IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDL8067

Publicized:2024/09/17

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

D2PT: Density to Point Transformer with Knowledge Distillation for Crowd Counting and Localization

Fan LI[†], **Enze YANG**[†a)], **Chao LI**[†], ***Nonmembers***, **Shuoyan LIU**[†], ***Member***, ***and*** **Haodong WANG**[†], ***Nonmember***

SUMMARY Crowd counting is a crucial task in computer vision, which poses a significant challenge yet holds vast potential for practical applications in public safety and transportation. Traditional crowd counting approaches typically rely on a single framework to predict density maps or head point distributions. However, the straightforward architectures often fall short in cases of over-counting or omission, particularly in diverse crowded scenes. To address these limitations, we introduce the Density to Point Transformer (D2PT), an innovative approach for effective crowd counting and localization. Specifically, D2PT employs a Transformer-based teacher-student framework that integrates the insights of density-based and head-point-based methods. Furthermore, we introduce feature-aligned knowledge distillation, formulating a collaborative training approach that enhances the performance of both density estimation and point map prediction. Optimized with multiple loss functions, D2PT achieves state-of-the-art performance across five crowd counting datasets, demonstrating its robustness and effectiveness for intricate crowd counting and localization challenges.

key words: *Crowd Counting, Head Point Localization, Vision Transformer, Knowledge Distillation*

1. Introduction

Crowd counting is a fundamental yet intricate task in computer vision, with significant applications in public safety and intelligent transportation systems. Research in this field has evolved into two primary methodologies: density-based and point-based approaches.

Density-based methods are widely adopted due to their ability to generate heatmaps that provides a quantitative assessment of crowd compositions [1]. Recent studies have focused on the visual features of highly congested scenes [2] and spatial context correlation utilizing pyramidal Convolutional Neural Networks (CNNs) [3]. However, these methods still confront challenges such as overestimation in densely populated areas and underestimation in less crowded regions.

Instead of density distributions, point-based methods directly predict head-point coordinates in crowd scenes [4]. As an intuitive approach, P2PNet [5] presents an end-to-end CNN for counting and locating individuals in crowded scenes. However, the drawback of point-based methods lies in the initialization stage, where the randomly generated proposals should encompass the ground truth targets. The optimization process is more challenging due to the classification of numerous ambiguous point proposals.

In recent years, the architecture of Vision Transformer (ViT) has demonstrated outstanding performance across various computer vision tasks. ViT-based crowd counting methods have surpassed traditional CNN architectures on multiple benchmarks due to their superior capacity to capture long-range dependencies and complex visual features [6]. Unfortunately, these methods typically replace the backbone network with ViT while employing a single loss function such as Mean Square Error (MSE), thereby overlooking the potential for feature fusion and collaborative training.

To address these limitations, we propose a Density to Point Transformer (D2PT) with Knowledge Distillation for crowd counting and localization. Our method innovatively integrates density estimation and point prediction with knowledge distillation, facilitating consistent alignment of the hidden features within the teacher-student network. Unlike the conventional paradigm where the student network sequentially learns from the output logits of the teacher network [7], this study presents a feature-aligned knowledge distillation method that leverages the advantages of both density-based and point-based approaches. This distillation paradigm has been proven to deliver exceptional performance, particularly when integrated with the ViT backbone [8]. Optimized with multiple loss functions, our D2PT achieves cutting-edge results on multiple crowd counting datasets, underscoring its potential as a formidable solution for complex crowd counting and localization tasks. The main contributions of this paper are summarized as follows:

- In this paper, we introduce a novel Density to Point Transformer (D2PT) that effectively integrates insights of density estimation and point localization. The comprehensive approach aims to address the challenges of overestimation and underestimation of crowd counting in practical scenarios.
- We adopt a feature-aligned knowledge distillation for teacher-student framework with the ViT backbone, which formulates a collaborative training approach that enhances the performance of both density estimation and point map prediction.
- We introduce multiple loss functions, including classification loss, localization loss, and disparity loss that measures the difference between the teacher and student networks. Evaluations across various public datasets reveal the state-of-the-art performance of D2PT among advanced crowd counting baselines.

[†]China Academy of Railway Sciences Corporation Limited, Daliushu Road, Beijing, China, 100081

a) E-mail: caetic@163.com

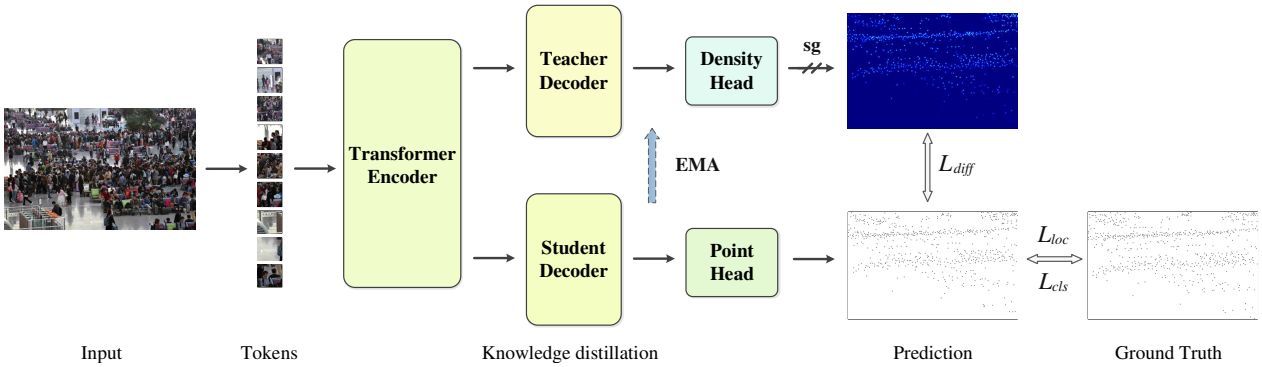

Fig. 1: The overall framework of the proposed D2PT. Note that the 'sg' behind the 'Density Head' indicates stop gradient, where the parameters of teacher decoder are only updated by EMA from student decoder.

2. Method

2.1 Preliminaries

Density-based methods: The continuous density distribution of an individual $\delta(x - x_i)$ is transformed with Gaussian kernel $G_\delta(x)$. The density-based networks are trained to predict the Gaussian distribution of the crowd, which can be formulated as: $F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\delta_i}(x), \sigma_i = \beta \bar{d}_i$ where $\overline{d_i}$ indicates the spread parameter of k nearest neighbors.

Point-based methods: For point-based approaches, let $p_i = (x_i, y_i)$ represent the center point of the i_{th} of N individuals. The collection of individuals can be denoted as $P = \{p_i | i \in 1, ..., N\}$. The point-based models predict $\hat{P}, \hat{C} = \{\hat{p}_j, \hat{c}_j | j \in 1, ..., M\}$, where $\hat{c}_j$ is the confidence score of the prediction $\hat{p}_j$. The network is trained to minimize the the Euclidean distance of $d(\hat{p}_j, p_i) = \|\hat{p}_j - p_i\|_2$.

2.2 Density to Point Transformer

The overall architecture of D2PT is shown in Fig. 1. The input image is tokenized into 16×16 patches, followed by patch embedding, the image tokens are fed into a shared-weight Transformer encoder. Subsequently, we employ two Transformer decoders with the same architecture in a teacher-student network setup. The teacher network is followed by a density head to generate Gaussian distribution, while the student network is connected to a point head that directly predicts the head-point locations.

The Transformer encoder contains several layers with a similar architecture. Each layer consists of a Self-Attention (SA) module and a Feed Forward Network (FFN). Layer normalization and residual connections are employed for each layer l. A Self-Attention module receives the input of Query (Q), Key (K), and Value (V), which can be defined as:

$$\begin{aligned} Q = Z_{l-1}W_Q, \quad K = Z_{l-1}W_K, \quad V = Z_{l-1}W_V \\ SA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{c}})V \\ MSA = [SA_1; SA_2; \cdots; SA_n]W_p \end{aligned} \tag{1}$$

where Q, K and V share the same size of input token Z, W_Q, W_K, and W_V are learnable matrices, and c is the channel dimension of Q and K. In particular, Multi-head Self-Attention (MSA) is utilized to capture long-range dependencies and spatial relationships among different regions, modeling the complex visual feature relations. According to equation (1), W_p is a re-projection matrix and n is the number of attention heads.

The structure of the teacher and student decoders is stacked with Slef-Attention (SA) layer, Cross Attention (CA) layer and FFNs. The CA layer takes two different embeddings X and input token Z, formulated as:

$$CA = SA(Q = XW_Q, K = ZW_K, V = ZW_V) \tag{2}$$

where the Q are concatenated by the trainable embeddings and content query from decoder. The teacher and student decoders output the decoded features F_{dec}.

Following the teacher decoder, we design a lightweight density projection head inspired by CrowdFormer [9]. It folds the output embedding F_{dec} into spatial feature maps, then a 1×1 convolution with the P-Sigmoid activation function is applied to fit the Gaussian density map. Note that the ViT encoder, teacher decoder and density head are previously trained for 90 epochs on the NWPU [10] dataset.

For point head projector after the student decoder, we employ two Multi-Layer Perceptrons (MLPs) as the point regression and classification layers. The first MLP revises the location of each head point, while the second MLP determines whether the point belongs to an individual.

Drawing inspiration from self-supervised knowledge distillation [8], our approach employs the teacher-student network structure to integrate robust location features from both branches. Given that the teacher and student networks

share an identical architecture, we implement a distinct Exponential Moving Average (EMA) as a momentum updater for cross-domain feature learning. The update rule can be represented as:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s \tag{3}$$

where λ is a cosine schedule from 0.996 to 1 during training. θ_t and θ_s indicate the weights of the teacher network and the student network, respectively.

As shown in Fig. 1, during the process of D2PT knowledge distillation, we stop the gradient update of the density map. The weights of the teacher decoder are updated solely through EMA of the student network. The diagram of knowledge distillation formulates the feature alignment of the underlying embeddings of intermediate layers of the teacher-student decoders. For the density projection head, we fix the parameters as it is exclusively responsible for generating Gaussian density distributions.

2.3 Loss Function

As shown in the right part of Fig. 1, the loss function for D2PT comprises three components: the classification loss L_{cls}, the localization loss L_{loc} within the student network, and the disparity loss between the teacher and student networks L_{diff}. Specifically, with the supervision of the ground truth points, we calculate the Cross Entropy loss L_{cls} for point classification. For point location loss L_{loc} in the student network, we employ the L1 loss with the KMO-based Hungarian method according to the rate of point matching. Further, we introduce the disparity loss L_{diff} to measure the consistency between the teacher and student networks. The overall loss function of the proposed D2PT is represented as:

$$\begin{aligned} L_{cls} &= -\frac{\sum_{i=1}^{N} log\hat{p}_{s(i)} + \mu\sum_{i=N+1}^{M} log(1-\hat{p}_{s(i)})}{M} \\ L_{loc} &= \left\| \hat{p}^{K}_{s(i)} - \hat{p}_{s(i)} \right\|_1 \\ L_{diff} &= \frac{|P_t - P_s|}{M} \\ L_{overall} &= L_{cls} + \nu L_{loc} + \xi L_{diff} \end{aligned} \tag{4}$$

where the M and N denote the number of predicted points and the ground truth points in L_{cls}. $\hat{p}_{s(i)}|i \in \{1,\ldots,N\}$ represents the predictions that matched with ground truth points while $\hat{p}_{s(i)}|i \in \{N+1,\ \ldots, M\}$ indicates the negative ones. μ represents the hyper-parameter of the Cross Entropy loss. In the formulation of L_{loc}, $\hat{p}^{K}_{s(i)}$ is the matched subset from predicted points of the student network $\hat{p}_{s(i)}$. As for the disparity loss L_{diff}, P_t and P_s indicate the quantitative prediction of the crowd image, which can be calculated by integrating the density map and point map of teacher and student networks. ν and ξ are hyper-parameters of the overall loss function.

3. Experimental Results

We utilize ViT-B/16 as the backbone, the number of Transformer encoder and teacher-student decoder layers are both set to 6. The Adam optimizer is implied with the batch size of 64 that trained on 4 Tesla V100 GPUs. The learning rate is set to $2e-4$ with a cosine decay scheduler.

We evaluate our D2PT and crowd counting baselines on ShanghaiTech includes Part A (SHT_A) and Part B (SHT_B) [1], UCF-QNRF (QNRF) [11], and NWPU-Crowd (NWPU) [10]. We set the crop size as 128×128 for SHT_A, 256×256 for the rest of datasets. Density-based models are validated with the metric of MAE and MSE, while the point-based methods are evaluated by average $\text{nAP}_{\{0.05:0.05:0.50\}}$.

Table 1: Experimental results of Density-basd methods evaluated on ShanghaiTech A, UCF-QNRF and NWPU datasets. The results highlighted in bold denote the best performance.

Methods	SHT_A		QNRF		NWPU	
	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [1]	110.2	173.2	277.0	426.0	232.5	764.9
CP-CNN [3]	73.6	106.4	199.4	340.4	-	-
CSRNet [2]	68.2	115.0	-	-	121.3	387.8
SFCN [12]	-	-	102.0	171.4	105.7	424.1
TransCrowd [6]	66.1	105.1	97.2	168.5	117.7	451.0
CLTR [13]	56.9	95.2	85.8	141.3	74.3	333.8
CrowdFormer [9]	56.9	97.4	78.8	136.1	67.1	301.6
D2PT (HP1)	57.5	97.8	96.0	140.0	77.9	343.8
D2PT (HP2)	56.2	95.0	74.7	127.0	72.1	301.6
D2PT (HP3)	56.0	94.5	73.9	125.4	68.5	282.1
D2PT (HP4)	**55.4**	**93.7**	**73.4**	**123.7**	**64.8**	**271.6**

Note: Hyper-Parameter (HP1-HP4) for D2PT are set as:
HP1: μ=0.1, ν=2, ξ=0.05
HP2: μ=0.5, ν=2, ξ=0.05
HP3: μ=0.5, ν=2.5, ξ=0.05
HP4: μ=0.5, ν=2.5, ξ=0.1

Results of Density based Methods Quantitative results of density-based methods are shown in Tabble 1. In ShanghaiTech A, D2PT outperforms the advanced CrowdFormer by 1.5 MAE and 3.7 MSE. For the extremely dense dataset UCF-QNRF, our method obtains 5.4 MAE and 12.4 MSE improvement compared with CrowdFormer. The NWPU dataset presents pronounced variations in inter-scene scale and density, compared with TransCrowd, the D2PT could significantly reduce the error rate by 54.9 MAE and 179.4 MSE. The ablations of hyper-parameters of our method are investigated at the bottom of Table 1, according to the results of 4 HP setting, the hyper-parameter or weight of L_{cls} (μ), L_{loc} (ν) and L_{diff} (ξ) indicate less error rate in the setting of 0.5, 2.5 and 0.1 respectively.

Results of Point based Methods The evaluation of point based methods is shown in Table 2. Compared with P2PNet,

Table 2: Experimental results of Head-Point predcition methods evaluated on ShanghaiTech A and B, UCF-QNRF, NWPU and JHU Crowd++ datasets. The results are evaluated by $nAP_{\{0.05:0.05:0.50\}}$.

Methods	SHT_A	SHT_B	QNRF	NWPU	JHU
CSRNet [2]	49.9	54.1	34.3	44.1	32.2
P2PNet [5]	64.4	76.3	53.1	65.0	58.4
TransCrowd [6]	62.2	72.4	50.7	66.1	57.6
CLTR [13]	65.8	73.9	55.6	67.6	59.2
D2PT	**68.2**	**77.9**	**58.1**	**69.8**	**60.9**

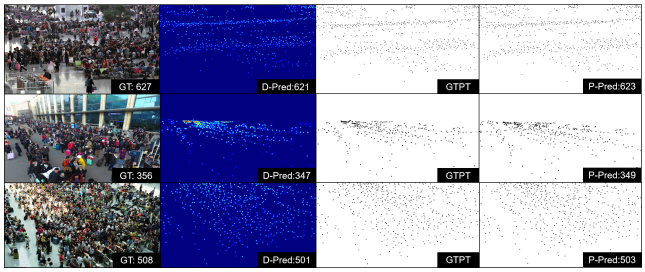

Fig. 2: Visualisation results on test set of NWPU dataset. The pictures arranged in columns are input images, density maps, ground truth points and point map predictions respectively.

D2PT introduces knowledge distillation by incorporating visual semantics of teacher and student network, which outperforms P2PNet by 1.2 to 4.0 nAP. Besides, characterized by the intricate inter-scene scale and density variations of NWPU dataset, our D2PT introduces a 2.2 nAP improvement than CLTR. These results suggest that D2PT is equipped with the capability to overcome the over-counting and omission challenges of varied scenarios.

Visualization Results To intuitively validate the performance of proposed method, we present part of the visualization results of the NWPU-Crowd dataset. As illustratesd in Fig. 2, the predicted density map (2nd column) and point map (4th column) closely resemble the ground truths in terms of density and point distributions, which demonstrates the robustness of D2PT in various crowd scenes ranging from sparse to densely populated.

4. Conclusion

This study presents the D2PT, a pioneering Transformer-based model for crowd counting and localization. By integrating the strengths and semantics of both density map-based and head-point localization techniques with feature-aligned knowledge distillation, D2PT effectively overcomes the prevalent limitations of over-counting or omission across diverse scenarios. Extensive experiments across five crowd counting datasets demonstrate that D2PT establishes a new crowd counting benchmark, indicating its robustness and reliability for critical real-world applications.

Acknowledgments

This work was supported by the fund of scientific research project of China Academy of Railway Sciences Corporation Limited (2023YJ125).

References

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp.589–597, 2016.

[2] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp.1091–1100, 2018.

[3] V.A. Sindagi and V.M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp.1879–1888, 2017.

[4] M. Rodriguez, I. Laptev, J. Sivic, and J. Audibert, "Density-aware person detection and tracking in crowds," IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, pp.2423–2430, 2011.

[5] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu, "Rethinking counting and localization in crowds: A purely point-based framework," 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp.3345–3354, 2021.

[6] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "Transcrowd: weakly-supervised crowd counting with transformers," Sci. China Inf. Sci., vol.65, no.6, pp.1–14, 2022.

[7] T. Furlanello, Z.C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born-again neural networks," Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, Proceedings of Machine Learning Research, vol.80, pp.1602–1611, 2018.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp.9630–9640, 2021.

[9] S. Yang, W. Guo, and Y. Ren, "Crowdformer: An overlap patching vision transformer for top-down crowd counting," Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, pp.1545–1551, 2022.

[10] Q. Wang, J. Gao, W. Lin, and X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," IEEE Trans. Pattern Anal. Mach. Intell., vol.43, no.6, pp.2141–2149, 2021.

[11] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Máadeed, N.M. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II, pp.544–559, 2018.

[12] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp.8198–8207, Computer Vision Foundation / IEEE, 2019.

[13] D. Liang, W. Xu, and X. Bai, "An end-to-end transformer model for crowd localization," Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I, pp.38–54, 2022.