Efficient Crowd Counting via Structured Knowledge Transfer

Lingbo Liu
Sun Yat-Sen University
liulingb@mail2.sysu.edu.cn

Jiaqi Chen
Sun Yat-Sen University
jadgechen@gmail.com

Hefeng Wu∗
Sun Yat-Sen University
wuhefeng@gmail.com

Tianshui Chen
DarkMatter AI Research
tianshuichen@gmail.com

Guanbin Li
Sun Yat-Sen University
liguanbin@mail.sysu.edu.cn

Liang Lin
Sun Yat-Sen University
DarkMatter AI Research
linliang@ieee.org

ABSTRACT
Crowd counting is an application-oriented task and its inference efficiency is crucial for real-world applications. However, most previous works relied on heavy backbone networks and required prohibitive run-time consumption, which would seriously restrict their deployment scopes and cause poor scalability. To liberate these crowd counting models, we propose a novel Structured Knowledge Transfer (SKT) framework, which fully exploits the structured knowledge of a well-trained teacher network to generate a lightweight but still highly effective student network. Specifically, it is integrated with two complementary transfer modules, including an Intra-Layer Pattern Transfer which sequentially distills the knowledge embedded in layer-wise features of the teacher network to guide feature learning of the student network and an Inter-Layer Relation Transfer which densely distills the cross-layer correlation knowledge of the teacher to regularize the student’s feature evolution. Consequently, our student network can derive the layer-wise and cross-layer knowledge from the teacher network to learn compact yet effective features. Extensive evaluations on three benchmarks well demonstrate the effectiveness of our SKT for extensive crowd counting models. In particular, only using around 6% of the parameters and computation cost of original models, our distilled VGG-based models obtain at least 6.5× speed-up on an Nvidia 1080 GPU and even achieve state-of-the-art performance. Our code and models are available at https://github.com/HCPLab-SYSU/SKT.

CCS CONCEPTS
• Computing methodologies → Machine learning; • Applied computing → Surveillance mechanisms.

KEYWORDS
crowd counting; knowledge transfer; network compression and acceleration

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

 ACM ’20, October 12–16, 2020, Seattle, WA, USA
 © 2020 Association for Computing Machinery.
 ACM ISBN 978-1-4503-7988-5/20/10...
 https://doi.org/10.1145/3394171.3413938

1 INTRODUCTION
Crowd counting, whose objective is to automatically estimate the total number of people in surveillance scenes, is an important issue of crowd analysis [27, 70]. With the rapid growth of urban population and the increasing demand for security analysis and early warning in large-scale crowded scenarios, this task has attracted extensive interest in academic and industrial fields, due to its wide-ranging applications in video surveillance [74, congestion alerting [50] and traffic prediction [33, 34].

Recently, deep neural networks [4, 6, 10, 32, 35, 41, 44, 58, 68, 71, 72, 76] have become mainstream in the task of crowd counting and have made remarkable progress. To acquire better performance, most of the state-of-the-art methods [13, 28, 31, 36, 40, 62, 66] utilized heavy backbone networks (such as the VGG model [56]) to extract features. Nevertheless, requiring large computation cost and running at low speeds, these models are exceedingly inefficient.

As shown in Table 1, the latest DISSNet [31] costs 3.7 seconds on an Nvidia 1080 GPU and 379 seconds on an Intel Xeon CPU to process an input image of size 2032×2912. This would seriously restrict their deployment scopes and cause poor scalability, particularly on edge computing devices [1] with limited computing resources. Moreover, to handle citywide surveillance videos in real-time, we may need thousands of high-performance GPUs, which are expensive and energy-consuming. Under these circumstances, a cost-effective model is extremely desired for crowd counting.

Thus, one fundamental question is how we can acquire an efficient crowd counting model from existing well-trained but heavy networks. A series of efforts [5, 8, 14, 37, 61] have been made to compress and speed-up deep neural networks. However, most of them either require cumbersome hyper-parameters search (e.g., the sensitivity in per layer for parameters pruning) or rely on specific hardware platforms (e.g., weight quantization and half-precision floating-point computing). Recently, knowledge distillation [19], which trains a small student network to acquire the knowledge of a complex teacher network, has become a desirable alternative for model compression due to its broad applicability areas. Numerous works [9, 42, 46, 49, 69, 75] have verified its effectiveness for image compression.
classification. However, it is difficult to apply knowledge distillation to the more challenging dense-labeling crowd counting, since this task requires the distilled models to maintain the discriminative abilities at every location. In this case, extensive knowledge is desired to be transferred to well equip the abilities of student networks for generating accurate crowd density maps.

To fully distill the knowledge of a pre-trained teacher network, we conduct a more in-depth analysis of the crowd counting models and observe that the structured knowledge of deep networks is implicitly embedded in i) layer-wise features that carry image content, and ii) cross-layer correlations that encode feature updating schemas. In this work, we develop a novel framework termed Structured Knowledge Transfer (SKT), which simultaneously distills the knowledge of layer-wise features and cross-layer correlation via two complementary transfer modules, i.e., an Intra-Layer Pattern Transfer (Intra-PT) and an Inter-Layer Relation Transfer (Inter-RT). First, our Intra-PT takes a set of representative features extracted from a well-trained teacher network to sequentially supervise the corresponding features of a student network, analogous to using the teacher’s knowledge to progressively correct the student’s learning deviation. As a result, the features of the student network exhibit similar visual or semantic patterns of its supervisor. Second, our Intra-PT densely computes the relationships between pairwise features of the teacher network and then utilizes such knowledge to help the student network regularize the long short-term evolution of its hierarchical features. Thereby, the student network can learn the solution procedure flow of its teacher. Thanks to the tailor-designed SKT framework, our lightweight student network can effectively learn compact and knowledgeable features, yielding high-quality crowd density maps.

In experiments, we apply the proposed SKT framework to compress and accelerate a series of existing crowd counting models (e.g., CSRNet [28], BL [40] and SANet [6]). Extensive evaluations on three representative benchmarks greatly demonstrate the effectiveness of our method. Under the condition of occupying nearly 6% of the original model parameters and computational cost, our distilled VGG-based models obtain at least 6.5x speed-up on GPU and 9x speed-up on CPU. Moreover, these lightweight models can preserve competitive performance, and even achieve state-of-the-art results on Shanghaitech [76] Part-A and UCF-QNRF [22] datasets. In summary, the major contributions of this work are three-fold:

- We propose a general and comprehensive Structured Knowledge Transfer framework, which can generate lightweight but effective crowd counting models. To the best of our knowledge, we are the first to focus on improving the efficiency of crowd counting models.
- Two cooperative knowledge transfers (i.e., an Intra-Layer Pattern one and an Inter-Layer Relation one) are incorporated to fully distill the structured knowledge of well-trained models to lightweight models for crowd counting.
- Extensive experiments on three benchmarks show the effectiveness of our method. In particular, our distilled VGG-based models have an order of magnitude speed-up while achieving state-of-the-art performance.

### 2 RELATED WORKS

**Crowd Counting:** Crowd counting has been extensively studied for decades. Early works [11, 26] estimated the crowd count by directly locating the people with pedestrian detectors. Subsequently, some methods [7, 47] learned a mapping between handcrafted features and crowd count with regressors. Only using image low-level information, these methods had high efficiencies, but their performance was far from satisfactory for real-world applications.

Recently, we have witnessed the great success of convolutional neural networks [29, 32, 43, 48, 53, 55, 63, 66, 71, 72] in crowd counting. Most of these previous approaches focused on how to improve the performance of deep models. To this end, they tended to use heavy backbone networks (e.g., the VGG model [56]) to extract representative features. For instance, Li et al. [28] combined a VGG-16 based front-end network and diluted convolutional back-end network to learn hierarchical features for crowd counting. Liu et al. [36] introduced an expanded context-aware network that learned both image features and geometry features with two truncated VGG16 models. Liu et al. [31] utilized three paralleled VGG16 networks to extract multiscale features and then conducted structured refinements. Recently, Ma et al. [40] proposed a new Bayesian loss for crowd counting and verified its effectiveness on VGG19. Although the aforementioned methods can make impressive progress, their performance advantages come with the cost of burdensome computation. Thus it is hard to directly apply these methods to practical applications. In contrast, we take into consideration both the performance and computation cost. In this work, we aim to improve the efficiency of existing crowd counting models under the condition of preserving the performance.

**Model Compression:** Parameters quantization [12], parameters pruning [14] and knowledge distillation [19] are three types of commonly-used algorithms for model compression. Specifically, quantization methods [23, 77] compress networks by reducing the number of bits required to represent weights, but they usually rely on specific hardware platforms. Pruning methods [17, 25, 78] removed redundant weights or channels of layers. However, most of them used weight masks to simulate the pruning and mass post-processing are needed to achieve real speed-up. By contrast, knowledge distillation [42, 49, 75] is more general and its objective is to transfer knowledge from a heavy network to a lightweight one. Recently, it has been widely studied. For instance, Hinton et al. [19]
trained a distilled network with the soft output of a large highly regularized network. Romero et al. [46] improved the performance of student networks with both the outputs and the intermediate features of teacher networks. Zagoruyko and Komodakis [69] utilized activation-based and gradient-based spatial attention maps to transfer knowledge between two networks. Nevertheless, most of these previous methods were proposed for image classification. Recently, for fast human pose estimation, Zhang et al. [73] directly used the final pose maps of teacher networks to supervise the interim and final results of student networks. [16] and [39] distilled the knowledge embedded in layer-wise features for semantic segmentation, but both of them neglected the the cross-layer correlation knowledge. In contrast, our method fully explores the structured knowledge (e.g., layer-wise one and across-layer one) of teacher networks to optimize student networks for compact and effective feature learning.

3 METHOD

In this work, a general Structured Knowledge Transfer (SKT) framework is proposed to address the efficiency problem of existing crowd counting models. Its architecture is shown in Fig. 1. Specifically, an Intra-Layer Pattern Transfer (Intra-PT) and an Inter-Layer Relation Transfer (Inter-RT) are incorporated into our framework to fully transfer the structured knowledge (e.g., layer-wise one and cross-layer one) from the teacher networks to the student networks.

In this section, we take the VGG16-based CSRNet [28] as an example to introduce the working modules of our SKT framework. The student network is a 1/n-CSRNet, in which the channel number of each convolutional layer (but except the last layer) is 1/n of the original one in CSRNet. Compared with the heavy CSRNet, the lightweight 1/n-CSRNet model only has 1/n^2 parameters and computation cost, but it suffers serious performance degradation. Thus, our objective is to improve the performance of 1/n-CSRNet as far as possible by transferring the knowledge of CSRNet. Notice that our SKT is general and it is also applicable to other crowd counting models (e.g., BL [40] and SANet [6]). Several distilled models are analyzed and compared in Section 4.

3.1 Feature Extraction

In general, teacher networks have been pre-trained on standard benchmarks. The learned knowledge can be explicitly represented as parameters or implicitly embedded into features. Similar to the previous works [46, 69], we perform knowledge transfer on the feature level. The knowledge in the features of teacher networks is treated as supervisory information and can be utilized to guide the representation learning of student networks. Therefore, before conducting knowledge transfer, we need to extract the hierarchical features of teacher/student networks in advance.

As shown in Fig. 1, given an unconstrained image I, we simultaneously feed it into CSRNet and 1/n-CSRNet for feature extraction. For convenience, the j-th dilated convolutional layer in the backbone network of CSRNet is renamed as “Conv5_j” in our work. Thus the feature at layer Convi_j of CSRNet can be uniformly denoted as t_i^j. Similarly, we use s_i^j to represent the Convi_j feature of 1/n-CSRNet. Notice that t_i^j and s_i^j share the same resolution, but s_i^j has only 1/n channels of t_i^j. Since the Inter-RT in Section 3.3 computes feature relations densely, we only perform distillation on some representative features, in order to reduce the computational cost during the training phase. Specifically, the selected features are shown as follows:

$$T = \{t_1^1, t_2^1, t_3^1, t_4^1, t_5^1, t_6^1, t_7^1, t_8^1, t_9^1\}, \quad S = \{s_1^1, s_2^1, s_3^1, s_4^1, s_5^1, s_6^1, s_7^1, s_8^1, s_9^1\},$$  (1)
where $T$ and $S$ are the feature groups of CSRNet and $1/n$-CSRNet, respectively.

### 3.2 Intra-Layer Pattern Transfer

As described in the above subsection, the extracted feature $t^i_j$ implicitly contains the learned knowledge of CSRNet. To improve the performance of the lightweight $1/n$-CSRNet, we design a simple but effective Intra-Layer Pattern Transfer (Intra-PT) module, which sequentially transfers the knowledge of selected features of CSRNet to the corresponding features of $1/n$-CSRNet. Formally, we enforce $s^j_i$ to learn the visual/semantic patterns of $t^i_j$ and optimize the parameters of $1/n$-CSRNet by maximizing their distribution similarity.

Specifically, our Intra-PT is composed of two steps. The first step is channel adjustment. As feature $s^j_i$ and $t^i_j$ have different channel numbers, it is unsuitable to directly compute their similarity. To eliminate this issue, we generate a group of interim features $H = \{h^1_1, h^2_1, h^1_2, h^2_2, h^1_3, h^2_3\}$ by feeding each feature $s^j_i$ in $S$ into a $1 \times 1$ convolutional layer, which is expressed as:

$$h^j_i = s^j_i \ast w_{1 \times 1},$$

where $w_{1 \times 1}$ denotes the parameters of the convolutional layer. The output $h^j_i$ is the embedding feature of $s^j_i$ and its channel number is the same as that of $t^i_j$.

The second step is similarity computation and knowledge transfer. Since Euclidean distance is too restrictive and may cause the rote learning of student networks, we adopt a relatively liberal metric, cosine, to measure the similarity of two features. Specifically, the similarity of $t^i_j$ and $h^j_i$ at location $(x, y)$ is calculated by:

$$S^j_i(x, y) = \cos(t^i_j(x, y), h^j_i(x, y)),$$

where $C^j_i$ denotes the channel number of feature $t^i_j$ and $h^j_i(x, y, c)$ is the response value of $t^i_j$ at location $(x, y)$ of the $c$-th channel. The symbol $|\cdot|$ is the length of a vector. Thus, the loss function of our Intra-PTD is defined as follows:

$$\mathcal{L}_{\text{Intra}} = \sum_{t^i_j, h^j_i} \sum_{x=1}^{h} \sum_{y=1}^{w} 1 - S^j_i(x, y),$$

where $H^j_i$ and $W^j_i$ are the height and width of feature $t^i_j$. By minimizing this simple loss and back-propagating the gradients, our method can effectively transfer knowledge and optimize the parameters of $1/n$-CSRNet. Compared with the pair-wise distillation [39] which is heavy computed and can only conducted on high-layer features with low resolutions, our Intra-PT is more efficient and can fully transfer the knowledge embedded at various layers.

### 3.3 Inter-Layer Relation Transfer

"Teaching one to fish is better than giving him fish". Thus, a student network should also be encouraged to learn how to solve a problem. Inspired by [67], the flow of solution procedure (FSP) can be modeled with the relationship between features from two layers. Such a relationship is a kind of meaningful knowledge. In this subsection, we develop an Inter-Layer Relation Transfer (Inter-RT) module, which densely computes the pairwise feature relationships (FSP matrices) of the teacher network to regularize the long short-term feature evolution of the student network.

Let’s introduce the detail of our Inter-RT. We first present the generation of FSP matrix. For two general feature $f_1 \in \mathbb{R}^{h \times w \times m}$ and $f_2 \in \mathbb{R}^{h \times w \times n}$, we compute their FSP matrix $\mathcal{F}(f_1, f_2) \in \mathbb{R}^{m \times n}$ with channel-wise inner product. Specifically, its value at index $(c_1, c_2)$ is calculated by:

$$\mathcal{F}_{c_1, c_2}(f_1, f_2) = \sum_{x=1}^{h} \sum_{y=1}^{w} f_1(x, y, c_1) \cdot f_2(x, y, c_2) / h \cdot w.$$

Notice that FSP matrix computation is conducted on features with same resolution. However, the features in $T$ have various resolutions. To address this issue and simultaneously reduce the FSP computation cost, we consistently resize all features in group $T$ to the resolution of $t^i_j$ with max pooling. The resized feature of $t^i_j$ is denoted as $R(t^i_j)$. In the same way, all features in group $H$ are also resized to the resolution of $h^j_i$.

Rather than only compute FSP matrices for adjacent features, we design a Dense FSP strategy to better capture the long-short term evolution of features. Specifically, we generate a FSP matrix $\mathcal{F}(R(t^i_j), R(h^j_i))$ for every pair of features $(t^i_j, h^j_i)$ in $T$. Similarly, a matrix $\mathcal{F}(R(h^j_i), R(h^j_i))$ is also computed for every pair of features $(h^j_i, h^j_i)$ in $H$. Finally, the loss function of our Inter-RT is calculated as follows:

$$\mathcal{L}_{\text{Inter}} = \sum_{t^i_j, h^j_i} \sum_{h^j_i} \|\mathcal{F}(R(t^i_j), R(h^j_i)) - \mathcal{F}(R(h^j_i), R(h^j_i))\|^2.$$ (6)

By minimizing the distances of these FSP matrices, the knowledge of CSRNet can be transferred to $1/n$-CSRNet.

### 3.4 Learn from Soft Ground-Truth

In our work, a density map generated from point annotations is termed as hard ground-truth. We find that the density maps predicted by the teacher network are complementary to hard ground-truths. As shown in Fig. 2, there may exist some blemishes (e.g., the inaccurate scales and positions of human heads, the unmarked heads) in some regions of hard ground-truths. Fortunately, with powerful knowledge, a well-trained teacher network may predict some relatively reasonable maps. These predicted density maps can also be treated as knowledge and we call them soft ground-truths.

In this work, we train our student network with both the hard and soft ground-truths. As shown in Fig. 1, we use $M$ to represent the hard ground-truth of image $I$. The predicted map of CSRNet is denoted as $M_s$ and the output map of $1/n$-CSRNet is denoted as $M_t$. Since $1/n$-CSRNet is expected to simultaneously learn the knowledge of hard ground-truth and soft ground-truth, we defined the loss function on density maps as follows:

$$\mathcal{L}_m = ||M_s - M||^2 + ||M_s - M_t||^2.$$ (7)
Figure 2: Illustration the complementarity of hard and soft ground-truth (GT). (b) is a hard GT generated from point annotations with geometry-adaptive Gaussian kernels. (c) is a soft GT predicted by CSRNet, while (d) is the estimated map of 1/n-CSRNet. The hard GT may exist some blemishes (e.g., the inaccurate scales and positions of human heads, the unmarked heads). For example, the red box in (b) shows the human heads with inaccurate scales. We find that the soft GT may be relatively reasonable in some regions and it is complementary to the hard GT. Thus, they can be incorporated to train the student network.

Finally, we optimize the parameters of 1/n-CSRNet by minimizing the total losses of all knowledge transfers:

$$\mathcal{L} = \alpha_1 \cdot \mathcal{L}_{\text{Intra}} + \alpha_2 \cdot \mathcal{L}_{\text{Inter}} + \alpha_3 \cdot \mathcal{L}_m,$$

where $\alpha_1$, $\alpha_2$ and $\alpha_3$ are weights of different losses.

## 4 EXPERIMENTS

### 4.1 Experiment Settings

In this work, we conduct extensive experiments on the following three public benchmarks for crowd counting.

**Shanghaitech** [76]: This dataset contains 1,198 images with 330,165 annotated individuals. It is composed of two parts: Part-A contains 482 images of congested crowd scenes, where 300 images are used for training and 182 for testing, and Part-B contains 716 images of sparse crowd scenes, with 400 images for training and the rest for testing.

**UCF-QNRF** [22]: As one of the most challenging datasets, UCF-QNRF contains 1,535 images captured from unconstrained crowd scenes with huge variations in scale, density and viewpoint. Specifically, 1,201 images are used for training and 334 for testing. There are about 1.25 million annotated people in this dataset and the number of persons per image varies from 49 to 12,865.

**WorldExpo’10** [72]: It contains 1,132 surveillance videos captured by 108 cameras during the Shanghai WorldExpo 2010. Specifically, 3,380 images from 103 scenes are used as the training set and 600 images from other five scenes as the test set. Region-of-Interest (ROI) are provided to specify the counting regions for the test set.

### 4.2 Ablation Study

#### 4.2.1 Exploration on Channel Preservation Rate

In this work, we compress the existing crowd counting models by reducing their channel numbers. A model can run more efficiently if it has fewer parameters/channels, however, at the expense of performance. Thus, the balance between efficiency and accuracy should be investigated. In this section, we first conduct ablation study to evaluate the influence of Channel Preservation Rate (CPR) on the performance of models.

In Table 2, we summarize the performance of various CSRNet trained with different CPRs. As can be observed, the original CSRNet has 16.26M parameters and achieves an RMSE of 105.99 using 205.88G FLOPs. When preserving half the number of channels, 1/2-CSRNet has a 4x reduction in parameters and FLOPs. However, without applying knowledge transfer, its performance seriously degrades, with the RMSE increasing to 137.32. By applying our SKT, 1/2-CSRNet exhibits an obvious performance gain. When CPR decreases to 1/4, we can observe the model is further reduced in model size and FLOP consumption with only 6.25% of the original one. What’s more, the 1/4-CSRNet with knowledge transfer just has a negligible performance drop. When CPR further decreases, 1/5-CSRNet meets a relatively large performance drop with a smaller gain in parameters and FLOPs reduction. Therefore, we consider it roughly reaches a balance when CPR is 1/4 and this setting is widely adopted in the following experiments.

#### 4.2.2 Effect of Different Transfer Configurations

We further perform experiments to evaluate the effect of different transfer configurations of our framework. This ablation study is conducted...
Figure 3: Visualization of the feature maps of different models on Shanghaitech Part-A. The first and fifth columns are the features of the complete CSRNet and the naive 1/4-CSRNet. The middle three columns show the student features of 1/4-CSRNet+SKT, 1/4-CSRNet+AB [18] and 1/4-CSRNet+AT [69]. The bottom three rows are the channel-wise average features at layers Conv3_1, Conv4_1 and Conv5_1, respectively. Thanks to the tailor-designed Intra-PT and Inter-RT, our 1/4-CSRNet+SKT can fully absorb the structured knowledge of CSRNet, thus the generated features are very similar to those of the teacher network.

Table 3: Performance of 1/4-CSRNet distilled with different transfer configurations on Shanghaitech Part-A. D-FSP and S-FSP refer to Dense FSP and Sparse FSP, respectively.

Table 4: Performance of 1/4-CSRNet trained with different ground-truths on Shanghaitech Part-A.

Table 5: Performance of different compression algorithms on Shanghaitech Part-A.
### 4.3 Comparison with Model Compression Algorithms

Undoubtedly, some existing compression algorithms can also be applied to compress the crowd counting models. To verify the superiority of the proposed SKT, we compare our method with ten representative compression algorithms.

In Table 5, we summarize the performance of different compression algorithms on Shanghaitech Part-A. Specifically, quantizing the parameters of CSRNet with 8 bits, DoReFa [77] and QAT [23] obtain an MAE of 80.02/75.50 respectively. When we employ the official setting of CP [17] to prune CSRNet, the compressed model obtains an MAE of 82.05 with 6.89M parameters. To maintain the same number of parameters of 1/4-CSRNet, L1Filter [25] and AGP [78] prunes 93.75% parameters, and their MAE are above 78. Furthermore, six distillation methods including our SKT are applied to distill CSRNet to 1/4-CSRNet. As can be observed, our method achieves the best performance w.r.t both MAE and RMSE. The feature visualization in Fig. 3 also shows that our features are much better than those of other compression methods. These quantitative and qualitative superiorities are attributed to that the tailor-designed Intra-PT and Inter-RT can fully distill the knowledge of the teacher networks. What’s more, the proposed SKT is easily implemented and the distilled crowd counting models can be directly deployed on various edge devices. In summary, our SKT fits most with the crowd counting task, among the various existing compression algorithms.

### 4.4 Comparison with Crowd Counting Methods

To demonstrate the effectiveness of the proposed SKT, we also conduct comparisons with state-of-the-art methods of crowd counting from both performance and efficiency perspectives. Besides CSRNet [28], we also apply our SKT framework to distill other two representative models BL [40] and SANet [6]. Specifically, the former is based on VGG19 and we obtain a lightweight 1/4-BL with the same transfer configuration of CSRNet. Similar to GoogLeNet [60], the latter SANet adopts multi-column blocks to extract features. For SANet, we transfer knowledge on the output features of each block, yielding a lightweight 1/4-SANet.

#### 4.4.1 Performance Comparison

The performance comparison with recent state-of-the-art methods on Shanghaitech, UCF-QNRF and WorldExpo’10 datasets are reported in Tables 6, 7 and 9, respectively. As can be observed, the BL model is the existing best-performing method, achieving the lowest MAE and RMSE on almost all these datasets. The CSRNet and SANet also show a relatively good performance among the compared methods. However, when reduced in model size to gain efficiency, the 1/4-BL, 1/4-CSRNet and 1/4-SANet models without knowledge transfer have a heavy performance degradation, compared with the original models. By applying our SKT method, these lightweight models can obtain comparable results with the original models, and even achieve better performance on some datasets. For example, as shown in Table 6,
## Table 8: The inference efficiency of state-of-the-art methods. #Param denotes the number of parameters, while FLOPs is the number of floating point operations. The execution time is computed on an Nvidia GTX 1080 GPU and a 2.4 GHz Intel Xeon E5 CPU. The units are million (M) for #Param, giga (G) for FLOPs, millisecond (ms) for GPU time, and second (s) for CPU time.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Param</th>
<th>Shanghaitech A (576×864)</th>
<th>Shanghaitech B (768×1024)</th>
<th>WorldExpo 10 (576×720)</th>
<th>UCF-QNRF (2032×2912)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLOPs</td>
<td>GPU</td>
<td>CPU</td>
<td>FLOPs</td>
</tr>
<tr>
<td>DSSINet [31]</td>
<td>8.86</td>
<td>729.20</td>
<td>296.32</td>
<td>32.39</td>
<td>1152.31</td>
</tr>
<tr>
<td>CAN [36]</td>
<td>18.10</td>
<td>218.20</td>
<td>79.02</td>
<td>7.99</td>
<td>344.80</td>
</tr>
<tr>
<td>CSRNet [28]</td>
<td>16.26</td>
<td>205.88</td>
<td>66.58</td>
<td>7.85</td>
<td>325.34</td>
</tr>
<tr>
<td>BL [40]</td>
<td>21.50</td>
<td>205.32</td>
<td>47.89</td>
<td>8.84</td>
<td>324.46</td>
</tr>
<tr>
<td>SANet [6]</td>
<td>0.91</td>
<td>33.55</td>
<td>35.20</td>
<td>3.90</td>
<td>52.96</td>
</tr>
<tr>
<td>1/4-CSRNet + SKT</td>
<td>1.02</td>
<td>13.09</td>
<td>8.88</td>
<td>0.87</td>
<td>20.69</td>
</tr>
<tr>
<td>1/4-BL + SKT</td>
<td>1.35</td>
<td>13.06</td>
<td>7.40</td>
<td>0.68</td>
<td>20.64</td>
</tr>
<tr>
<td>1/4-SANet + SKT</td>
<td>0.058</td>
<td>2.52</td>
<td>11.83</td>
<td>1.10</td>
<td>3.98</td>
</tr>
</tbody>
</table>

## Table 9: MAE of different methods on the WorldExpo'10 dataset. The models with symbol * are our reimplemented teacher networks.

<table>
<thead>
<tr>
<th>Method</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang et al [72]</td>
<td>9.8</td>
<td>14.1</td>
<td>14.3</td>
<td>22.2</td>
<td>3.7</td>
<td>12.9</td>
</tr>
<tr>
<td>MCNN [76]</td>
<td>3.4</td>
<td>20.6</td>
<td>12.9</td>
<td>13.0</td>
<td>8.1</td>
<td>11.6</td>
</tr>
<tr>
<td>Shang et al [51]</td>
<td>7.8</td>
<td>15.4</td>
<td>14.9</td>
<td>11.8</td>
<td>5.8</td>
<td>11.7</td>
</tr>
<tr>
<td>IG-CNN [2]</td>
<td>2.6</td>
<td>16.1</td>
<td>10.1</td>
<td>20.2</td>
<td>7.6</td>
<td>11.3</td>
</tr>
<tr>
<td>ConvLSTM [64]</td>
<td>7.1</td>
<td>15.2</td>
<td>15.2</td>
<td>13.9</td>
<td>3.5</td>
<td>10.9</td>
</tr>
<tr>
<td>IC-CNN [45]</td>
<td>17.0</td>
<td>12.3</td>
<td>9.2</td>
<td>8.1</td>
<td>4.7</td>
<td>10.3</td>
</tr>
<tr>
<td>SwitchCNN [48]</td>
<td>4.4</td>
<td>15.7</td>
<td>10.0</td>
<td>11.0</td>
<td>5.9</td>
<td>9.4</td>
</tr>
<tr>
<td>DNCL [55]</td>
<td>1.9</td>
<td>12.1</td>
<td>20.7</td>
<td>8.3</td>
<td>2.6</td>
<td>9.1</td>
</tr>
<tr>
<td>CP-CNN [58]</td>
<td>2.9</td>
<td>14.7</td>
<td>10.5</td>
<td>10.4</td>
<td>5.8</td>
<td>8.86</td>
</tr>
<tr>
<td>PGCNet [66]</td>
<td>2.5</td>
<td>12.7</td>
<td>8.4</td>
<td>13.7</td>
<td>3.2</td>
<td>8.1</td>
</tr>
<tr>
<td>TEDnet [24]</td>
<td>2.3</td>
<td>10.1</td>
<td>11.3</td>
<td>13.8</td>
<td>2.6</td>
<td>8.0</td>
</tr>
<tr>
<td>SANet*</td>
<td>2.92</td>
<td>15.22</td>
<td>14.86</td>
<td>14.73</td>
<td>4.20</td>
<td>10.39</td>
</tr>
<tr>
<td>1/4-SANet</td>
<td>3.77</td>
<td>19.93</td>
<td>19.33</td>
<td>18.42</td>
<td>6.36</td>
<td>13.56</td>
</tr>
<tr>
<td>1/4-SANet + SKT</td>
<td>3.42</td>
<td>16.13</td>
<td>15.82</td>
<td>15.37</td>
<td>4.91</td>
<td>11.13</td>
</tr>
<tr>
<td>1/4-CSNet</td>
<td>1.96</td>
<td>15.70</td>
<td>20.59</td>
<td>8.52</td>
<td>3.70</td>
<td>10.09</td>
</tr>
<tr>
<td>1/4-CSNet + SKT</td>
<td>1.77</td>
<td>12.32</td>
<td>14.49</td>
<td>7.87</td>
<td>3.10</td>
<td>7.91</td>
</tr>
<tr>
<td>BL*</td>
<td>1.79</td>
<td>10.70</td>
<td>14.12</td>
<td>7.08</td>
<td>3.19</td>
<td>7.37</td>
</tr>
<tr>
<td>1/4-BL</td>
<td>1.97</td>
<td>18.39</td>
<td>28.95</td>
<td>8.12</td>
<td>3.94</td>
<td>12.27</td>
</tr>
<tr>
<td>1/4-BL + SKT</td>
<td>1.41</td>
<td>10.45</td>
<td>13.10</td>
<td>7.63</td>
<td>4.08</td>
<td>7.34</td>
</tr>
</tbody>
</table>

## 5 CONCLUSION

In this work, we propose a general Structured Knowledge Transfer (SKT) framework to improve the efficiencies of existing crowd counting models. Specifically, an Intra-Layer Pattern Transfer and an Inter-Layer Relation Transfer are incorporated to fully transfer the structured knowledge (e.g., layer-wise one and cross-layer one) from a heavy teacher network to a lightweight student network. Extensive evaluations on three standard benchmarks show that the proposed SKT can efficiently compress extensive models of crowd counting (e.g., CSRNet, BL and SANet). In particular, our distilled VGG-based models can achieve at least 6.5× speed-up on GPU and 9.0× speed-up on CPU, and meanwhile preserve very competitive performance.

## ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No. U1811463, 61876045 and 61976250, in part by the Natural Science Foundation of Guangdong Province under Grant No. 2017A030312006, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2020B1515020048, and in part by Zhjujiang Science and Technology New Star Project of Guangzhou under Grant No. 201906010057.
REFERENCES


Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. 2017. Spatiotemporal modeling for crowd counting in videos. In ICCV. IEEE.


