International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 3
ISSN: 2088-5334

Lightweight Fuss-Free Network-Based Crowd Counting Model Using Knowledge Distillation

Chuho Yi [a], Jungwon Cho [b,*]

[a] Department of AI Convergence, Hanyang Women's University, Seoul, Republic of Korea
[b] Department of Computer Education, Jeju National University, Jeju, Republic of Korea
Corresponding author: *jwcho@jejunu.ac.kr

Abstract— **This paper presents FFNet-S, a lightweight crowd counting model built on the simple and efficient architecture of FFNet, but enhanced via knowledge distillation (KD). The student model employs MobileNetV3 as the backbone with preservation of the multi-scale feature fusion structure of FFNet. To guide the student effectively, a composite distillation loss is introduced. This combines soft target regression, intermediate feature alignment, and attention transfer. A two-stage training strategy is adopted. Initial training on the ground truth ensures stable convergence. Next, gradual incorporation of distillation losses enhance performance. Experiments on benchmark datasets, including the ShanghaiTech Part A (SHA) and Part B (SHB), show that FFNet-S is over 90% smaller than the teacher model, but the accuracy is comparable. Moreover, FFNet-S makes inferences in real time, rendering it suitable for deployment on edge devices with limited computational resources. The proposed approach shows that a carefully designed KD framework enables compact models to exhibit the capacities of larger more complex networks without a significant loss of accuracy. Balancing of speed, accuracy, and efficiency renders FFNet-S very applicable in real-world scenarios such as surveillance systems, drones, and Internet of Things platforms. We present a practical and scalable solution for efficient crowd counting. This encourages further exploration of lightweight models for computer vision tasks when resources are constrained.**

Keywords— **Crowd counting; knowledge distillation; fuss-free network.**

Manuscript received 7 Aug. 2024; revised 29 Dec. 2024; accepted 16 Feb. 2025. Date of publication 30 Jun. 2025.

I. INTRODUCTION

Crowd counting using computer vision has become very important in terms of surveillance and public safety. This enables real-time monitoring and management of crowd densities. Traditional approaches to crowd counting can be classified into detection- and regression-based methods, including density map estimation and point regression [1]. In recent years, increasingly complex deep neural networks, such as multi-column CNNs, attention mechanisms, and transformer-based models, have been introduced, leading to significant improvements in counting accuracy. However, such models are often complex and have high computational costs.

The Fuss-Free Network (FFNet), which aims to address these issues, features simple yet effective architecture consisting of only a backbone and a multi-scale feature fusion module [2]. Despite the simple design, FFNet performs comparably to more complex models, demonstrating that simplicity and efficiency are not mutually exclusive.

Recently, as the need to deploy crowd counting models on resource-constrained devices such as CCTV cameras, drones, and Internet of Things (IoT) sensors have grown, model compression and acceleration have become increasingly important. Knowledge distillation (KD) compacts models by transferring knowledge from a large teacher to a smaller student. KD was first introduced by Hinton et al [3] and guides the student using softened teacher outputs (soft targets) [3], [4]. Since then, KD has been refined to include intermediate feature alignment, attention map imitation, and self-distillation [5], [6].

KD-based models have demonstrated promising results in crowd counting. Some student models that use only about 6% of the teacher model parameters are as accurate as, or more exact than, the whole model [7]. For example, Wang et al. [8], [9] proposed the dual knowledge distillation (DKD) framework that mitigates teacher errors by combining teacher- and self-guided learning. Chen et al. [10], [11] introduced Review KD. The "review phase" further refines the student's density map estimation. The student outperforms the teacher.

A. Crowd Counting

Early crowd counting methods used either detection or regression. Detection-based approaches, such as those employing Haar wavelets or a histogram of oriented gradients, were used to identify individual people or heads; however, they often failed when some individuals occluded others in large crowds [12], [13]. In contrast, regression-based methods used density maps to bypass explicit detection. For example, Chen et al. employed multi-scale convolutional neural networks with a multi-column CNN architecture to address scale variations [14], [15]. Subsequent studies employed multi-branch architectures, dilated convolutions, and pyramid pooling to obtain multi-scale receptive fields, effectively tackling issues such as perspective distortion and scale diversity [16] as shown in Fig. 1.

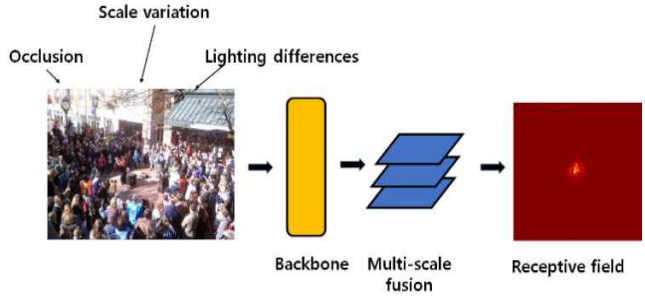

Fig. 1 Challenges associated with crowd counting and the role of FFNet.

Input crowd images often vary in terms of scale, occlusion, and lighting. FFNet effectively addresses these issues using simple architecture that features a backbone and multi-scale feature fusion. This generates robust density maps without requiring a complex model structure.

In particular, the multi-scale fusion approach combines features from different levels of a network. For example, the feature pyramid network inspired crowd counters that integrate hierarchical features [17], [18]. Transformer-based models have also emerged. For example, "Crowd Counting with Transformer" utilizes transformer blocks to model contextual information effectively [19]. The scale-adaptive selection network uses a segmentation approach to handle scale variations more effectively [20].

Recent methods have emphasized the need to strike a balance between accuracy and efficiency. For example, MobileCount uses MobileNetV2 as the backbone, and models that leverage GhostNet have also been introduced [21], [22]. Neural architecture search techniques such as the "Efficient Crowd Counting Neural Architecture Search" have been used to design efficient architectures automatically for crowd counting [23]. Additionally, point-based approaches such as P2PNet have been proposed. These directly predict individual head positions; they do not generate density maps.

Despite such advances, a significant gap remains between large, high-performing models and lightweight models for edge devices. This motivates our present study. We aimed further to compress the efficient yet strong architecture (FFNet) using knowledge distillation (KD).

B. Knowledge Distillation

Hinton et al. [3] introduced KD to compress classification models. A student network is trained using the softened outputs (soft labels) of a large teacher network. Soft labels contain richer "dark knowledge" (example: inter-class relationships) than do traditional hard labels and, thus, improve the generalizability of the student model.

Soft targets aside, Murata et al. [24] proposed FitNets. Intermediate feature hints allowed the student to learn more effectively the internal representations of the teacher. Zagoruyko and Komodakis [25] and Liu et al. [26] utilized attention transfer (AT) to assist students in mimicking the teacher's attention maps. Various loss functions have since been explored. These include the L2 distance and advanced metrics, such as the optimal transport distance (OTD) between teacher and student features, the use of cosine similarity for directional alignment, and activation boundary loss. In the fields of semantic segmentation and object detection, structured distillation methods have been applied to align pixel- or region-wise outputs [27].

In the context of crowd counting, KD poses unique challenges given the continuous nature of a density map and the significant scale variation. Some recent studies have developed distillation strategies specifically for crowd counting. For example, Wang et al. [28] introduced dual KD (DKD). In the first stage, knowledge is transferred from teacher to student by aligning the density distributions via feature alignment and use of the OTD loss. During the second stage, the student employs self-distillation to refine their predictions. This corrects any teacher-imparted bias. Remarkably, the DKD model's performance was comparable to that of a complete model, but it used only about 6% of the teacher's parameters.

Huang et al. [29] presented a KD method for IoT deployment. The approach focuses on selecting hints to robust features that withstand real-world noise. The cited studies highlight the significant potential of KD in terms of crowd counting, while also emphasizing the capacity gap—an overly small student model struggles to match the performance of a larger teacher model. In this paper, we address this issue by carefully designing a student that is moderately smaller than FFNet and leveraging a combination of distillation losses to exploit the teacher's knowledge fully.

In this paper, we propose a lightweight crowd counting model based on FFNet, combined with knowledge distillation (KD) techniques. The main contributions of this study are:

a. We designed a student model that significantly reduces the number of parameters required while preserving the simple, multi-scale feature fusion structure of FFNet.
b. We developed a composite distillation loss function that combines soft target regression, feature map alignment, and AT to transfer knowledge effectively from the teacher.
c. Experiments using two benchmark datasets, ShanghaiTech Part A and B (SHA/SHB), revealed that the distilled student model is as accurate as the full teacher FFNet. Still, both the model size and computational cost are significantly reduced.

We evaluated the proposed method in terms of the mean absolute error (MAE), mean squared error (MSE), number of model parameters, floating-point operations per image (FLOPs), and frames per second (FPS). These metrics verified the model's efficiency in making real-time inferences. The Materials and Methods describe the architecture of the FFNet student model and the distillation methodology. The Results and Discussion detail the experimental settings, the results,

and the analysis. Finally, the Conclusion summarizes the study and indicates the directions of our future work.

II. Materials and Method

A. Teacher and Student Model Architectures

Our teacher model is the original FFNet. This is structurally simple and performs well. FFNet uses the ConvNeXt-Tiny backbone to extract features that are then passed to a multi-scale feature fusion module with three parallel branches [30]. Each branch processes features from a different stage of the backbone through a focus transition module (FTM) that employs dynamic convolution and attention mechanisms to refine features and reduce dimensionality, thereby addressing any misalignment between different scales. The outputs of the three FTMs are concatenated and passed through a final convolutional layer to generate a density map that shows the crowd density at each pixel. The total crowd count is obtained by integrating the entire map.

The student model, termed Student FFNet (FFNet-S), retains the overall FFNet structure but replaces the backbone with a lighter network and reduces the number of channels within the FTM modules. Specifically, as our focus was on model efficiency in low-power devices, we used MobileNet-v3 as the model backbone. This required only 3.05 million parameters, much less than the 29.0 million of the ConvNeXt-T teacher model. However, FFNet-S retains the three-branch fusion structure, but each branch now features a simplified FTM that uses fewer filters. The output of the student model is also a density map.

B. Knowledge Distillation Framework

Our KD approach uses multiple loss components to effectively transfer knowledge. Below, T is the teacher network and S the student network. The pretrained weights of T are frozen during KD training. S is trained from scratch by T. Fig. 2 shows the overall KD process.

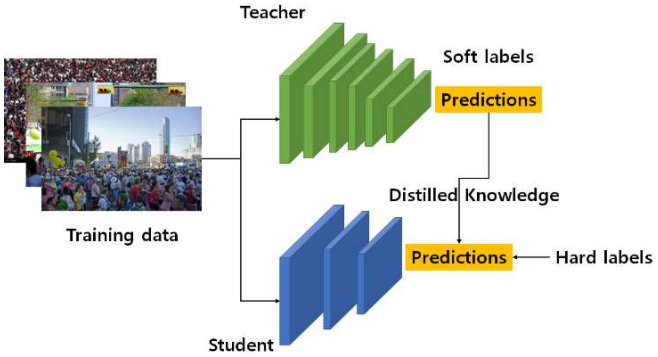

Fig. 1 The KD framework for crowd counting.

The training data, (diverse crowd scenes) are simultaneously fed into both T and S. T, typically a large and very accurate model, generates soft labels—refined predictions that capture richer distributional knowledge on the crowd density. These soft predictions supervise/guide the training of S, as do hard labels derived from the ground truth density maps.

The distilled knowledge imparted by T allows S to learn not only the final output distribution but also the implicit structural understanding captured by T during training. Such dual supervision via soft and hard labels improves the generalizability of S, even when the capacity of S is significantly lower than that of T. As Fig. 2 shows, the aim was to allow a lightweight S to mimic the predictive behavior of a high-capacity T. S produces accurate density maps using far fewer parameters and less computational time than T. Such a framework is particularly effective for crowd counting, where all scale variations, occlusions, and visual complexities pose significant challenges, and where real-time performance on edge devices is a practical necessity.

1) Soft Target Distillation Loss (L_{soft}): We obtain the output density maps from the teacher D_T and the student D_S. Instead of training the student solely with the ground truth density map (hard labels), we include a loss term that encourages D_S to mimic D_T. Following Hinton's KD formulation, we treat D_T as the "soft target" for the student S. We use a standard L2 loss on the density maps, defined as:

$$L_{soft} = |D_S - D_T|_2^2 \quad (1)$$

This loss encourages S to imitate the final predictions of T, which reflect T's refined understanding of crowd distribution. If necessary, D_T can be subjected to temperature scaling (often applied during classification tasks), but we found that direct application of the L2 loss to the density maps was sufficient.

2) Feature Distillation Loss (L_{feat}): To address the capacity gap, we aligned the intermediate representations. We extracted feature maps from key layers (backbone stages and FTM outputs) of T and the corresponding layers of S. For each teacher-student feature pair (F_T^I, F_S^I), we applied two strategies:

a. Regression loss: A 1×1 convolution (a learnable linear projection) is applied to F_S^I, and the L2 loss with F_T^I is computed. This helps the student learn feature representations close to those of the teacher.
b. Attention Transfer: We compute attention maps as $A_T = Normalize(|F_T^I|)$ and $A_S = Normalize(|F_S^I|)$, where the maps are derived by summing the absolute feature values across channels and then applying L2 normalization. The L2 loss between the two attention maps is then calculated.

The final feature loss L_{feat} is the sum of the regression and attention losses across all selected layers. This not only encourages S to match the values of T but also to focus on similar spatial regions requiring attention.

3) Hard Label Loss (L_{hard}): To ensure that S does not drift from the actual task, we retained traditional supervision using the ground truth. This is the standard regression loss between the S output D_S and the ground truth density map D_{GT}, derived using the MAE or MSE. We found that the MAE was slightly better in terms of count accuracy. This loss anchors S to the actual objective—accurate crowd counting—S does not overfit any errors made by T. As some T outputs may be inaccurate, L_{hard} is corrective.

4) Total Loss: The overall loss is defined as a weighted sum:

$$L_{total} = \alpha \cdot L_{hard} + \beta \cdot L_{soft} + \gamma \cdot L_{feat} \quad (2)$$

where α, β, and γ are hyperparameters controlling the influence of each component. Typically, α is set relatively high (e.g., 1) to prioritize ground truth supervision, and β and γ to about 0.5 each, to balance the distillation effects without overwhelming the true labels.

The experimental results showed that S is sensitive to this balance. If β or γ is too high, S may overfit errors of T; if either parameter is too low, knowledge transfer becomes ineffective. Our composite loss function is similar to the multi-term loss structure of DKD but tailored to the output characteristics, features, and attention alignment of FFNet.

C. Training

We initialize S randomly. The pretrained T remains fixed during training. The dataset images are fed into both networks, and T generates output density maps D_T with intermediate features F_T^I. However, no gradient propagates through T; only the parameters of S are updated to minimize the total loss L_{total}.

We use the AdamW optimizer with a moderate learning rate of 10^{-4}. Training continues until the validation error of S becomes stable. We adopted a two-stage training schedule inspired by DKD. In the first stage, the model is trained primarily using $L_{soft} + L_{feat}$ so that S quickly aligns with the features and outputs of T. In the second stage, the weight of L_{hard} is gradually increased to fine-tune the model toward the ground truth density maps. This helps correct any errors introduced by the pseudo-labels of T and ultimately reduces the MAE of S.

During inference, only S is used to enable fast lightweight crowd counting. As T is fixed and therefore may not adapt to shifts in the input domains, the batch normalization layers within S must be carefully tuned either by selecting an appropriate batch size or via recalibration. This ensures stable performance.

III. Results And Discussion

A. Experimental Setup

We evaluated our model using two standard benchmark datasets [31] that differ in terms of crowd density and scene diversity. SHA contains highly congested crowd scenes (482 images). SHB contains moderately dense street scenes (716 images). Following a standard protocol, both SHA and SHB were evaluated using predefined training/test splits.

We used two primary metrics commonly adopted in crowd counting: the MAE and MSE that are based on the differences between the predicted and actual head counts. MAE measures prediction accuracy (lower is better). MSE penalizes larger errors more heavily. To evaluate model efficiency, we also measured the numbers of parameters and the FLOPs.

Our teacher model, T (FFNet with the ConvNeXt-T backbone) was loaded from the website of the authors earlier cited and achieved near state-of-the-art performance, as they reported. The student model, S (FFNet-S with MobileNet-v3 as the backbone) was initialized using pretrained weights from ImageNet (excluding the final output layers). The entire implementation employed PyTorch. To augment data, we used both random horizontal flipping and cropping to obtain 256×256 image patches.

The learning rate was warmed up for 1 epoch and then decayed by a factor of 0.1 every 20 epochs. Training was terminated after 100 epochs. Ground truth density maps were generated using Gaussian kernels with fixed spread values specific to each dataset.

The loss weights were tuned and set to $\alpha = 1, \beta = 0.5, \gamma = 0.5$. KD was applied only after S had learned an initial alignment. As early application of KD sometimes triggered instability, we trained S using only L_{hard} for the first five epochs, and then added both L_{soft} and L_{feat} during later epochs.

B. Quantitative Results

Table 1 compares the performance of our T (FFNet), S (FFNet-S), and several recent crowd counting models across all datasets. Despite a near 90% reduction in the number of parameters, S exhibited only a minimal drop in performance. More importantly, S outperformed many previous models. This slight decrease in accuracy is considered acceptable, especially when weighed against the practical benefit of a 10× improvement in efficiency.

TABLE I
THE PERFORMANCES OF CROWD COUNTING MODELS USING A TEACHER MODEL (FFNET) AND A STUDENT MODEL (FFNET-S).

Method	SHA		SHB		Parameters	FLOPs
	MAE	MSE	MAE	MSE		
P2PNet	5.27	85.1	6.3	9.9	19.2M	104.7G
SASNet	53.5	88.3	6.3	9.9	38.9M	232.9G
FFNet	48.3	80.5	6.1	9.0	29M	23.7G
FFNet-S(ours)	50.4	83.7	6.5	9.3	3.1M	1.5G

In terms of efficiency, S offers significant advantages. The number of parameters is reduced from 29.0 M (T) to 3.1M (S) and the FLOPs decrease from 23.7G (T) to 1.5 G (S), substantially boosting the inference speed. S is well-suited for edge devices and can engage in real-time crowd counting (for example, surveillance applications).

C. Ablation Study

We conducted an ablation study using the SHA and SHB datasets to examine the contributions of each component in our framework. First, we removed the various distillation losses. When the soft target loss (L_{soft}) was eliminated and S was trained only on ground-truth density maps (hard labels), the MAE increased by approximately 10, indicating significant degradation in performance and highlighting the crucial role played by T-delivered soft target guidance. In contrast, when L_{soft} was retained but the feature distillation loss (L_{feat}) was excluded, the MAE rose by about 3. This suggests that feature alignment plays a vital role in stabilizing learning. Additionally, we compared attention transfer (AT) to vanilla L2 feature loss. The absence of AT increased the MAE by about 1 point, implying that AT afforded a slight but meaningful performance gain. Next, we evaluated the importance of FTM. Removal of the S FTM meant that S now engaged in only simple multi-scale fusion. The MAE

increased by approximately 5, confirming that the FTM was useful even in the lightweight *S*.

Last, we assessed the effects of different KD training schedules. Our two-stage training strategy—where L_{soft} and L_{feat} were not applied during the initial epochs, which reduced the MAE by around two compared to the figure when all losses were applied from the commencement. Delaying the distillation loss avoided instability during early training. We next tested a self-distillation strategy; S refined its predictions after initial KD training. However, this did not further improve performance, possibly because S was already strong. This result is not in line with the findings of the first DKD paper and may be attributable to differences in the datasets or model capacity.

D. Discussion

Our experimental results demonstrate that a lightweight FFNet student model was nonetheless very accurate, thanks to its feature of knowledge distillation (KD). The performance meets the practical requirements of real-world applications. For example, an MAE of approximately 50 on the SHA dataset indicates an average error of about 50 people, which is considered acceptable during trend analysis of high-density crowds.

Notably, S is capable of real-time operation, rendering S useful in embedded systems (such as drones) that monitor crowd sizes at outdoor events. Sometimes—particularly on the SHB dataset, which includes urban scenes— S performed as well as T, possibly attributable to the regularization effect of KD. As S is smaller than T, S is less prone to overfitting and may therefore generalize better during training. Similar findings have been reported in other studies on ReviewKD and DKD, where student models sometimes surpassed teacher models under specific conditions. Nevertheless, in most cases, the teacher model still serves as the upper bound in terms of performance.

IV. Conclusion

We leveraged KD when designing a lightweight crowd counting model based on the FFNet architecture. By effectively combining soft target regression with feature-level distillation, a simple S was as accurate as a complex T. Both model size and the computational cost were significantly reduced.

Extensive evaluations were conducted using several benchmark datasets. The MAE/MSE figures demonstrated that the distilled FFNet-S model performed as well as the original FFNet but with only about 10% of the parameters. Moreover, the new model yields real-time inferences using only standard hardware. This study demonstrates that, when a well-designed distillation strategy is employed, it is possible to create a simple yet high-performing network architecture without compromising accuracy.

In the future, we plan to explore an online distillation approach in which a teacher guides a student in a single session, potentially reducing the training time. Additionally, we can utilize structured pruning and quantization to further compress the already distilled model, aiming to optimize deployment on low-power devices. We plan to extend the method beyond simple counting. We can distill spatial distribution data that localizes crowds. We plan to use point-supervised losses to help the student learn not only how many people are present but also where they are located. We aim to contribute to the development of more efficient and practical crowd counting systems that can be deployed on real-world edge platforms.

Acknowledgment

Hanyang Women's University supported this research for research funding [grant number: 2024-1-018].

References

[1] B. Li, X. Zhang, X. Li, and H. Lu, "Approaches on crowd counting and density estimation: A review," *Pattern Anal. Appl.*, vol. 24, pp. 853–874, 2021, doi: 10.1007/s10044-021-00959-z.

[2] L. Chen, X. Gao, Y. Liu, and J. Wang, "The effectiveness of a simplified model structure for crowd counting," *IEEE Trans. Instrum. Meas.*, early access, 2025, doi: 10.1109/tim.2025.3554288.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv*, Mar. 2015, doi: 10.48550/arXiv.1503.02531.

[4] S. Umirzakova et al., "Simplified knowledge distillation for deep neural networks bridging the performance gap with a novel teacher-student architecture," *Electronics*, vol. 13, no. 22, Art. no. 4530, 2024, doi: 10.3390/electronics13224530.

[5] H.-B. Bak and S.-H. Bae, "Knowledge distillation based on internal/external correlation learning," *J. Korean Soc. Comput. Inf.*, vol. 28, no. 4, pp. 31–39, 2023.

[6] D. Lee, "Knowledge distillation for recommender systems," in *Proc. Korean Inst. Inf. Sci. Eng. Conf. (KIISE)*, Seoul, South Korea, 2021, pp. 48–52.

[7] R. Wang, Y. Zhao, T. Huang, and W. Zhang, "Efficient crowd counting via dual knowledge distillation," *IEEE Trans. Image Process.*, vol. 33, pp. 569–583, 2023, doi: 10.1109/tip.2023.3343609.

[8] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 8190–8199, doi: 10.1109/cvpr.2019.00839.

[9] Z. Huo et al., "Domain adaptive crowd counting via dynamic scale aggregation network," *IET Comput. Vis.*, vol. 17, no. 7, pp. 814–828, 2023, doi: 10.1049/cvi2.12198.

[10] P. Chen et al., "Distilling knowledge via knowledge review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5008–5017, doi: 10.1109/cvpr46437.2021.00497.

[11] A. N. Alhawsawi, S. D. Khan, and F. U. Rehman, "Crowd counting in diverse environments using a deep routing mechanism informed by crowd density levels," *Information*, vol. 15, no. 5, Art. no. 275, 2024, doi: 10.3390/info15050275.

[12] V. Sindagi and V. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, 2018, doi: 10.1016/j.patrec.2017.07.007.

[13] G. Gao et al., "A survey of deep learning methods for density estimation and crowd counting," *Vicinagearth*, vol. 2, no. 1, Art. no. 2, 2025, doi: 10.1007/s44336-024-00011-8.

[14] L. Chen, G. Wang, and G. Hou, "Multi-scale and multi-column convolutional neural network for crowd density estimation," *Multimedia Tools Appl.*, vol. 80, pp. 6661–6674, 2021, doi: 10.1007/s11042-020-10002-8.

[15] Y. Zhang et al., "Congested crowd counting via adaptive multi-scale context learning," *Sensors*, vol. 21, no. 11, Art. no. 3777, 2021, doi: 10.3390/s21113777.

[16] F. Zhu et al., "Real-time crowd counting via lightweight scale-aware network," *Neurocomputing*, vol. 472, pp. 54–67, 2022, doi: 10.1016/j.neucom.2021.11.099.

[17] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 1879–1888, doi: 10.1109/iccv.2017.206.

[18] C.-C. Lien and P.-C. Wu, "A crowded object counting system with self-attention mechanism," *Sensors*, vol. 24, no. 20, Art. no. 6612, 2024, doi: 10.3390/s24206612.

[19] T. Ye, X. Chu, and H. Wang, "CCTrans: Simplifying and improving crowd counting with transformer," *arXiv*, Sep. 2021. [Online]. Available: https://arxiv.org/abs/2109.14483.

[20] C. Haldız et al., "Crowd counting via joint SASNet and a guided batch normalization network," in *Proc. Signal Process. Commun. Appl. Conf. (SIU)*, Istanbul, Türkiye, 2023, pp. 1–4, doi: 10.1109/SIU59756.2023.10223901.

[21] M. Sandler et al., "MobilenetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 4510–4520, doi: 10.1109/cvpr.2018.00474.

[22] L. Zhao et al., "A lightweight deep neural network with higher accuracy," *PLoS ONE*, vol. 17, no. 8, Art. no. e0271225, 2022, doi: 10.1371/journal.pone.0271225.

[23] Y. Wang et al., "Eccnas: Efficient crowd counting neural architecture search," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 1, pp. 1–19, 2022, doi: 10.1145/3465455.

[24] R. Murata et al., "Recurrent neural network-FitNets: Improving early prediction of student performance by time-series knowledge distillation," *J. Educ. Comput. Res.*, vol. 61, no. 3, pp. 639–670, 2023, doi: 10.1177/07356331221129765.

[25] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv*, Dec. 2016. [Online]. Available: https://arxiv.org/abs/1612.03928

[26] H. Liu, Y. Zhang, and Y. Chen, "A symmetric efficient spatial and channel attention (ESCA) module based on convolutional neural networks," *Symmetry*, vol. 16, no. 8, Art. no. 952, 2024, doi: 10.3390/sym16080952.

[27] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2604–2613.

[28] Z. Li et al., "Dual teachers for self-knowledge distillation," *Pattern Recognit.*, vol. 151, Art. no. 110422, 2024, doi: 10.1016/j.patcog.2024.110422.

[29] L. Huang et al., "Context-aware multi-scale aggregation network for congested crowd counting," *Sensors*, vol. 22, no. 9, Art. no. 3233, 2022, doi: 10.3390/s22093233.

[30] S. Woo et al., "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 16133-16142, doi:10.1109/cvpr52729.2023.01548.

[31] Y. Zhang et al., "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 589–597, doi: 10.1109/cvpr.2016.70.