Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P4779-cd

P2PNeXt: Advancing Crowd Counting and Localization Using an Enhanced P2PNet Architecture

Thomas Golda

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Germany.
E-mail: thomas.golda@iosb.fraunhofer.de

Jann Sänger

University of Applied Sciences Karlsruhe (HKA), Germany.
E-mail: jann.saenger@hka.de

John Hildenbrand

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Germany.
E-mail: john.hildenbrand@iosb.fraunhofer.de

Jürgen Metzler

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Germany.
E-mail: juergen.metzler@iosb.fraunhofer.de

Accurate crowd counting and localization are essential for ensuring public safety and managing risks in densely populated areas, such as during large events or in urban environments. They enable authorities to monitor and manage large gatherings effectively, thereby preventing overcrowding and potential accidents. In emergency situations, accurate crowd data can facilitate quicker and more efficient responses by enabling the identification of high-density areas that may require immediate attention. From the computer vision perspective, these are crucial capabilities, demanding both precision in object counting and accurate spatial localization of individuals. In this study, we propose an enhancement to the P2PNet, a point-based framework for crowd counting, by integrating a modern neural network architecture, ConvNeXt, as the backbone. We explored two primary directions for the backbone integration: utilizing a feature pyramid to combine various feature maps, and employing a single feature map from ConvNeXt, bypassing the feature pyramid. Initial experiments indicated that the single-feature-map approach, particularly with the very first feature map, yielded superior results. However, through a few critical modifications to the feature pyramid module — including bilinear interpolation for upsampling, batch normalization across convolutions, and the inclusion of ReLU in the decoder — the feature pyramid approach ultimately outperformed the single feature map method. The revised feature pyramid, especially the first feature map output from the decoder module, achieved the best results across multiple datasets. This way our research contributes to the broader understanding of risk assessment and management, offering a robust solution for precise crowd density estimation and localization.

Keywords: Crowd Counting, Computer Vision, Machine Learning, ConvNeXt, P2PNet, Point-Based Framework, Public Safety.

1. Introduction

Large-scale public gatherings, such as festivals, sports events, demonstrations, and religious ceremonies, are defining aspects of human interaction and a potential issue for public safety. Although these events foster cultural and social engagement, they can also present substantial challenges in terms of safety and logistic management. Past incidents, including stampedes and overcrowded venues, have tragically underscored the need for effective crowd monitoring systems. For example, failures in crowd management have led to significant loss of life during mass gatherings, highlighting the critical need for proactive tools that can assist authorities in real-time decision-making to prevent similar occurrences. Figure 1 shows an exemplary case of a very crowded situation, in which inappropriate monitoring and crowd management might lead to crucial incidents.

Fig. 1.: Exemplary view of a festival location. A large crowd gathering in a limited space, resulting in extremely crowded conditions with high pedestrian densities. Such scenarios show an increased risk to visitor safety.

Traditional methods for crowd counting relied heavily on manual observation or simple image analysis techniques. These approaches, however, are inherently limited when dealing with high-density crowds or complex spatial arrangements. Manual methods are cost-effective but labor-intensive, while sensor-based technologies offer automation and real-time data, but come with higher costs and complexity.

In this context, automated crowd monitoring has emerged as a promising solution. By leveraging advancements in computer vision and machine learning, these systems can process visual data to estimate crowd density, track movements, and count individuals with remarkable accuracy. Among the various tasks within this domain, crowd counting and localization play a particularly significant role. They provide foundational metrics for understanding spatial distributions and planning interventions, whether for crowd dispersal, resource allocation, or emergency responses.

For this reason, we investigate the so-called P2PNet by Song et al. (2021) that is one such model that stands out for its point-based approach to crowd counting.

To address its current limitations, this paper proposes a novel extension of P2PNet by replacing its VGG16 backbone, a neural network architecture introduced by Simonyan and Zisserman (2014), with ConvNeXt introduced by Liu et al. (2022), a state-of-the-art convolutional architecture. ConvNeXt incorporates architectural advancements inspired by vision transformers while retaining the simplicity and efficiency of convolutional neural networks (CNNs). However, the transition to ConvNeXt necessitates adjustments to the feature extraction and decoding modules of P2PNet.

This paper introduces several key modifications to the P2PNet framework to enable the integration of ConvNeXt. In summary, this paper makes the following contributions:

- It integrates the ConvNeXt into the P2PNet, addressing the limitations of the VGG16 backbone and enhancing feature extraction.
- It introduces an updated feature pyramid.
- It provides a comprehensive evaluation of the modified framework, demonstrating its superiority over the reference approach on different benchmark datasets.

Our approach distinguishes itself from traditional crowd counting methods by integrating a state-of-the-art architecture, ConvNeXt, into the P2PNet framework. This integration not only addresses the limitations of the previous VGG16 backbone but also enhances feature extraction capabilities, enabling more accurate crowd counting and localization in complex scenarios.

2. Related Work

2.1. *Crowd Counting*

Crowd counting methods are generally divided into detection-based and density-based approaches. Detection-based methods aim to identify and localize individual objects, such as heads or bodies, in an image. These methods work well in sparse crowds but struggle with dense scenes due to heavy occlusions and overlapping individuals. To overcome these issues, density-based methods were introduced. Instead of detecting individuals, they estimate continuous density maps that predict the spatial distribution of the crowd. By integrating these maps, the total count can be calculated, as shown in works like Li et al. (2018); Wang et al. (2020). Density-based methods perform well in dense crowds and provide accurate counts. However, they lack the ability to precisely localize individuals, which is a key strength of

detection-based methods. This limitation has motivated research into combining both approaches to leverage their respective strengths, particularly through point-based methods.

Point-based crowd counting focuses on localizing individuals in an image by directly predicting specific points that represent their heads. The general workflow starts with generating a set of point proposals, with a popular strategy being the prediction of offsets from a grid of fixed reference points to account for the translation invariance of convolutional layers. These proposals are then matched to the ground-truth labels, often using a cost matrix that considers both classification and regression values for establishing one-to-one matching. Based on this matching, the loss is calculated, allowing for the joint optimization of regression and classification tasks. Compared to other approaches, point-based frameworks offer the advantage of precise target localization and a simpler pipeline by avoiding intermediate representations like density maps or bounding boxes.

An early implementation of point-based crowd counting is P2PNet, introduced by Song et al. (2021). It utilizes a CNN backbone as a feature extractor and outputs a feature pyramid processed in two parallel branches: one for point regression and another for classification. During training, the model learns both tasks simultaneously and produces predictions with confidence scores during inference.

Building upon this foundation, subsequent publications have introduced refinements to improve the performance of point-based frameworks. For instance, Jia et al. (2024) incorporated multi-scale feature extraction and fusion techniques to address the challenges posed by scale variations of targets, achieving significant performance gains. Ma et al. (2023) proposed an improved matching strategy that enhanced the overall accuracy of predictions.

Other contributions focus on specific aspects of the point-based workflow. The authors in Chen et al. (2024) introduced auxiliary point guidance to stabilize the proposal-target matching process and developed a feature interpolation method for adaptive feature extraction, significantly improving the robustness and accuracy of the evaluated models. Uysal and Bayazıt (2023) enhanced an existing point-based approach proposed by Zand et al. (2022) by implementing a dynamic weight assignment mechanism. This method ensures that weight parameters are updated dynamically during training and optimized jointly with the model parameters, rather than being constant, leading to improved performance.

Recently, Ryu and Song (2024) proposed PSLNet, which achieves state-of-the-art results across multiple benchmarks by introducing pseudo square labels and an anchor-free detection mechanism. This approach predicts the probability of a center point within a responsible grid and employs box regression and centerness estimation to detect individuals outside the grid. The anchor-free methodology eliminates the dependency on predefined anchor boxes, simplifying the detection process and enhancing overall performance.

These developments highlight the evolution of point-based crowd counting approaches, demonstrating their potential for achieving high accuracy and efficiency in various crowd monitoring scenarios. Our work builds on this tradition by integrating the ConvNeXt backbone into the P2PNet framework, refining its feature extraction and processing capabilities for improved performance across diverse datasets.

2.2. *Traditional Approaches to Crowd Counting*

Safety authorities and event organizers employ various traditional and technological methods to count pedestrians during large gatherings. Manual counting, using tally sheets or clickers, is a common but labor-intensive method. Counting stations are often set up at entry and exit points, staffed by counters or equipped with laser and LiDAR systems to register individuals. Mobile applications utilizing GPS technology can track participant movements and aggregate data for estimating crowd sizes, while RFID technology allows attendees to carry tags that are scanned at checkpoints for precise tracking. Additionally, mobile data from cellphone companies provides insights by analyzing aggregated and anonymized location data to estimate the number of individuals

in specific areas over time.

For larger crowds, image data from CCTV cameras or drones can assess pedestrian counts, with manual counting in predefined areas allowing for density extrapolation to estimate total counts across the event space. However, these methods often lack robustness in diverse crowd scenarios, are prone to errors, and may result in rough estimates or limited availability for arbitrary institutions. Moreover, some approaches are restricted to closed events and may not be suitable for open-area festivities.

2.3. *Semantic Feature Extraction*

Traditional convolutional neural networks like Residual Neural Networks (ResNets) introduced by He et al. (2015) and models by the Visual Geometry Group (VGG) Simonyan and Zisserman (2014) remain widely used for feature extraction in image-based tasks. These architectures have demonstrated strong performance and continue to serve as the backbone for many applications. However, recent advances in deep learning have introduced more modern architectures, such as transformer-based models Dosovitskiy et al. (2020); Liu et al. (2021) and next-generation CNNs like ConvNeXt Liu et al. (2022). ConvNeXt, in particular, offers a compelling alternative to transformer-based architectures by achieving comparable performance while maintaining the simplicity and efficiency of CNNs Jiang et al. (2024). This makes ConvNeXt a suitable choice for tasks that demand high accuracy with minimal computational overhead.

In this work, we specifically utilize the ConvNeXt-tiny variant. Its lightweight design, with fewer parameters and lower computational requirements, makes it easier to deploy in real-world applications while still delivering robust performance.

3. Methodology

3.1. *The P2PNet Model*

P2PNet is designed to directly regress the coordinates of head positions in images, eliminating the need for density maps or intermediate representations during training and inference. The model takes an image as input and produces numerical outputs, specifically pedestrian counts and image coordinates for predicted head positions, utilizing evenly distributed anchor points across the image.

For each anchor point, it computes a delta value representing the distance to the predicted head position and a confidence score assessing the detection quality. This setup can be envisioned as a grid of points, each associated with a distance vector and a confidence score, where the distance vector points towards the true head position.

The original implementation uses a VGG16 backbone pre-trained on ImageNet Deng et al. (2009) as its feature extractor. P2PNet processes four distinct feature maps from various stages of the backbone, integrating them into a feature pyramid within the decoder. Only the last three maps are utilized, with the regression branch predicting head positions and the classification branch computing confidence scores to filter out false positives.

3.2. *ConvNeXt Backbone*

To enhance P2PNet's performance, we replace the VGG16 backbone with the more contemporary ConvNeXt architecture. This transition requires adjustments due to the differing shapes of the feature maps generated by ConvNeXt. We upsample the feature pyramid outputs to match the dimensions of the corresponding VGG16 feature maps, facilitated by an upsampling module. The principal modifications as shown in Figure 2 include:

- Bilinear interpolation for upsampling instead of nearest neighbor,

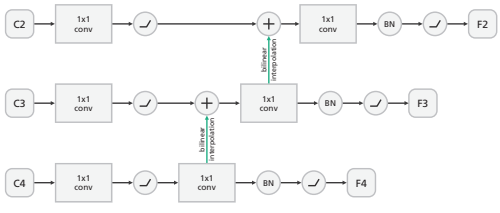


Fig. 2.: Resulting feature pyramid computation as implemented using the ConvNeXt as backbone, where C2 to C4 are the outputs of the last three ConvNeXt stages.

- Incorporation of batch normalization (BN) within the feature pyramid,
- Addition of Rectified Linear Unit (ReLU) activation in the decoder module.

These modifications aim to improve the quality of feature maps and stabilize training, enhancing the model's robustness and generalization performance across diverse datasets and real-world scenarios. In the following sections, we evaluate both single feature map and feature pyramid approaches.

4. Experiments

4.1. *Experimental Setup*

The model is implemented in PyTorch and trained for 1,000 epochs on an Nvidia L40 GPU using the Adam optimizer introduced by Kingma and Ba (2014). We adhere to the configuration used by Song et al. (2021) for the original model settings. The best-performing weights from each experiment are utilized for evaluation.

For the evaluation, we conducted $K = 5$ experiments for each configuration and report the average results, including the confidence intervals. All results reported in Tables 1-4 were obtained through our own experiments. For the comparison with the original P2PNet, we employed the available code[a] provided by Song et al. (2021).

4.2. *Evaluation Metrics*

As stated earlier, typical safety applications require two main insights: the number of individuals within a monitored area and their distribution. Hence, we evaluate the counting and localization performance of the models. To measure counting performance, we compute the mean average error (MAE) between the ground truth and the estimated pedestrian counts. Similarly, localization performance is evaluated in accordance with the methodology described in Song et al. (2021) as used in the P2PNet paper. We calculate the mean average precision (mAP), which is a standard metric for object and pedestrian detection.

[a] https://github.com/TencentYoutuResearch/CrowdCounting-P2PNet

4.3. *Datasets*

In our experiments, we utilize two widely recognized public datasets: JHU-Crowd++ by Sindagi et al. (2020) and UCF-QNRF by Idrees et al. (2018), along with an internal dataset that focuses on typical scenarios encountered in the monitoring of public events.

The JHU-Crowd++ dataset consists of approximately 4,372 images, featuring over 1.5 million annotated individuals and capturing a variety of scenes under different weather and lighting conditions. Conversely, UCF-QNRF comprises 1,535 high-resolution images with around 1.25 million annotations, distinguished by its extensive range of crowd densities and complex backgrounds.

Although our internal dataset is smaller than the aforementioned datasets (comprising 97 images with approximately 41,400 annotated individuals), it is specifically utilized for cross-domain evaluation. This evaluation is particularly relevant, as it closely reflects real-world applications, where training is often conducted on data from diverse domains prior to deployment in the final application domain.

Together, these three datasets provide critical benchmarks for assessing the performance of crowd counting models in real-world scenarios.

4.4. *ConvNeXt Backbone Comparison*

We first compare the different available versions of ConvNeXt, namely *tiny*, *small*, and *base*, excluding the *large* model, before making a final comparison with the original P2PNet.

4.4.1. *Crowd Counting*

Initially, we examine the different feature levels and their impact on counting performance. We investigate the updated feature pyramid, specifically focusing on the first feature map, the second feature map, and the complete feature pyramid. Table 1 presents an overview of our results.

When comparing the individual feature maps, we observe that utilizing the first feature map (F2) generally leads to superior performance compared to the second feature map (F3). This phenomenon can be attributed to the fact that F2 encompasses information from subsequent feature maps. We

Table 1.: Comparison of the different feature maps and the overall feature pyramid, for three ConvNeXt-architectures: *base*, *small* and *tiny*. All experiments were conducted on the JHU dataset. Best results are shown in bold.

Configuration	Feature Map 1		Feature Map 2		Feature Pyramid	
	count ↓	loc. ↑	count ↓	loc. ↑	count ↓	loc. ↑
ConvNeXt-B	70.2 ± 0.2	76.3% ± 0.2%	57.4 ± 0.1	76.3% ± 0.2%	63.7 ± 0.4	75.8% ± 0.1%
ConvNeXt-S	66.7 ± 0.4	74.6% ± 0.3%	79.0 ± 0.8	63.0% ± 0.1%	63.9 ± 0.3	74.3% ± 0.1%
ConvNeXt-T	67.4 ± 0.3	74.7% ± 0.1%	58.5 ± 0.1	76.5% ± 0.1%	**54.2 ± 0.2**	**77.6% ± 0.1%**

find that the regression and classification branches of our architecture struggle with the deeper, more complex information contained within the second feature map. This difficulty may be due to the smaller size of these feature maps and the particularly condensed nature of the information, which complicates our ability to extract relevant local information for the anchor points. In contrast, we note that the complete feature pyramid, which integrates both levels with an extended range of information, demonstrates even better results than using only the first feature map. This trend is consistently observed across all backbone variants.

Interestingly, the largest version of ConvNeXt (i.e., *base*) benefits from the second feature map, which might be due to the model size. However, this observation may indicate potential overfitting, as our cross-domain evaluation reveals poorer performance compared to the other models.

Consequently, we use the complete feature pyramid to examine the counting performance of the newly introduced feature pyramid against the original version for each of the ConvNeXt backbones. Table 2 displays the results, consistent with the experimental setup described in Section 4.1. On average, the new feature pyramid improves pedestrian counts by reducing the error by approximately 14% over the initial feature pyramid.

4.4.2. *Localization*

Analogously to the counting experiments, we evaluate the localization performance of the different backbones. The results are also contained in Table 1. As indicated in Section 4.4.1, F2 outperforms F3 not only in counting but also in localization performance. This observation is reasonable, as counting and localization are closely related tasks. The localization results obtained for the feature pyramid that incorporates both feature maps consequently reinforce our findings from Section 4.4.1. Table 3 presents a comparison between the old and new versions of the feature pyramid for the localization task.

Table 2.: Counting performance comparison between the original and proposed versions of the feature pyramid. Experiments are conducted on the JHU dataset, reporting both the mean (μ_{MAE}) and standard error. Lower values indicate better performance.

Configuration	old ↓	new ↓
ConvNeXt-B	74.9 ± 1.7	63.7 ± 0.4
ConvNeXt-S	62.0 ± 0.4	63.9 ± 0.3
ConvNeXt-T	71.0 ± 1.2	**54.2 ± 0.2**

Table 3.: Localization performance comparison between the original and proposed versions of the feature pyramid. Results are evaluated on the JHU test split, with both the mean (μ_{mAP}) and standard error reported for each experiment. Higher values indicate better performance.

Configuration	old ↑	new ↑
ConvNeXt-B	65.0% ± 0.3%	75.8% ± 0.1%
ConvNeXt-S	69.5% ± 0.3%	74.3% ± 0.1%
ConvNeXt-T	66.0% ± 0.5%	**77.6% ± 0.1%**

We observe an average improvement of 14% over the original feature pyramid in localization performance. These results underscore the advantages of employing a modern neural network architecture as a backbone.

4.4.3. *Backbone Choice*

In the preceding sections, we investigated the counting and localization performance of P2PNet using the different ConvNeXt backbones. Although the *base* version consistently demonstrates strong performance in all experiments and achieves the best results, our final choice is the *tiny* version. The improvement of the *base* over the *tiny* ConvNeXt is marginal, yet it incurs nearly four times the computational cost. This consideration is particularly important for real-world applications, where computational resources are often limited.

4.5. *P2PNeXt versus P2PNet*

The results in Table 4 compare the counting and localization performance of P2PNeXt and P2PNet across JHU and UCF-QNRF, as well as our internal dataset.

On the JHU dataset, the P2PNeXt outperforms P2PNet, indicating a relative improvement in localization of approximately 3%. In terms of MAE, P2PNeXt reduces the counting error, representing a relative improvement of about 2%. These observations are underlined by the results obtained on the UCF-QNRF dataset. P2PNeXt shows better performance, attaining an mAP improvement of about 4% over P2PNet. Furthermore, P2PNeXt has a lower MAE, reflecting a relative improvement of approximately 20%. For our internal dataset, P2PNeXt excels with an mAP of 80%, outperforming P2PNet by around 6%. The MAE for P2PNeXt is also better at 66.9, compared to 71.4 for P2PNet, translating to a relative improvement of about 9%. Since both the results on the internal dataset and the JHU dataset were obtained with the same model, these findings indicate that our proposed update to P2PNet leads to an improved generalization.

In summary, P2PNeXt demonstrates superior counting accuracy across all evaluated datasets,

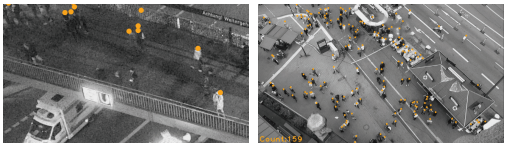
Table 4.: Comparison of P2PNeXt and P2PNet with respect to counting performance on two widely used public datasets. All results were obtained in our experiments.

Dataset	Model	mAP ↑	MAE ↓
JHU	our	**77.6% ± 0.1%**	**54.2 ± 0.2**
	P2PNet	75.5% ± 0.1%	55.0 ± 0.1
UCF-QNRF	our	**72.8% ± 0.1%**	**138.9 ± 0.9**
	P2PNet	70.1% ± 0.5%	174.4 ± 1.5
internal	our	**80.0% ± 0.2%**	**64.9 ± 1.2**
	P2PNet	75.1% ± 0.3%	71.4 ± 0.5

including JHU, UCF-QNRF, and our internal dataset. This enhanced performance is particularly notable, as it results in significant improvements in localization accuracy as well. Overall, the architectural changes in P2PNeXt yield significant advancements in typical crowd counting tasks.

4.6. *Qualitative Evaluation*

Finally, we examine examples from the internal dataset as given in Figure 3. Overall, the model



(a) Due to low contrast and strong image noise certain pedestrians are missed.

(b) Many matched people with almost no false negatives and false positives.

Fig. 3.: Exemplary results obtained using the P2PNeXt on our internal evaluation dataset. The images have been converted to grayscale to enhance the contrast.

demonstrates strong performance in cross-domain tests. However, specific instances lead to inaccuracies, particularly in missed detections of individual pedestrians, including children in strollers, or in low-light conditions where contrast is diminished. Notably, false negatives where actual pedestrians are not detected occur more frequently

than false positives, which are incorrect identifications of non-existent pedestrians. This tendency underscores the model's challenges in detecting certain groups under difficult conditions.

5. Conclusion

This paper introduces an enhanced version of the P2PNet framework by integrating the ConvNeXt architecture to improve crowd counting and localization. Our results show a 4% increase in mAP on the UCF-QNRF dataset and a 20% reduction in MAE, demonstrating significant improvements in both accuracy and reliability for crowd management applications. Additionally, evaluations on an internal dataset indicate that P2PNeXt performs well in cross-domain scenarios, achieving an mAP of 80% and an MAE of 64.9, which highlights its robustness for real-world applications. These improvements not only boost benchmark performance but also enhance the model's ability to generalize. Overall, this work marks a significant step forward in automated crowd counting for public safety applications.

Acknowledgement

This project received funding from the German Federal Ministry of Education and Research through the project ESCAPE-PRO on visitor flow simulations at concurrent large-scale events (grant no. 13N16641).

References

Chen, I.-H., W.-T. Chen, Y.-W. Liu, M.-H. Yang, and S.-Y. Kuo (2024). Improving point-based crowd counting and localization based on auxiliary point guidance.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

He, K., X. Zhang, S. Ren, and J. Sun (2015). Deep residual learning for image recognition.

Idrees, H., M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah (2018). *Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds*, pp. 544–559. Springer International Publishing.

Jia, C., Z. Cheng, Y. Leng, J. Wang, and Y. Tang (2024). *MPRNet: Multi-scale Pointwise Regression Network for Crowd Counting and Localization*, pp. 180–191. Springer Nature Singapore.

Jiang, M., S. Khorram, and L. Fuxin (2024, 6). Comparing the decision-making mechanisms by transformers and cnns via explanation methods. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9546–9555. IEEE.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization.

Li, Y., X. Zhang, and D. Chen (2018, 6). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100. IEEE.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo (2021). Swin transformer: Hierarchical vision transformer using shifted windows.

Liu, Z., H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie (2022, 6). A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ma, X., X. Wang, D. Li, and J. Piao (2023, 4). A p2pnet based on adaptively spatial feature fusion and confidence differentiation in crowd counting and profiling. In *2023 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pp. 327–331. IEEE.

Ryu, J. and K. Song (2024). Crowd counting and individual localization using pseudo square label. *IEEE Access 12*, 68160–68170.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition.

Sindagi, V. A., R. Yasarla, and V. M. Patel (2020). Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*.

Song, Q., C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu (2021). Rethinking counting and localization in crowds:a purely point-based framework.

Uysal, D. and U. Bayazıt (2023, 9). Learning weight of losses in multi-scale crowd counting. In *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, pp. 1–6. IEEE.

Wang, B., H. Liu, D. Samaras, and M. Hoai (2020). Distribution matching for crowd counting. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zand, M., H. Damirchi, A. Farley, M. Molahasani, M. Greenspan, and A. Etemad (2022). Multiscale crowd counting and localization by multitask point supervision.