicates the results reported in [2] and "baseline" means the results we reproduce with a different geometry-adaptive density map generation method.

TABLE I
ABLATION STUDY
BOLDEN ARE THE BEST RESULTS AND UNDERLINED THE 2nd BEST.

| GAG | WVM | AKM | SF | Standing | Sitting | **CMAE** | OMAE |
|---|---|---|---|---|---|---|---|
| | | | | 8.36 | 5.56 | 6.96 | - |
| ✓ | | | | 7.28 | 5.38 | 6.33 | 7.56 |
| ✓ | ✓ | | | 7.23 | 5.27 | 6.25 | 7.36 |
| ✓ | | ✓ | | <u>6.89</u> | <u>5.02</u> | <u>5.96</u> | **6.51** |
| ✓ | ✓ | ✓ | | 7.26 | **4.71** | 5.99 | 7.46 |
| ✓ | ✓ | ✓ | ✓ | **6.58** | 5.15 | **5.87** | <u>6.68</u> |

TABLE II
COMPARISON AMONG THE VARIANTS OF AKM

| | Standing | Sitting | **CMAE** | OMAE |
|---|---|---|---|---|
| Baseline | 7.28 | 5.38 | 6.33 | 7.56 |
| Common convs | 7.11 | 5.61 | 6.36 | 7.22 |
| 1×3 | 7.62 | 5.78 | 6.70 | 7.45 |
| 1×2+1×2 | **6.80** | 5.22 | <u>6.01</u> | 7.04 |
| 3×2+1×2 | <u>6.83</u> | 5.19 | <u>6.01</u> | 7.29 |
| 1×2+3×2 | 7.70 | 6.18 | 6.94 | 7.50 |
| 5×2+1×2 | 7.36 | 5.36 | 6.36 | <u>6.58</u> |
| 5+9+1×2 | 7.16 | 5.31 | 6.24 | 6.77 |
| 9+5+1×2 | 7.24 | **4.91** | 6.07 | 6.71 |
| 9×2+1×2 | 6.89 | 5.02 | **5.96** | **6.51** |
| 9+17+1×2 | 7.42 | <u>4.93</u> | 6.18 | 7.30 |
| 17+9+1×2 | 6.86 | 5.29 | 6.07 | 7.14 |
| 17×2+1×2 | 7.46 | 5.30 | 6.38 | 7.09 |
| 33×2+1×2 | 7.22 | 5.06 | 6.14 | 6.76 |
| 65×2+1×2 | 7.11 | 5.43 | 6.27 | 7.09 |

### C. Ablation Study

We first conduct an ablation study to validate the effectiveness of the key components in our approach DSR. The experiments are performed on the Standing/Sitting subtask of the dataset. "GAG" means the geometry-adaptive density map generation method and "SF" indicates the stacked feature. As shown in Table I, the results reveal that: 1) The baseline (GAG) together with either WVM or AKM alone can achieve better performance, which confirm the effectiveness of the two proposed modules. 2) we need stacked feature to provide more global and contextual information to enable the two complementary modules for the best performance.

### D. Comparison among the Variants of AKM

We then evaluate the effectiveness of AKMs with different number of kernels. As shown in Table II, "common convs" represents the method replacing AKM with static convolution layers in a hierarchical way. "1×3" means that the AKM generates 3 adaptive kernels and directly get the semantic maps without the hierarchical structure. "5×2+1×2" represents the AKM with 5 background kernels, 5 foreground kernels and 2 kernels for each category, "5+9+1×2" represents the AKM with 5 background kernels, 9 foreground kernels and 2 kernels for each category, and the rest can be deduced.

TABLE III
COMPARISON OF THE FUSION RATIO
FOR OUTPUTS FROM TWO PROPOSED MODULES

| Method | Standing | Sitting | **CMAE** | OMAE |
|---|---|---|---|---|
| Baseline | 7.28 | 5.38 | 6.33 | 7.56 |
| $\lambda = 0.2$ | 8.03 | 7.16 | 7.60 | 8.09 |
| $\lambda = 0.4$ | 7.26 | 4.71 | **5.99** | 7.46 |
| $\lambda = 0.6$ | 6.80 | 5.22 | <u>6.01</u> | 7.04 |
| $\lambda = 0.8$ | 7.11 | 5.61 | 6.36 | 7.22 |

The result of "1x2+1x2" outperforms that of "1x3", which confirm the effectiveness of hierarchical structure inside AKM. The result under "1x2+3x2" setting becomes clear worse possibly because one kernel setting, e.g., "1x2+1x2", is enough to generate good semantic maps. Generally, the result of "9x2+1x2" achieves the best overall performance, so we set $k = 9$ in our experiments. The two experiments with more kernels for the background unexpectedly outperforms corresponding experiments with more kernels for the foreground. In our opinion, more kernels do not directly lead to better utilization of semantic information (we have tried up to 65 kernels in the paper). Therefore, although the foreground contains more semantic information, more kernels for the foreground may not bring better result.

### E. Comparison with the Fusion Ratio

We conduct experiments to explore the fusion ratio $\lambda$ for outputs from our proposed WVM and AKM. As shown in Table III, the best CMAE is achieved when $\lambda = 0.4$, so the hyperparameter $\lambda$ is set as 0.4 in our experiments.

### F. Comparison with the State-of-the-arts

We compare our DSR with state-of-the-art: FCC [2]. In Table IV, our approach outperforms the best competitor FCC by 15.6%, 6.1% and 8.4% on Standing/Sitting, Waiting/Not waiting, and Violent/Non-violent subtasks respectively. Due to the particularity of Towards/Away subtask where the images have clear low resolution (238×158) and are all grayscale, our approach does not perform as well as the other subtasks, which remains further improvement in future. Generally, our DSR has achieved a new state-of-the-art on fine-grained crowd counting benchmark, and can be served as a new solid baseline for future research.

To better understand the effectiveness of our proposed approach, we display three groups of fine-grained density maps generated by the baseline and our proposed approach in Fig. 5 (images with the same color box belong to the same group). In each group, the upper image shows the visualization of the baseline, while the lower one is for our approach. In the red-box group, it can be observed from the original picture on the right that people in the right of the red box are sitting. While the baseline misclassifies them as standing, ours can correctly classify them and people standing nearby in such a small and dense scene. In the brown-box group, the baseline judges the person at the right edge as standing, which is obviously wrong,

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART
ON FINE-GRAINED CROWD COUNTING DATASET

| Method | FCC | Baseline | Ours | Method | FCC | Baseline | Ours | Method | FCC | Baseline | Ours | Method | FCC | Baseline | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Standing | 8.36 | 7.28 | **6.58** | Towards | **1.30** | 1.43 | 1.68 | Waiting | 2.61 | 2.30 | **2.24** | Violent | 4.34 | **3.92** | 3.95 |
| Sitting | 5.56 | 5.38 | **5.15** | Away | **1.89** | 1.94 | 1.96 | Not waiting | **2.67** | 2.81 | 2.71 | Non-violent | 3.32 | 3.32 | **3.06** |
| **CMAE** | 6.96 | 6.33 | **5.87** | **CMAE** | **1.60** | 1.68 | 1.82 | **CMAE** | 2.64 | 2.56 | **2.48** | **CMAE** | 3.83 | 3.62 | **3.51** |
| OMAE | - | 7.56 | **6.68** | OMAE | - | 1.36 | **1.05** | OMAE | - | 2.69 | **2.48** | OMAE | - | **3.62** | 4.03 |

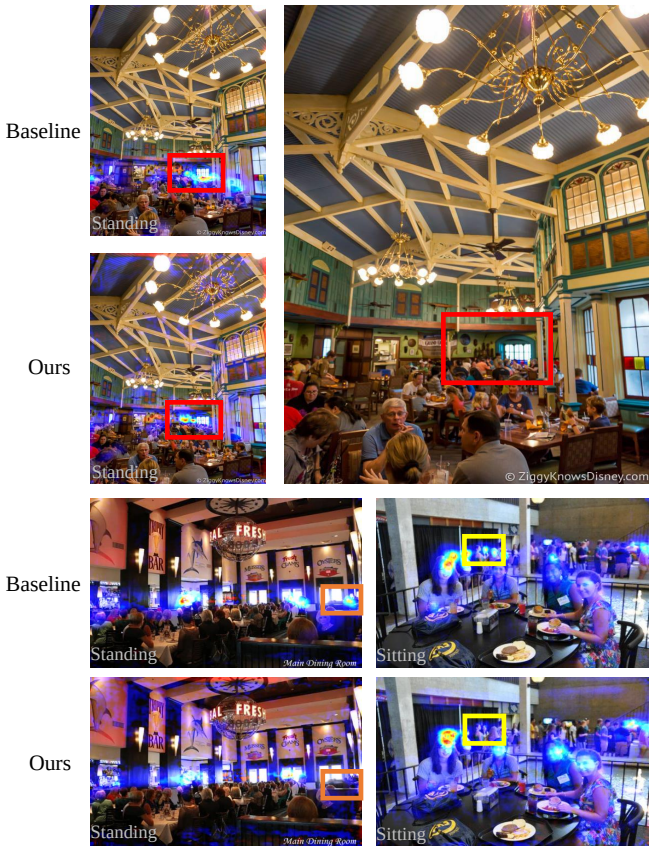


Fig. 5. The visualization of fine-grained density maps generated by the baseline and our proposed approach.

but ours makes the right decision. In the yellow-box group, some people in the queue are considered by baseline as sitting possibly due to being close to the sitting people, while ours still outputs the correct results. More visulizations are provided in the supplementary file.

## V. CONCLUSION

In this paper, we propose **DSR** (**D**iscriminative **S**emantic **R**elationship), a novel approach for fine-grained crowd counting which contains two complementary modules, Word Vector Module (WVM) and Adaptive Kernel Module (AKM). Moreover, stacked feature has been introduced to further improve the performance of both modules. By making use of the semantic information by WVM and learning to generate more reasonable adaptive kernels by AKM, we can better model the discriminative semantic relationship in fine-grained crowd counting task. Our approach generally outperforms state-of-the-art approaches and provides a new solid baseline for fine-grained crowd counting task. Worthwhile and related future work can be spawned from above technical contributions.

## REFERENCES


[1] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *2008 19th international conference on pattern recognition*. IEEE, 2008, pp. 1–4.

[2] Jia Wan, Nikil Senthil Kumar, and Antoni B Chan, "Fine-grained crowd counting," *IEEE transactions on image processing*, vol. 30, pp. 2114–2126, 2021.

[3] Bastian Leibe, Edgar Seemann, and Bernt Schiele, "Pedestrian detection in crowded scenes," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 878–885.

[4] Antoni B Chan and Nuno Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Transactions on image processing*, vol. 21, no. 4, pp. 2160–2177, 2011.

[5] Victor Lempitsky and Andrew Zisserman, "Learning to count objects in images," *Advances in neural information processing systems*, vol. 23, 2010.

[6] 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*, 2018, pp. 1091–1100.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] Hyojun Go, Junyoung Byun, Byeongjun Park, Myung-Ae Choi, Seunghwa Yoo, and Changick Kim, "Fine-grained multi-class object counting," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 509–513.

[9] Justin Kay, Catherine M Foley, and Tom Hart, "Fine-grained counting with crowd-sourced supervision," *arXiv preprint arXiv:2205.11398*, 2022.

[10] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[11] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.