Human-centric Computing and Information Sciences

November 2023 | Volume 13

www.hcisjournal.com

KIPS
Korea Information Processing Society

KIPS CSWRG
Korea Information Processing Society
Computer Software Research Group

HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2023) 13:50
DOI: https://doi.org/10.22967/HCIS.2023.13.050
Received: July 14, 2022; Accepted: August 28, 2022; Published: November 15, 2023

Crowd Counting via Attention and Multi-Feature Fused Network

Xiangyu Guo[1], Mingliang Gao[1,*], Jinfeng Pan[1], Jianrun Shang[1], Alireza Souri[2], Qilei Li[3], and Alessandro Bruno[4]

Abstract

With the rapid development of Internet of Everything and artificial intelligence techniques and massive amounts of video surveillance data, crowd counting has drawn extensive attention in computer vision. Inspired by deep learning methods, convolutional neural networks (CNN) have been dedicated to improving the effectiveness of crowd counting. As CNN is unable to capture the continuous size changes of heads in images, the large-scale variations impede the development of crowd counting. To solve this problem, this paper presents an attention and multi-feature fused network (AMFNet) containing a multi-level feature extractor and four attentional density estimator (ADE) modules. The multi-level extractor is used to extract the features of different sizes and various kinds of context information based on a deep network backbone. The existing ADE modules are built to merge different level features to generate a high-quality density map. A channel attention unit is adopted in the ADE modules to identify the head accurately. Then, four ADE modules are applied to exploit multi-level features and generate a fine-grained density map for coping with various scales. The experiment results show that the proposed AMFNet performs well in dense crowd scenarios, and that it is comparable to mainstream methods in terms of accuracy and robustness.

Keywords

Crowd Counting, Convolutional Neural Network, Attentional Mechanism, Information Fusion

1. Introduction

Crowd counting makes great sense in numerous applications, e.g., video surveillance [1], smart city governance [2–5], and public safety management [6]. Although well-developed thanks to the massive amounts of video surveillance, big data [7], and artificial intelligence (AI) techniques [8], crowd counting is limited by various challenges, e.g., perspective distortion, large-scale variations, occlusion, and background clutter.

Researchers have made unrelenting efforts to address these problems. Early work mainly concentrated on detection-based [9, 10] and regression-based [11] methods, which are effective on sparse scenarios but whose performance is seriously affected by the scale variations and occlusion in the dense crowd.

***Corresponding Author:** Mingliang Gao (mlgao@sdut.edu.cn)
[1]School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China
[2]Department of Software Engineering, Halic University, Istanbul, Turkey
[3]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
[4]Department of Business, Law, Economics, Consumer Behaviour 'Carlo A. Ricciardi', Faculty of Communication, IULM University, Milan, Italy

Nowadays, convolution neural network (CNN)-based methods [12–14] feature various architectures for crowd counting. These methods aim to regress a density map given the ground truth, where the summed pixel values are equal to the final counts.

Some methods [15, 16] enhance the counting performance by fusing multi-scale information. The attention mechanism is also utilized to improve the model performance in crowd counting [17, 18]. Despite the improved performance, these methods still have some drawbacks in dealing with large-scale variations in congested crowd scenarios. Such scale variation is attributable to the indiscriminate placement of cameras. As illustrated in Fig. 1, the heads in images or across diverse images have scale variations.

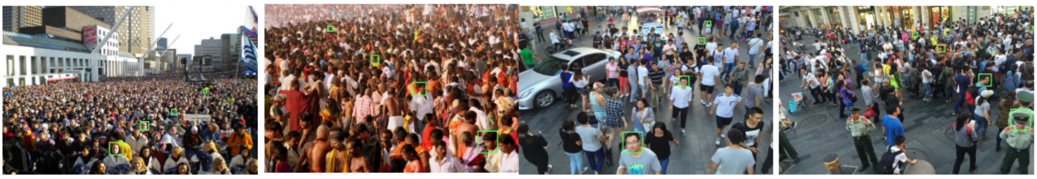


Fig. 1. Large-scale variations in crowd scenarios (the green box indicates the head region).

To tackle the large-scale variations in crowded scenarios, we propose an attention and multi-feature fused network (AMFNet) for crowd counting. The proposed AMFNet mainly consists of a multi-level feature extractor and four attentional density estimator (ADE) modules. The multi-level feature extractor is aimed at generating features having different levels and different sizes and containing various kinds of context information. The ADE module is adopted to merge different level features to output a high-quality density map. A channel attention (CA) unit is adopted in the ADE module to identify the head accurately. To exploit the multi-level features sufficiently, four ADE modules are connected serially. In sum, the contributions of this paper are as follows:

- An ADE module is designed to generate an elaborate density map for crowd counting. A CA unit is adopted in the ADE module to identify the head accurately.
- By cascading four ADE modules, the proposed AMFNet can fuse features with different levels sufficiently.
- Extensive experiments are carried out on five benchmark datasets to evaluate the counting performance in accuracy and robustness.

The rest of this paper is organized as follows: the related work is depicted in Section 2; the proposed method is described in Section 3; and the experiment discussion and conclusion are provided in Sections 4 and 5, respectively.

2. Related Work

CNN-based approaches have widely developed in the crowd counting domain [12–14], benefitting from the strong ability of feature representation. Here, we primarily analyze three types of CNN-based approaches that are closely related to the proposed AMFNet, i.e., multi-scale-based approaches, multi-level-based approaches, and attention-based approaches.

2.1 Formal Concept Analysis

These approaches are aimed at fusing multi-scale or multi-context information to resolve the large-scale variations in dense crowds. Zhang et al. [12] designed a three-branch network by utilizing various sizes of kernels to acquire features with various receptive fields. The final prediction is generated by fusing these features. Sam et al. [13] built a switching-CNN that leverages a switch classifier to tackle large-scale variations. In addition, it adopts recurrent networks to fuse features from multi-column CNN.

Chen et al. [19] proposed a multi-scale semantic refined strategy to capture more semantic features of various scales so as to address the scale variations. Sindagi and Patel [15] designed a contextual pyramid (CP)-CNN to acquire features with multiple scales. By encoding the local and global context, the network can be aware of different density classes. Gao et al. [20] designed a DULR (down, up, left, and right) module that can handle extremely dense crowds by fusing global, local, and pixel-level features.

To deal with large crowd-density variations, Sajid et al. [21] built a patch rescaling module that is beneficial in fusing multi-scale information. Dai et al. [22] proposed a dense scale network by connecting different dilated convolutional layers that could capture multi-scale information. Generally, the approaches above are difficult to train due to the cumbersome structures [23].

2.2 Multi-Level-based Approaches

These approaches take full advantage of the multi-level information of the backbone network to improve the counting performance. Liu et al. [24] designed a multi-column model embedded in structured feature enhancement modules to integrate multi-level information to address large-scale variations. Chen et al. [25] built a scale pyramid network to acquire multi-level features by using dilated convolutions with different rates. Furthermore, Sindagi and Patel [16] proposed a multi-level bottom-top and top-bottom fusion network (MBTTBF) to mix information with different levels, which is extracted from shallow to deep layers. Song et al. [26] designed a scale-adaptive selection network (SASNet) to generate feature representations with multiple levels, which can build the correspondence relation between feature levels and head scales. Wang et al. [27] built a semi-supervised multi-level auxiliator to exploit shared characteristics at multiple levels to address the scale variation. Zhu et al. [28] built a scale and level aggregation module to leverage the multi-level information. The methods above directly use multi-level information, but there is a gap between these different levels of information; thus, resulting in information loss. The proposed method first enhances low-level information and then merges it with high-level information, which is more conducive to the retention of detailed information.

2.3 Attention-based Approaches

These approaches improve counting performance by applying the visual attention mechanism to enable the counting models to concentrate on focused information deliberately. Liu et al. [29] built DecideNet to handle scenarios with varying densities. They incorporated an attention module to measure the dependability of different kinds of estimations. Amirgholipour et al. [30] proposed a pyramid density-aware attention network (PDANet). They employed a classification attention module to deliver the multi-scale features, and two decoder modules were employed to provide a two-scale density map. Wang et al. [31] built a hybrid attention module to concentrate on the discriminative region and alleviate the background clutter. Rong and Li [32] presented a coarse- and fine-grained attention network (CFANet) to restrain the background clutter and adjust weights as the crowd density levels change. Jiang et al. [17] built an attention scaling network to alleviate the performance differences in various areas, and it improves the estimations by exploiting attention masks. Lin et al. [33] merged a global, learnable local, and instance attention scheme into the network to address the large-scale variations. Wang and Breckon [34] utilized an attention layer to predict a high-resolution feature map. These attention-based approaches use an attention module to enhance the counting performance. Unlike the methods above, the proposed AMFNet improves the counting accuracy by embedding the attention unit in multiple modules.

3. The Proposed Approach

This section presents an overview of the AMFNet architecture. The ADE module is illustrated based on two main steps including coarse density map generation and head identification. Finally, a detailed description of the loss function and ground truth generation is presented.

3.1 Network Design

The structure of AMFNet is depicted in Fig. 2. It includes a multi-level feature extractor and four ADE modules. The ADE module is split into two stages: first, it takes two feature maps as input and produces a low-level feature map M_i ; the second step refines M_i and produces the high-quality density map P_i. The final prediction is generated by a convolutional operation.

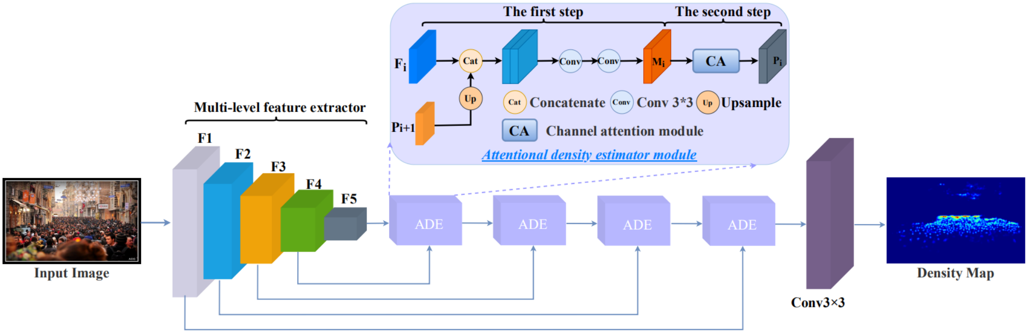


Fig. 2. Structure of the proposed AMFNet.

The first 13 layers of VGG-16 are utilized to extract multi-level features. Five basic feature maps denoted as $\{F_1\ , F_2\ \ F_3\ \ , F_4\ \ , F_5\}$ are generated hierarchically through the extractor. The ADE module is designed to produce an intermediate density map by fusing multi-level features. Moreover, four ADE modules are serially connected to exploit the multi-level features sufficiently. In this case, the final prediction M can be obtained through a convolutional operation and is formulated as:

$$M = Conv(4 \times O_f(X_h, X_l)), \tag{1}$$

where O_f denotes a function of the ADE module, with X_h and X_l as the high-level and low-level feature maps, respectively.

3.2 Attentional Density Estimator Module

The multi-level feature extractor can extract five feature maps with different levels. The high-level feature map P_{i+1} has more context information and small size, whereas the low-level feature map F_i has less context information and large size. The ADE module is designed to merge features with multiple levels.

As depicted in Fig. 2, the ADE module contains two steps. First, it can produce a coarse density map M_i and recognize a region proposal by fusing feature maps. Thus, it can mitigate errors caused by background. The first step can be formulated as:

$$M_i = Conv_2^{3\times3}(Conv_1^{3\times3}(Cat(F_i, Up(P_{i+1})))), \tag{2}$$

where Up and Cat denote the up-sample and concatenate operation, respectively. The first convolutional operation is used for integrating features, and the second convolutional operation is utilized for channel reduction. After the first process, the generated density map M_i can highlight a discriminative area. However, it causes overestimation by misidentifying other objects in the region.

The goal of the second step is to identify a head in the previously recognized region. Channel attention can be regarded as an object (head in this work) selection through the adaptive adjustment of the weight of each channel process [35]. Thus, a channel attention module is embedded in the ADE module to identify the heads.

As shown in Fig. 3, the CA unit operates on the channel dimension and outputs a channel-wise weight W that can emphasize the head features. The CA unit is calculated by:

$$P_i = M_i \otimes Sig(Conv1d(GAP(M_i))), \tag{3}$$

where P_i represents the optimized feature map and M_i is the coarse feature map. $Conv1d$ denotes a fast 1-dimension convolutional operation. Through this, the CA unit can suppress the influence of misidentifying the object.

To exploit the basic feature maps and generate an optimized density map sufficiently, four ADE modules are serially connected, and an estimated map is obtained through the final convolution layer.

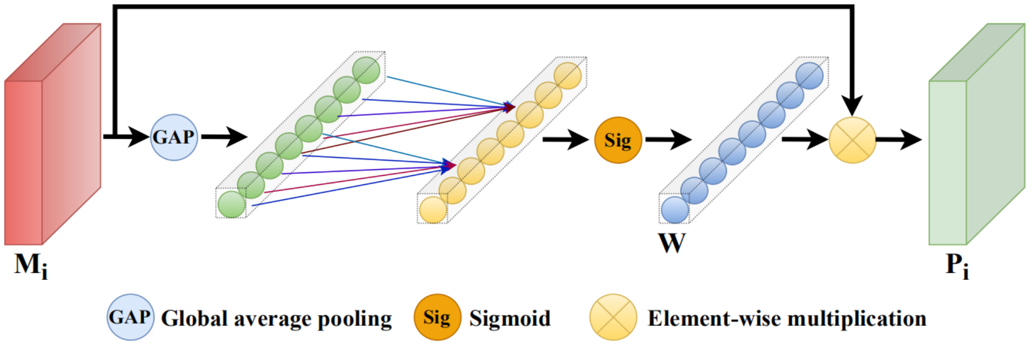


Fig. 3. Architecture of the channel attention unit.

3.3 Loss Function

The proposed AMFNet is optimized by employing the mean squared error (MSE) loss, which aims to minimize the distance between the prediction and ground truth. It can be formulated as:

$$\text{loss} = \frac{1}{N}\sum_{i=1}^{N}\|E(X^i;\theta) - GT^i\|_2^2, \tag{4}$$

where N represents the batch size, X^i denotes the i-th input image, $E(X^i;\theta)$ and GT^i are the predicted density map and the corresponding ground truth, respectively, and θ is a series of parameters to be learned during training.

3.4 Ground Truth Map Generation

The geometry-adaptive Gaussian kernel is employed to cope with the dense crowds [12]. Each labeled head h_i is blurred by using a Gaussian kernel. Suppose each annotated head is represented as $\delta(h - h_i)$. Then, a normalized Gaussian kernel gets convolved with the delta function and generates the Gaussian density map. In a nutshell, it is formulated as:

$$D_{gt} = \sum_{i=1}^{H}\delta(h - h_i) * G_{\sigma_i}(h),\ \sigma_i = \beta \overline{d_\iota}, \tag{5}$$

where H is the number of head annotations and h represents the position pixel. $\delta(h - h_i)$ depicts a target head, and $\overline{d_\iota}$ indicates the mean distance of k-nearest neighbors (k=3 in this work). σ_i represents the variance that is positively correlated with $\overline{d_\iota}$, and β is a hyperparameter and is set as 0.3 in the work.

4. Experiments and Discussion

4.1 Evaluation Metrics

Similar to other methods, we employ the most widely used mean absolute error (MAE) and root mean square error (RMSE) to measure the counting performance. They are formulated as:

$$MAE = \frac{1}{C}\sum_{j=1}^{C}\left|\widehat{y_j} - y_j\right|, \tag{6}$$

$$RMSE = \frac{1}{C}\sqrt{\sum_{j=1}^{C}\left\|\widehat{y_j} - y_j\right\|^2}, \tag{7}$$

where C represents the number of test samples, $\widehat{y_j}$ denotes the predicted count of the j-th image, and y_j denotes the ground truth count of the j-th image. They can reflect the accuracy and robustness of the model.

4.2 Datasets

The ShanghaiTech dataset [12] contains two parts: SHHA and SHHB. The SHHA sub-dataset is arbitrarily captured from the web, presenting dense scenes, whereas the SHHB sub-dataset comes from the street in Shanghai and presents sparse crowds.

The UCF-QNRF [36] contains 1,535 high-quality images with 1.25 million head annotations. The images present a broader variety of scenes in an extremely dense crowd.

The UCF_CC_50 dataset [37] includes 50 images collected from the web, and the head annotations per image vary from 94 to 4,543. As there are limited training samples and the crowds are extremely dense, it is a challenging crowd dataset.

The WorldExpo'10 dataset [38] was captured by 108 cameras during the Shanghai WorldExpo 2010. It has a total number of 9,920 frames, from which 199,923 heads are annotated. The test set is divided into five parts, and each part has a region-of-interest.

NWPU-Crowd is composed [39] of 5,109 images with more than 2M head annotations. Compared with the aforementioned datasets, it is much more challenging as it contains negative instances, high resolution (2311×3383), and large-scale variation.

4.3 Implementation Details

To ensure that the proposed network is sufficiently trained during the training stage, we randomly crop the original images and flip them horizontally. The training and test are conducted on two NVIDIA RTX3060 GPUs in the PyTorch framework. The backbone network is the pre-trained VGG-16 from ImageNet. For the WorldExpo'10 dataset, the training size is set as 512×672, and other datasets are set as 576×768. As the NWPU-Crowd dataset has higher resolution compared with other datasets, a large batch size may lead to out-of-memory during the training stage. The batch size of NWPU-Crowd is set as 4, and the batch size is set as 8 for other datasets. The Adam [40] algorithm is employed to optimize the model. Learning rate L_r is initially set as 10^{-5}, and decay rate D_r is set as 0.995. Training epoch T is set as 1,500. In a nutshell, the pseudocode of AMFNet is depicted in Algorithm 1.

Algorithm 1. Pseudocode of the proposed AMFNet

```
Input: Original image
Initialize $L_r$, $D_r$, $B$, and $T$.
Extract features $\{F_1,F_2,F_3,F_4,F_5\}$ from VGG-16,
function1 CA unit
optimized feature $P_i = M_i \otimes Sig(Conv1d(GAP(M_i)))$ ,
end function1
function2 ADE module
```

```
        feature1    $M_i = Conv_2^{3\times3}(Conv_1^{3\times3}(Cat(F_i, Up(P_{i+1}))))$,
        feature2    $P_i = function1(M_i)$,
    end function2
    for $i$ = 1:4
        initial feature $X_h, X_l = function2(extract\_feature)$
        estimated map    $M = Conv(O_f(X_h, X_l))$
    End
    return final map $M$
Output: Estimated density map $M$
```

4.4 Comparative Analysis

To verify the effectiveness of the proposed AMFNet, comparative experiments are conducted with state-of-the-art methods. The experiment results are presented in Tables 1 and 2.

For the SHHA sub-dataset, AMFNet ranks first in MAE (66.8) and third in RMSE (107.6). For the SHHB sub-dataset, the proposed method ranks first in MAE (7.7) and RMSE (12.2) compared with the competitors. The visual results are depicted in Fig. 4. The first, second, and third rows denote the original image, ground truth, and estimated map, respectively. This demonstrates that the counting value obtained by the proposed method is very close to the ground truth in both congested scene (SHHA) and sparse scene (SHHB).

For the UCF_CC_50 dataset, AMFNet scores 217.3 and 354.6 in MAE and RMSE, respectively. Compared with CSRNet [41], which also settled down to the scale variation in congested scenes, it improves MAE and RMSE by 18.4% and 9.5%, respectively. The subjective evaluations are shown in Fig. 5. This proves that AMFNet performs well in scenarios with large-scale variations.

For the UCF-QNRF dataset, the proposed AMFNet outperforms all the competitors. Specifically, AMFNet reduces MAE and RMSE by 28.1% and 20.8%, respectively, compared with PCCNet [21], which also adopts the method of multi-level information fusion. Furthermore, although the images in UCF-QNRF have large-scale variation, the result also proves that the proposed AMFNet can cope well with this problem.

Table 1. Comparative results on the SHHA, SHHB, UCF_CC_50, QNRF, and NWPU-Crowd datasets

Method	SHHA		SHHB		UCF_CC_50		QNRF		NWPU-Crowd	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Zhang et al. [38]	181.8	277.7	32.0	49.8	467.0	498.5	-	-	-	-
MCNN [12]	110.2	173.2	26.4	41.3	3776	509.1	277.0	426.0	232.5	714.6
CMTL [14]	101.3	152.4	20.0	31.1	322.8	341.4	252.0	514.0	-	-
A-CCNN [42]	85.4	124.6	19.2	31.5	367.3	423.7	-	-	176.5	520.6
Switch-CNN [13]	90.4	135.0	21.1	30.1	318.1	439.2	228.0	445.0	-	-
SaCNN [43]	83.8	139.2	16.2	25.8	314.9	424.8	-	-	-	-
MobileCount [44]	81.4	133.3	8.1	12.7	284.5	421.2	117.9	207.6	-	-
ACM-CNN [45]	72.2	**103.5**	17.5	22.7	291.6	320.9	-	-	-	-
PCCNet [20]	73.5	124.0	11.0	19.0	240.0	315.5	148.7	247.3	-	-
ic-CNN [46]	68.5	116.2	10.7	16.0	266.1	397.5	-	-	-	-
IG-CNN [47]	72.5	118.2	13.6	21.1	291.4	349.4	-	-	-	-
CSRNet [41]	68.2	115.0	10.6	16.0	266.1	397.5	-	-	121.3	387.4
DecideNet [29]	-	-	20.8	29.4	-	-	-	-	-	-
SANet [48]	67.0	104.5	8.4	13.6	258.4	**334.9**	-	-	190.6	491.4
AMFNet (proposed)	**66.8**	107.6	**7.7**	**12.2**	**217.3**	354.6	**106.8**	**195.9**	**115.2**	**379.3**

The best results are marked in bold.

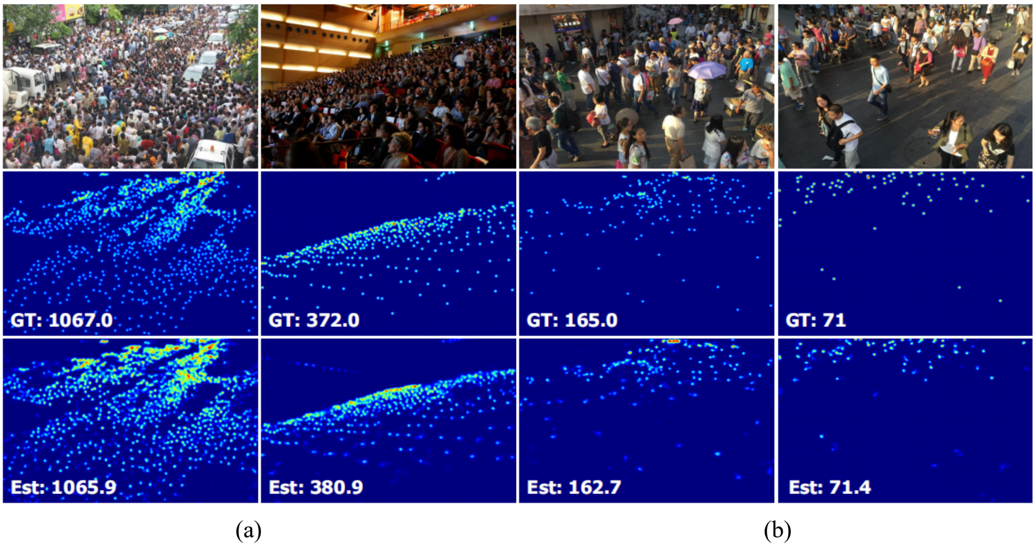


Fig. 4. Visualization of the estimated results on the ShanghaiTech dataset: (a) SHHA and (b) SHHB.

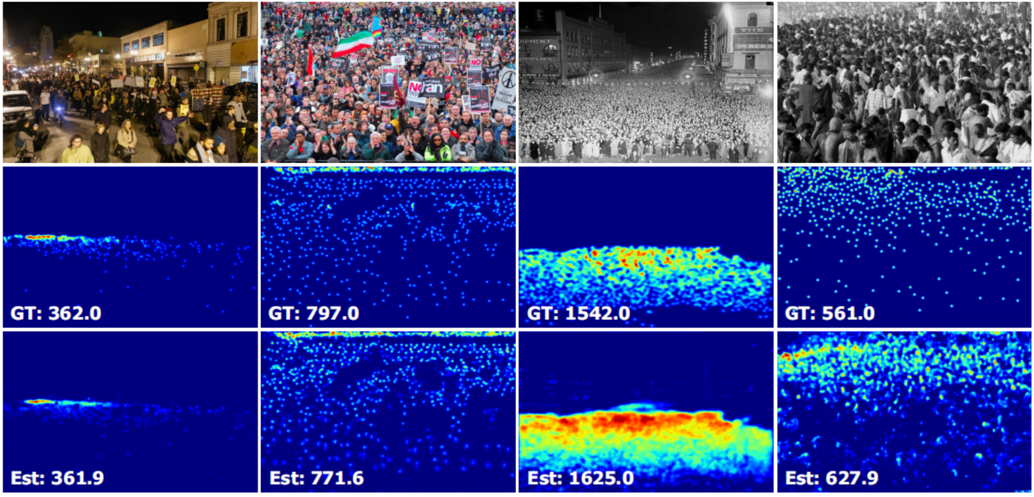
Table 2. Comparative results on the WorldExpo'10 dataset

Method	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	MAE (Avg.)
Zhang et al. [38]	9.8	14.1	14.3	22.4	3.7	12.9
MCNN [12]	3.4	20.6	12.9	13.0	8.1	11.6
CMTL [14]	3.8	32.3	19.5	20.6	6.6	16.6
Switch-CNN [13]	4.4	15.7	10.0	11.0	5.9	9.4
SaCNN [43]	2.6	13.5	10.6	12.5	3.3	8.5
ACM-CNN [45]	2.4	**10.4**	11.4	15.6	3.0	8.6
PCCNet [20]	1.9	18.3	10.5	13.4	3.4	9.5
ic-CNN [46]	17.0	12.3	9.2	**8.1**	4.7	20.3
IG-CNN [47]	2.6	16.1	10.2	20.2	7.6	11.3
CSRNet [41]	2.9	11.5	**8.6**	16.6	3.4	8.6
DecideNet [29]	2.0	13.1	8.9	17.4	4.8	9.2
SANet [48]	2.6	13.2	9.0	13.3	3.0	8.2
AMFNet (proposed)	**0.8**	12.0	9.0	9.1	**2.9**	**6.7**

The best results are marked in bold.

The proposed method performs best in Scene 1 (sparse crowd) and Scene 5 (congested crowd) with MAE score of 0.8 and 2.9, respectively. Meanwhile, the average MAE is achieved using the best overall methods with a score of 6.7, demonstrating that the overall performance of the proposed AMFNet is remarkable compared to other methods across various scenes. The qualitative results on the WorldExpo'10 dataset are depicted in Fig. 6. The visualization shows that AMFNet performs well in different crowd scenarios.

On the NWPU-Crowd dataset, the proposed AMFNet scores 115.2 and 379.3 in MAE and RMSE, respectively, showing an improvement of 39.5% and 22.8 % in MAE and RMSE, respectively, compared with SANet [48], which also extracts multi-level features for crowd counting. The visualized results are shown in Fig. 7. This proves that AMFNet can suppress the influence of background.



(a) (b)

Fig. 5. Visualization of the estimated results on the (a) UCF-QNRF and (b) UCF_CC_50 datasets.

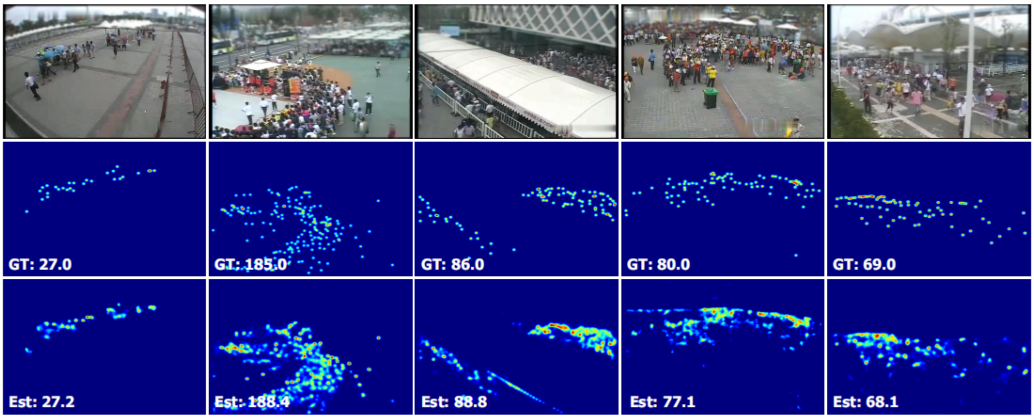


Fig. 6. Visualization of the estimated results on the WorldExpo'10 dataset.

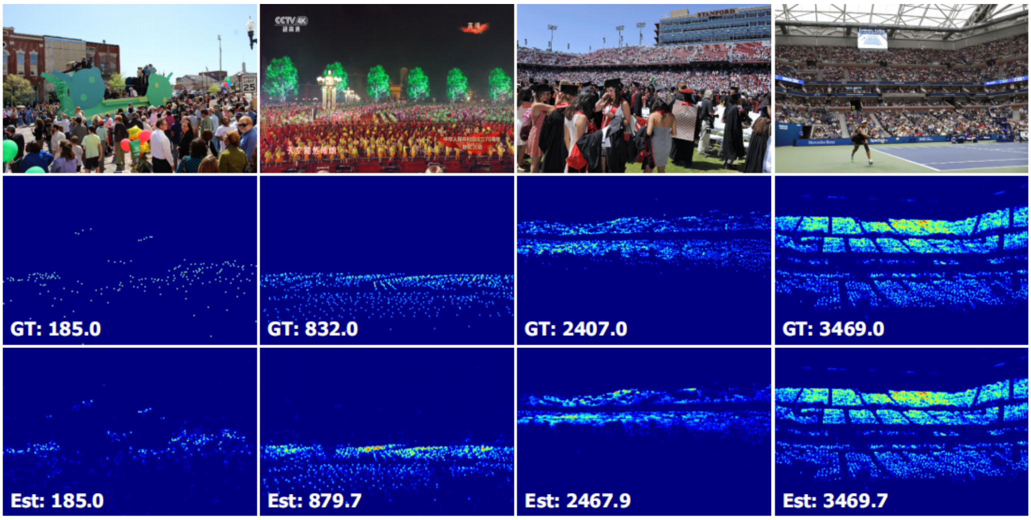


Fig. 7. Visualization of the estimated results on the NWPU-Crowd dataset.

4.5 Ablation Study

To verify the effect of various configurations and modules on crowd counting performance, extensive ablation experiments are conducted on the SHHA dataset. The experiments explore the proposed model from two aspects: the ADE module and the CA unit. The corresponding items are listed as follows:

1. “Baseline” denotes the basic model that adopts only VGG-16.
2. “Baseline + ADE(1)” represents adding one ADE module to item 1.
3. “Baseline + ADE(2)” refers to adding two ADE modules to item 1.
4. “Baseline + ADE(3)” denotes adding three ADE modules to item 1.
5. “Baseline + ADE(4)” represents the proposed AMFNet.
6. “Baseline + ADE(*n*) (w/o CA)” denotes adding *n* ADE modules without the CA module to the “baseline.”

Table 3 shows that the baseline model scores 81.7 and 126.7 in MAE and RMSE, respectively, which are the worst across all entries in the table. After adding an ADE module to the baseline model, MAE and RMSE decrease by 6.0% and 2.8%, respectively. This demonstrates that the ADE module is helpful in enhancing the counting performance. When the number of ADE modules reaches 4, MAE and RMSE reach the lowest scores of 66.8 and 107.6, respectively. This proves that fusing the features of different scales is essential in enhancing the performance of counting heads. When the CA module is removed from the ADE module, the scores of MAE and RMSE decrease as shown in the even rows (2, 4, 6, and 8) in Table 3. This proves that the CA unit is beneficial in boosting the counting accuracy.

Table 3. Ablation analysis of the key components in AMFNet

Methods	MAE	RMSE
Baseline	81.7	126.7
Baseline + ADE(1) (w/o CA)	76.8	123.2
Baseline + ADE(1)	75.1	122.0
Baseline + ADE(2) (w/o CA)	74.0	120.7
Baseline + ADE(2)	73.3	118.8
Baseline + ADE(3) (w/o CA)	71.3	116.6
Baseline + ADE(3)	70.6	111.7
Baseline + ADE(4) (w/o CA)	68.8	109.2
Baseline + ADE(4) (proposed)	**66.8**	**107.6**

The best results are marked in bold.

5. Conclusion

This study presented an AMFNet to address the large-scale variations in crowd counting. The key component in AMFNet is the ADE module, which can fuse different level features and produce an elaborate density map. By cascading four ADE modules, various features are sufficiently used for information fusion. The experiment results illustrate that the AMFNet outperforms many related methods in terms of accuracy and robustness. In future work, other evaluation factors such as semantic interoperability and crowd mobility shall be explored to improve the extraction quality of crowd counting. Likewise, meta-heuristic algorithms can be applied for improving feature selection methods in attention-based density estimation.

Author’s Contributions

Conceptualization, XG, MG. Funding acquisition, MG. Investigation and methodology, XG, JS. Project administration, MG, AB. Resources, JP, AS. Supervision, MG, XG, QL. Writing of the original draft, XG, QL, JS. Writing of the review and editing, XG, MG, QL. Software, MG, AS, AB. Validation,

XG, JP, AS. Formal analysis, XG, MG, AB. Data curation, XG, MG. Visualization, XG, JS, AS.

Funding

This work is supported by the National Natural Science Foundation of China (No. 61601266 and 61801272) and the National Natural Science Foundation of Shandong Province (No. ZR2021QD041 and ZR2020MF127).

Competing Interests

The authors declare that they have no competing interests.

References

[1] S. Wen, X. Zhang, R. Cao, B. Li, Y. Li, "MSSRM: a multi-embedding based self-attention spatio-temporal recurrent model for human mobility prediction," *Human-Centric Computing and Information Sciences*, vol. 11, article no. 37, 2021. https://doi.org/10.22967/HCIS.2021.11.037

[2] J. H. Park, M. M. Salim, J. H. Jo, J. C. S. Sicato, S. Rathore, and J. H. Park, "CIoT-Net: a scalable cognitive IoT based smart city network architecture," *Human-centric Computing and Information Sciences*, vol. 9, article no. 29, 2019. https://doi.org/10.1186/s13673-019-0190-9

[3] S. K. Singh, A. E. Azzaoui, T. W. Kim, Y. Pan, and J. H. Park, "DeepBlockScheme: a deep learning-based blockchain driven scheme for secure smart city," *Human-centric Computing and Information Sciences*, vol. 11, article no. 12, 2021. https://doi.org/10.22967/HCIS.2021.11.012

[4] J. Vanus, P. Kucera, R. Martinek, and J. Koziorek, "Development and testing of a visualization application software, implemented with wireless control system in smart home care," *Human-centric Computing and Information Sciences*, vol. 4, article no. 18, 2014. https://doi.org/10.1186/s13673-014-0019-5

[5] H. Jo and Y. I. Yoon, "Intelligent smart home energy efficiency model using artificial TensorFlow engine," *Human-centric Computing and Information Sciences*, vol. 8, article no. 9, 2018. https://doi.org/10.1186/s13673-018-0132-y

[6] L. Zhao, Y. Zhang, and Y. Cui, "A multi-scale U-shaped attention network-based GAN method for single image dehazing," *Human-centric Computing and Information Sciences*, vol. 11, article no. 38, 2021. https://doi.org/10.22967/HCIS.2021.11.038

[7] D. Cao, X. Ren, M. Zhu, and W. Song, "Visual question answering research on multi-layer attention mechanism based on image target features," *Human-centric Computing and Information Sciences*, vol. 11, article no. 11, 2021. https://doi.org/10.22967/HCIS.2021.11.011

[8] M. Malekshahi Rad, A. M. Rahmani, A. Sahafi, and N. Nasih Qader, "Social Internet of Things: vision, challenges, and trends," *Human-centric Computing and Information Sciences*, vol. 10, article no. 52, 2020. https://doi.org/10.1186/s13673-020-00254-6

[9] Y. Verdie and F. Lafarge, "Detecting parametric objects in large scenes by Monte Carlo sampling," *International Journal of Computer Vision*, vol. 106, pp. 57-75, 2014. https://doi.org/10.1007/s11263-013-0641-0

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017. https://doi.org/10.1109/TPAMI.2016.2577031

[11] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: counting people without people models or tracking," in *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-7. https://doi.org/10.1109/CVPR.2008.4587569

[12] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 589-597. https://doi.org/10.1109/CVPR.2016.70

[13] D. B. Sam, S. Surya, and R. Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 4031-4039. https://doi.org/10.1109/CVPR.2017.429

[14] V. A. Sindagi and V. M. Patel, V. M. "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proceedings of 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, 2017, pp. 1-6. https://doi.org/10.1109/AVSS.2017.8078491

[15] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 1879-1888. https://doi.org/10.1109/ICCV.2017.206

[16] V. A. Sindagi and V. M. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 1002-1012. https://doi.org/10.1109/ICCV.2019.00109

[17] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, "Attention scaling for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4705-4714. https://doi.org/10.1109/CVPR42600.2020.00476

[18] Y. Zhang, C. Zhou, F. Chang, A. C. Kot, and W. Zhang, "Attention to head locations for crowd counting," in *Image and Graphics*. Cham, Switzerland: Springer, 2019, pp. 727-737. https://doi.org/10.1007/978-3-030-34110-7_61

[19] J. Chen, K. Wang, W. Su, and Z. Wang, "SSR-HEF: crowd counting with multiscale semantic refining and hard example focusing," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6547-6557, 2022. https://doi.org/10.1109/TII.2022.3160634

[20] J. Gao, Q. Wang, and X. Li, "PCC Net: perspective crowd counting via spatial convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3486-3498, 2020. https://doi.org/10.1109/TCSVT.2019.2919139

[21] U. Sajid, W. Ma, and G. Wang, "Multi-resolution fusion and multi-scale input priors based crowd counting," in *Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 5790-5797. https://doi.org/10.1109/ICPR48806.2021.9412406

[22] F. Dai, H. Liu, Y. Ma, X. Zhang, and Q. Zhao, "Dense scale network for crowd counting," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, Taipei, Taiwan, 2021, pp. 64-72. https://doi.org/10.1145/3460426.3463628

[23] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "CNN-based density estimation and crowd counting: a survey," 2020 [Online]. Available: https://arxiv.org/abs/2003.12783.

[24] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 1774-1783. https://doi.org/10.1109/ICCV.2019.00186

[25] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, 2019, pp. 1941-1950. https://doi.org/10.1109/WACV.2019.00211

[26] Q. Song, C. Wang, Y. Wang, Y. Tai, C. Wang, J. Li, J. Wu, and J. Ma, "To choose or to fuse? scale selection for crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2576-2583, 2021. https://doi.org/10.1609/aaai.v35i3.16360

[27] M. Wang, H. Cai, X. Han, J. Zhou, and M. Gong, "STNet: scale tree network with multi-level auxiliator for crowd counting," *IEEE Transactions on Multimedia*, vol. 25, pp. 2074-2084, 2022. https://doi.org/10.1109/TMM.2022.3142398

[28] F. Zhu, H. Yan, X. Chen, T. Li, and Z. Zhang, "A multi-scale and multi-level feature aggregation network for crowd counting," *Neurocomputing*, vol. 423, pp. 46-56, 2021. https://doi.org/10.1016/j.neucom.2020.09.059

[29] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "DecideNet: counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 5197-5206. https://doi.org/10.1109/CVPR.2018.00545

[30] S. Amirgholipour, W. Jia, L. Liu, X. Fan, D. Wang, and X. He, "PDANet: pyramid density-aware attention based network for accurate crowd counting," *Neurocomputing*, vol. 451, pp. 215-230, 2021. https://doi.org/10.1016/j.neucom.2021.04.037

[31] F. Wang, J. Sang, Z. Wu, Q. Liu, and N. Sang, "Hybrid attention network based on progressive embedding scale-context for crowd counting," *Information Sciences*, vol. 591, pp. 306-318, 2022. https://doi.org/10.1016/j.ins.2022.01.046

[32] L. Rong and C. Li, "Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Salt Lake City, UT, 2021, pp. 3674-3683. https://doi.org/10.1109/CVPR.2018.00545

[33] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting crowd counting via multifaceted attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, 2022, pp. 19628-19637. https://doi.org/10.1109/CVPR52688.2022.01901

[34] Q. Wang and T. P. Breckon, "Crowd counting via segmentation guided attention networks and curriculum loss," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15233-15243, 2022. https://doi.org/10.1109/TITS.2021.3138896

[35] M. H. Guo, T. X. Xu, J. J. Liu, Z. N. Liu, P. T. Jiang, T. J. Mu, et al., "Attention mechanisms in computer vision: a survey," 2021 [Online]. Available: https://arxiv.org/abs/2111.07624.

[36] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 532-546.

[37] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 2547-2554. https://doi.org/10.1109/CVPR.2013.329

[38] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 833-841. https://doi.org/10.1109/CVPR.2015.7298684

[39] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: a large-scale benchmark for crowd counting and localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2141-2149, 2021. https://doi.org/10.1109/TPAMI.2020.3013269

[40] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[41] Y. Li, X. Zhang, and D. Chen, "CSRNet: dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 1091-1100. https://doi.org/10.1109/CVPR.2018.00120

[42] S. Amirgholipour, X. He, W. Jia, D. Wang, and M. Zeibots, "A-CCNN: adaptive CCNN for density estimation and crowd counting," in *Proceedings of 2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 948-952. https://doi.org/10.1109/ICIP.2018.8451399

[43] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, 2018, pp. 1113-1121. https://doi.org/10.1109/WACV.2018.00127

[44] P. Wang, C. Gao, Y. Wang, H. Li, and Y. Gao, "MobileCount: an efficient encoder-decoder framework for real-time crowd counting," *Neurocomputing*, vol. 407, pp. 292-299, 2020. https://doi.org/10.1016/j.neucom.2020.05.056

[45] Z. Zou, Y. Cheng, X. Qu, S. Ji, X. Guo, and P. Zhou, "Attend to count: crowd counting with adaptive capacity multi-scale CNNs," *Neurocomputing*, vol. 367, pp. 75-83, 2019. https://doi.org/10.1016/j.neucom.2019.08.009

[46] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 278-293.

[47] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, "Divide and grow: capturing huge diversity in crowd images with incrementally growing CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 3618-3626. https://doi.org/10.1109/CVPR.2018.00381

[48] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 757-773.