

Article

Hierarchical Inverse Distance Transformer for Enhanced Localization in Dense Crowds

Xiangfeng Qiu ^{1,2}, Jin Ye ¹, Siyu Chen ¹ and Jinhe Su ^{1,*} 

¹ School of Computer Engineering, Jimei University, Xiamen 361021, China; qxf@kingtopinfo.com (X.Q.); yj@jmu.edu.cn (J.Y.); chensy@jmu.edu.cn (S.C.)

² Xiamen Kingtop Information Technology Co., Ltd., Xiamen 361008, China

* Correspondence: sujh@jmu.edu.cn

Abstract: Achieving precise individual localization within densely crowded scenes poses a significant challenge due to the intricate interplay of occlusions and varying density patterns. Traditional methods for crowd localization often rely on convolutional neural networks (CNNs) to generate density maps. However, these approaches are prone to inaccuracies stemming from the extensive overlaps inherent in dense populations. To overcome this challenge, our study introduces the Hierarchical Inverse Distance Transformer (HIDT), a novel framework that harnesses the multi-scale global receptive fields of Pyramid Vision Transformers. By adapting to the multi-scale characteristics of crowds, HIDT significantly enhances the accuracy of individual localization. Incorporating Focal Inverse Distance techniques, HIDT adeptly addresses issues related to scale variation and dense overlaps, prioritizing local small-scale features within the broader contextual understanding of the scene. Rigorous evaluation on standardized benchmarks has unequivocally validated the superiority of our approach. HIDT exhibits outstanding performance across various datasets. Notably, on the JHU-Crowd++ dataset, our method demonstrates significant improvements over the baseline, with MAE and MSE metrics decreasing from 66.6 and 253.6 to 59.1 and 243.5, respectively. Similarly, on the UCF-QNRF dataset, performance metrics increase from 89.0 and 153.5 to 83.6 and 138.7, highlighting the effectiveness and versatility of our approach.



Citation: Qiu, X.; Ye, J.; Chen, S.; Su, J.

Hierarchical Inverse Distance

Transformer for Enhanced

Localization in Dense Crowds.

Electronics **2024**, *13*, 2289. <https://doi.org/10.3390/electronics13122289>

Academic Editors: Chiman Kwan and Silvia Liberata Ullo

Received: 14 May 2024

Revised: 4 June 2024

Accepted: 9 June 2024

Published: 11 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: crowd localization; Pyramid Transformer; FIDT

1. Introduction

Crowd localization [1] is a crucial component in crowd analysis, designed to accurately identify and pinpoint individual positions in densely populated environments. This capability is essential for analyzing crowd behavior, compiling pedestrian statistics, and enhancing safety monitoring. Traditional methods often employ bounding boxes to detect individuals, but these can falter in congested areas where deep learning-based detectors struggle to deliver accurate results. Additionally, the cost of creating bounding box annotations for dense crowds is often prohibitively high, leading many crowd datasets to include only point-level annotations.

To overcome these challenges, models like PSDDN [2] and LSC-CNN [3] have adopted strategies using the nearest-neighbor head distance to establish pseudo-ground-truth bounding boxes, akin to those used in detection models. However, these models still depend on bounding boxes during training and utilize complex detection frameworks like Faster R-CNN [4], which can result in pseudo-ground-truth boxes that inaccurately represent actual head sizes, leading to suboptimal performance.

Building on this, Dinggang Liang and his team have introduced an innovative approach that utilizes Focal Inverse Distance Transform Maps (FIDT) [5] based on the HRNet backbone. This method, complemented by an Independent Structural Similarity Loss function (I-SSIM loss) and implemented through a Local Maximum Detection Strategy

(LMDS), effectively addresses the typical limitations found in dense crowd analysis. FIDT Maps, created by applying an inverse distance transform to crowd density maps, effectively extract and highlight the distance between each pixel and the nearest human head, providing precise location coordinates for crowd localization. The use of HRNet [6], known for maintaining high-resolution representations throughout the network, is crucial for preserving the spatial details necessary for accurate individual detection in densely populated areas. The integration of FIDT Maps with the HRNet backbone provides a robust solution for the traditional challenges in crowd localization, marking a significant advancement in the field.

Due to its reliance solely on the last layer of HRNet features, the Focal Inverse Distance Transform Maps method faces challenges in accurately calculating focal neighborhoods, particularly in dense crowds or under complex weather conditions such as heavy fog or rain. To overcome these limitations, this paper introduces the Hierarchical Inverse Distance Transformer, utilizing a Transformer-based backbone to enhance global modeling capabilities and substantially improve global perception in densely populated areas. Additionally, a Pyramid Vision Transformer [7] generates multi-scale feature maps, which are incrementally fused with deep residuals recovered from various levels of a Laplacian pyramid. Our proposed approach significantly mitigates localization issues in overlapping crowd areas, as evidenced by its performance on the ShanghaiTech dataset, as shown Figure 1, with mean absolute error (MAE) and mean squared error (MSE) of 53.4 and 6.5, and 93.5 and 10.7 for parts A and B, respectively. Moreover, on the JHU-Crowd++ dataset [8], our method shows substantial improvements over the baseline, with MAE and MSE reducing to 62.1 and 243.5 from 66.6 and 253.6. Similarly, on the UCF-QNRF dataset [9], our metrics improved from an MAE of 89.0 and an MSE of 153.5 to 83.6 and 153.7, underscoring the effectiveness and adaptability of our model.

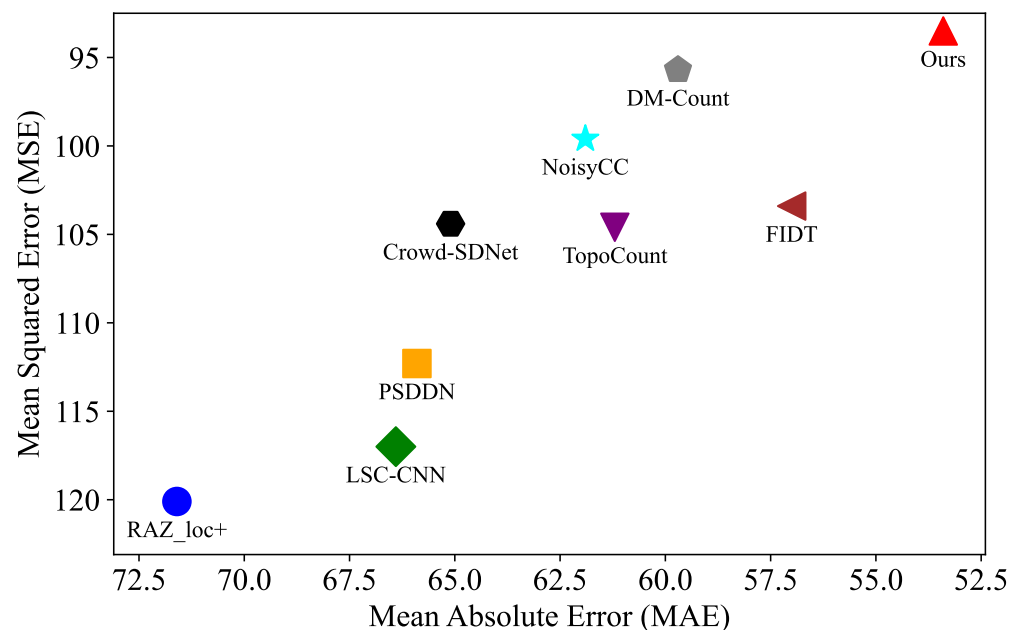


Figure 1. Comparison of MAE and MSE performance in ShanghaiTech A Part.

The key contributions of this paper are summarized as follows:

- (1) Introduction of the Hierarchical Inverse Distance Transformer (HIDT): A novel Transformer-based backbone that enhances global modeling capabilities, significantly improving global perception in densely populated areas;
- (2) Implementation of the Pyramid Vision Transformer: This approach generates multi-scale feature maps that are incrementally fused with deep residuals from various levels of a Laplacian pyramid, effectively mitigating localization issues in overlapping crowd regions;

(3) Significant Performance Improvements: The proposed method demonstrates superior performance across multiple datasets, including ShanghaiTech, JHU-Crowd++, and UCF-QNRF. It substantially reduces mean absolute error (MAE) and mean squared error (MSE) metrics, underscoring the model's effectiveness and adaptability.

2. Related Work

Contemporary crowd analysis methodologies primarily focus on counting, employing Convolutional Neural Networks (CNNs) [10,11] to create density maps that estimate crowd sizes through data synthesis. To increase the precision of these maps, various strategies merge features from disparate layers or scales. These frameworks also incorporate attention mechanisms to significantly enhance focus on critical areas, thereby boosting accuracy. The CLIP-EBC [12] framework has developed the first entirely CLIP-based model, addressing issues in existing classification methods such as improper discretization strategies and label noise. By enhancing the discretization strategy, implementing label correction methods, and introducing a distance-aware cross-entropy loss function, the framework significantly improves the accuracy of crowd counting. Additionally, the use of multi-head layers has been shown to aggregate features from CNN architectures more effectively. Diversifying the representation of density maps during training phases is critical for improving model performance. With the aim of reducing the extensive demand for manual labeling, semi-supervised [13] and weakly supervised [14] learning approaches have been implemented. Despite these advances, most counting methods still only provide aggregate counts or basic density maps without accurately identifying individual locations, greatly limiting their real-world applicability. The recently proposed Focal Inverse Distance Transform Maps primarily utilize CNNs to generate basic density maps for crowd size estimation but typically fail to precisely locate individuals, especially in dense or complex settings. Addressing these deficiencies, this paper introduces the Hierarchical Inverse Distance Transformer. This method leverages a Transformer-based backbone to significantly enhance global modeling capabilities and markedly improve perceptual accuracy in densely populated areas. This method contrasts sharply with the earlier Focal Inverse Distance Transform Maps approach, which relies solely on the final layer features of HRNet and struggles with accurate focal neighborhood calculations in extremely dense crowds and adverse conditions such as heavy fog. On the JHU-Crowd++ dataset, our method has shown substantial enhancements over the baseline, with the mean absolute error (MAE) and mean squared error (MSE) decreasing to 62.1 and 243.5 from initial values of 66.6 and 253.6, respectively. These results clearly demonstrate the enhanced precision and flexibility of our model in diverse and challenging conditions, reflecting significant enhancements in the field of crowd analysis.

3. Methodology

3.1. Hierarchical Inverse Distance Transformer

The structure of the Hierarchical Inverted Distance Transformer (HIDT) is illustrated in Figure 2, featuring a distinctive encoding strategy. The encoder begins by using Atrous Spatial Pyramid Pooling (ASPP) to perform Patch Embedding calculations. The purpose of Patch Embedding is to transform the input image into a series of fixed-size patches and subsequently compute global attention. ASPP, or Atrous Spatial Pyramid Pooling, involves calculating dilated convolutions with multiple dilation factors, concatenating the results into a single tensor, and then reducing its dimensionality via a 1×1 convolution. The aim of this module is to capture receptive fields of different scales and extract multi-scale information. Following the Patch Embedding, the process enters the Transformer block, which is divided into four stages. These stages are designed to further process and refine feature information, ensuring comprehensive handling of the data collected by the encoder.

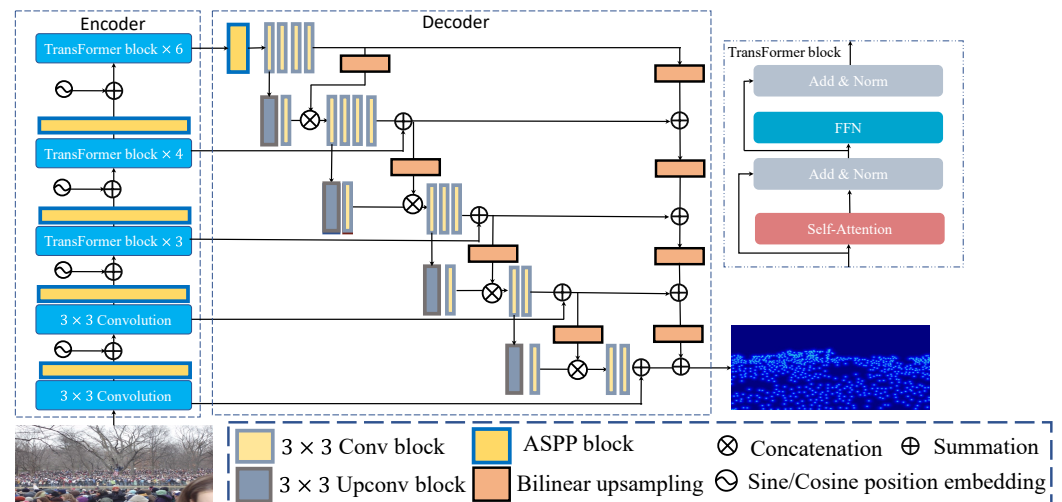


Figure 2. Model structure of Hierarchical Inverse Distance Transformer. The encoder of the Hierarchical Inverse Distance Transformer (HIDT) employs global attention computation based on Transformers, while the decoder combines dense skip connections with a Laplacian pyramid structure. The encoder utilizes an Atrous Spatial Pyramid Pooling (ASPP) module to compute Patch Embedding, which integrates local image information. This local–global computation enhances the model’s robustness.

In the decoding phase of the HIDT, the architecture is thoughtfully partitioned into several branches, each tailored to correspond with a distinct level of a Laplacian pyramid. This segmentation is designed to optimize feature reconstruction across various scales. The uppermost branch of the decoder is specifically tasked with reconstructing the global layout of features, essentially providing a broad overview of the scene’s spatial structure. Below this top layer, additional branches focus on extracting depth residuals from the latent features uncovered by the encoder. These branches are critical for refining the finer details within the image, allowing the model to discern subtle variations in crowd density and individual groupings. By employing this hierarchical approach, the decoder adeptly handles the complexity of densely populated scenes, ensuring that both the overarching patterns and the intricate details are accurately represented. This structured decoding strategy enhances the model’s ability to perform precise crowd localization by capturing essential spatial dynamics and textural nuances.

In addition to the aforementioned model architecture, the application of the FIDT technique to generate a Focused Inverse Distance Transform Map is of paramount importance. To ensure the fidelity of local features and the integrity of structures, the model employs a dual-loss system that combines mean squared error (MSE) and Independent Structural Similarity (I-SSIM) loss. This approach not only aims to minimize pixel error, but also emphasizes maintaining the structural accuracy of crowd features, thereby significantly enhancing the model’s performance in accurately locating individuals in densely populated environments. This dual focus on detail and structure substantially improves the reliability and precision of the crowd localization function within the HIDT model.

3.2. Focal Inverse Distance

The integration of the Focal Inverse Distance Transform (FIDT) map within the Hierarchical Inverse Distance Transformer (HIDT) framework significantly enhances the precision of crowd distribution modeling. This advancement is vital for applications such as crowd counting, density estimation, and individual localization in complex environments. The FIDT Map ensures that the model’s focus remains on crucial regions, thereby reducing the impact of irrelevant background information and improving overall performance.

The distance transform map of implicit information is defined as follows:

$$P(x, y) = \min_{(x', y') \in B} \sqrt{(x - x')^2 + (y - y')^2} \quad (1)$$

In Equation (1), B represents the comprehensive set of head annotations. For any pixel located at (x, y) , $P(x, y)$ measures the shortest distance from that pixel to the nearest annotated head position. Directly regressing the distance transform map poses challenges due to the wide range of distance values, which can extend from zero to the maximum dimension of the image.

To address these challenges and manage the variation in distances, we employ an inverse function, formalized in Equation (2):

$$I = \frac{1}{(P(x, y))^{\alpha \times P(x, y) + \beta} + C'} \quad (2)$$

This formula modulates the attenuation rate within the transform, prioritizing head regions and swiftly diminishing background noise. The parameters α and β control the attenuation rate, while C' is a constant that prevents division by zero, ensuring stability in the transformation. By fine-tuning these parameters, the FIDT Map can effectively focus on areas of high importance, thereby enhancing the model's ability to distinguish between densely populated regions and sparse background areas.

The strategic application of the FIDT Map within the HIDE framework allows for more accurate modeling of crowd distributions. This methodology is particularly beneficial for tasks such as crowd counting, density estimation, and individual localization in complex scenarios. By concentrating the model's attention on critical regions and minimizing the influence of irrelevant background information, the FIDT Map significantly improves the reliability and precision of the crowd localization function in the HIDE model.

3.3. Loss Function Calculation

Using only MSE loss for training can lead to negative impacts, such as blurring effects and loss of local structure information. To address these issues, SSIM loss has been proven to enhance the quality of the predicted map. The SSIM is defined as follows:

$$SSIM(E, G) = \frac{(2\mu_E\mu_G + \lambda_1)(2\sigma_{EG} + \lambda_2)}{(\mu_E^2 + \mu_G^2 + \lambda_1)(\sigma_E^2 + \sigma_G^2 + \lambda_2)} \quad (3)$$

Here, E and G denote the estimated map and ground-truth map, respectively. The variables μ and σ represent the mean and variance, while λ_1 and λ_2 are set to 0.0001 and 0.0009 to prevent division by zero. The SSIM ranges from -1 to 1 , with $SSIM = 1$ indicating an exact match between the estimated and ground-truth maps. The SSIM loss is thus expressed as

$$L_S(E, G) = 1 - SSIM(E, G). \quad (4)$$

Typically, SSIM loss employs a sliding window to scan the predicted map without differentiating between the foreground (head region) and the background. However, for localization tasks focusing on local maxima, this approach may highlight false local maxima in the background. To address this, we propose the Independent SSIM (I-SSIM) loss, defined as

$$L_{I-SSIM} = \frac{1}{N} \sum_{n=1}^N L_S(E_n, G_n), \quad (5)$$

where N is the total number of persons, and E_n and G_n are the estimated and ground-truth maps for the n -th instance, respectively. By using a window size of 30×30 , we ensure

that it captures the entire head region without including unnecessary background for most instances. The final training objective is thus given by

$$L = L_{MSE} + L_{I-SSIM}, \quad (6)$$

where L_{MSE} represents the MSE loss and L_{I-SSIM} denotes the proposed I-SSIM loss.

4. Experimental Results

4.1. Datasets and Metric

JHU-Crowd++, published by Johns Hopkins University, provides a diverse array of scenes set in complex environmental conditions, featuring different weather, lighting, and levels of crowd density. It includes a total of 4372 images, split into 1515 for training, 635 for validation, and 2227 for testing. The dataset offers detailed annotations, including the marking of individual head positions and annotations related to crowd behavior and directional movements, making it exceptionally suited for advanced crowd behavior analysis and counting research.

UCF-QNRF, released by the University of Central Florida, features extremely high-density crowd scenes with over 1.25 million fine-grained head annotations. This collection includes 1535 images, with 1201 allocated for training and 334 for testing, lacking a separate validation subset. The high resolution and extensive annotations make this dataset ideal for evaluating and improving the performance of crowd counting models in challenging conditions.

ShanghaiTech, issued by the Shanghai University of Science and Technology, is categorized into Part A and Part B. Part A contains high-density crowd scenes on urban streets, with a total of 482 images: 300 for training and 182 for testing. Part B includes lower-density settings, such as parks and small streets, with 716 images in total: 400 for training and 316 for testing. This structure supports varied research needs in crowd density analysis and model performance testing.

4.2. Implementation Details

Our model was trained on hardware comprised of four NVIDIA 4090 GPUs, each assigned a batch size of 32. Initially, the network was pre-trained on the ImageNet-1K dataset. The training process of the model begins with the preprocessing of images. After preprocessing, the images are fed into HIFT, which outputs a feature map. This feature map is then compared with the ground-truth labels to compute the loss. Specifically, the feature map produced by the HIFT is used to generate a Focused Inverse Distance Transform map. Subsequently, we compute the mean squared error (MSE) and Independent Structural Similarity (I-SSIM) losses. This approach ensures that the model maintains accuracy, not only at the pixel level, but also in preserving structural consistency.

4.3. Training Strategy and Hyperparameters

Input Data: The input to the system consists of image data, which are preprocessed before being used for training.

Feature Details: The model extracts features that include multi-scale information. These features are embedded using Atrous Spatial Pyramid Pooling (ASPP) and are processed and refined across multiple stages of the Transformer.

Data Augmentation: Data augmentation techniques such as random flipping, color jittering, and Gaussian blurring are employed to improve the generalization capability of the model.

Optimizer: The AdamW optimizer is used, with weight decay set to 0.05.

Learning Rate: The initial learning rate is set to 1×10^{-4} , utilizing a cosine annealing schedule, and includes a warm-up phase.

Loss Function: We adopt a dual-loss system that combines mean squared error (MSE) and Independent Structural Similarity (I-SSIM) loss to simultaneously minimize pixel error and preserve structural accuracy.

4.4. Ablation Experiment

Table 1 shows the ablation study results on the JHU-Crowd++ dataset. The analysis emphasizes the impact of the different components on the model performance, which is evaluated using the mean absolute error (MAE) and mean squared error (MSE). Secondly, since the Decoder requires multi-scale features for computation, the ablation experiments do not evaluate the Decoder in isolation but rather in conjunction with FIDT as Decoder and FIDT.

Table 1. Ablation experiments of model components on the JHU-Crowd++ dataset. The best results are in bold.

Hierarchical Transformer Encoder	Laplacian Pyramid Decoder	Focal Inverse Distance Transform	MAE	MSE
✓			72.1	286.2
✓		✓	69.3	272.5
✓	✓		63.1	257.5
✓	✓	✓	59.1	243.5

The baseline model with only the Hierarchical Transformer Encoder achieves an MAE of 72.1 and an MSE of 286.2. Adding the Focal Inverse Distance Transform improves the results to an MAE of 69.3 and an MSE of 272.5. Incorporating the Laplacian Pyramid Decoder further reduces the MAE to 63.1 and the MSE to 257.5. The best performance is obtained by integrating all components, achieving an MAE of 59.1 and an MSE of 243.5. This demonstrates that each component contributes to enhancing the model's accuracy, with the full configuration providing the most significant improvement.

4.5. Comprehensive Evaluation

In this section, we undertake an exhaustive evaluation of our model's performance on three benchmark datasets: JHU-Crowd++, UCF-QNRF, and ShanghaiTech. We also provide visual representations from these datasets to enable a comprehensive comparison with current methods.

Our analysis aims to demonstrate the robustness and adaptability of our model under the diverse conditions encapsulated by each dataset. The JHU-Crowd++ dataset, with its varied scenes and complex environmental conditions, serves as a testbed for our model's ability to manage variations in crowd density and environmental dynamics such as weather and lighting conditions. The UCF-QNRF dataset, which is characterized by extremely high-density crowds, poses a challenge to our model's accuracy in crowd counting under congested scenarios. Meanwhile, the ShanghaiTech dataset offers a diverse array of environments from high-density urban streets to less crowded park areas, making it ideal for assessing our model's versatility.

The visual comparisons provided showcase not only the quantitative improvements that our model achieves but also the qualitative enhancements in crowd detection and analysis. These visualizations specifically highlight the accuracy and detail improvements that our model introduces compared to existing approaches, emphasizing our contributions to the field of crowd analysis.

4.5.1. JHU-Crowd++ Results Analysis

As indicated in Table 2, our approach outperforms current mainstream models on similarity metrics, achieving reductions of 7.5 in mean absolute error (MAE) and 9.9 in mean squared error (MSE) relative to the formerly best model. Figure 3 (panels 2, 3, and 5) shows that the FIDT model can sometimes predict considerable irrelevant distances. Moreover, Figure 3 (panels 1, 3, and 6) demonstrates a tendency to produce numerous erroneous points, resulting in redundant data points.

Table 2. Main results on the JHU-Crowd++ and UCF-QNRF datasets. The best results are in bold. The data in the table are derived from FIDT [5].

Method	JHU		QNRF	
	MAE	MSE	MAE	MSE
CSRNet [15]	85.9	309.2	112.7	189.7
SFCN [16]	77.5	297.6	102.0	171.4
L2SM [17]	79.3	316.4	104.7	173.6
CG-DRCN [18]	82.3	328.0	112.2	176.3
MUD-iKNN [19]	97.6	394.4	104.0	172.0
DSSI-Net [20]	133.5	416.5	99.1	159.2
MBTTBF [21]	81.8	299.1	97.5	165.2
BL [22]	75.0	299.9	88.7	154.8
KDMG [23]	69.7	268.3	99.5	173.0
NoisyCC [24]	67.7	258.5	85.8	150.6
DM-Count [25]	75.3	279.4	85.6	148.3
RAZ_loc+ [25]	89.7	320.6	118.0	198.0
PSDDN [2]	91.3	292.4	97.3	162.0
LSC-CNN [3]	112.7	454.4	120.5	218.2
Crowd-SDNet [26]	72.1	263.1	86.7	187.0
AutoScale [27]	85.6	356.1	104.4	174.2
TopoCount [1]	60.9	267.4	89.0	159.0
FIDT [5]	66.6	253.6	89.0	153.5
Ours	59.1	243.5	83.6	138.7

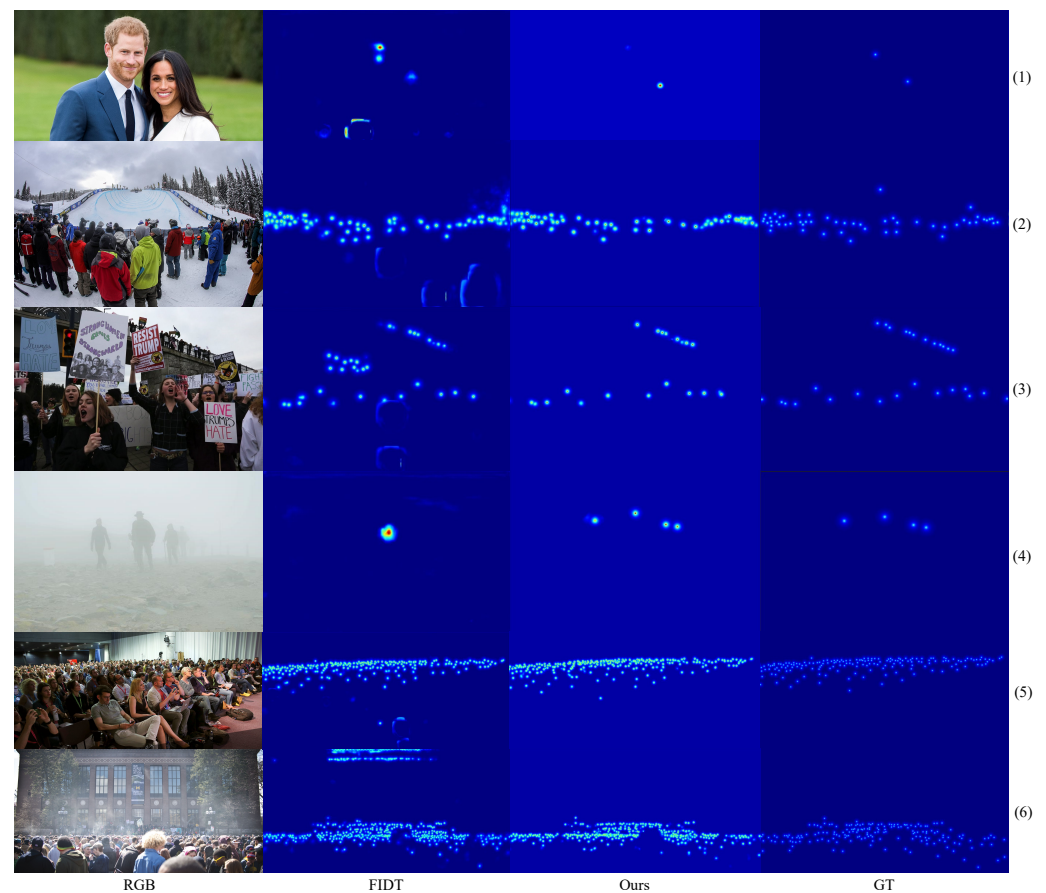


Figure 3. Qualitative results on JHU-Crowd++. In comparison to the FIDT, our method achieves more accurate and refined localization.

Particularly, in Figure 3 (panel 1), our model demonstrates weak responses to the points on the left, indicating a potential shortfall in capturing comprehensive spatial features in that segment of the image. In panel 3, there is evident dispersion in the model's predictions for smaller targets, leading to a failure in accurately locating more distant individuals within the crowd. This issue may point to limitations in the model's sensitivity or resolution, especially when dealing with smaller or less distinct features in complex scenes.

4.5.2. UCF-QNRF Results Analysis

As shown in Table 2, our method outperforms existing mainstream models in terms of similarity metrics, with reductions in mean absolute error (MAE) by 5.4 and mean squared error (MSE) by 14.8 compared to the previous best model. Nonetheless, Figure 4 reveals significant misdetections in the FIDT model's performance. Notably, panel 1 displays a proliferation of redundant localizations, while the subsequent panels also exhibit numerous computation errors that lead to incorrect mapping distances.

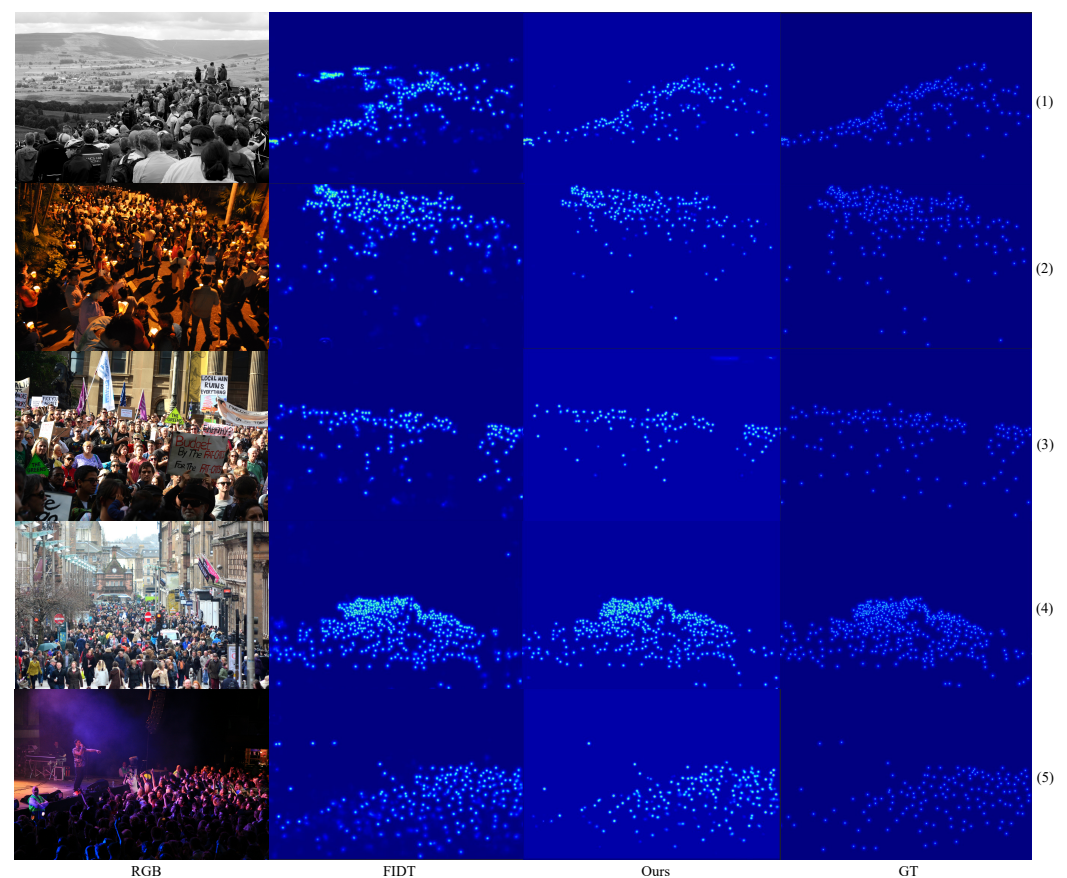


Figure 4. Qualitative results on UCF-QNRF. In comparison to the FIDT, our method achieves more accurate and refined localization.

This comprehensive evaluation emphasizes the dual nature of our model's performance: its success in reducing overall error metrics and its challenges in specific operational contexts. In particular, Figure 4 (panel 1) illustrates an issue with the model's propensity to over-detect in densely populated areas, potentially due to an oversensitivity to particular features or inadequate filtering processes. The computation errors in other panels may reflect difficulties in managing spatial relationships and accurately calculating distances in environments with high feature density or variable dynamics.

4.5.3. ShanghaiTech Results Analysis

As indicated in Table 3, our approach has shown significant achievements for both Part A and Part B of the ShanghaiTech dataset. Relative to the FIDT model, our method decreased the mean absolute error (MAE) in Part A by 3.6 and the mean squared error (MSE) by 9.9. In Part B, we reduced the MAE from 6.9 to 6.4 and the MSE from 11.8 to 10.7. Despite the MSE in Part B not achieving the lowest possible value, it was close to the optimal result.

Table 3. Main results on the ShanghaiTech Part A and Part B datasets. The best results are in bold. The data in the table are derived from FIDT [5].

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
CSRNet [15]	68.2	115.0	10.6	16.0
SFCN [16]	64.8	107.5	7.6	13.0
L2SM [17]	64.2	98.4	7.2	11.1
CG-DRCN [18]	64.0	98.4	8.5	14.4
MUD-iKNN [19]	68.0	117.7	13.4	21.4
DSSI-Net [20]	60.6	96.0	6.9	10.3
MBTTBF [21]	60.2	94.1	8.0	15.5
BL [22]	62.8	101.8	7.7	12.7
KDMG [23]	63.8	99.2	7.8	12.7
NoisyCC [24]	61.9	99.6	7.4	11.3
DM-Count [25]	59.7	95.7	7.4	11.8
RAZ_loc+ [25]	71.6	120.1	9.9	15.6
PSDDN [2]	65.9	112.3	9.1	15.6
LSC-CNN [3]	66.4	117.0	8.1	12.7
Crowd-SDNet [26]	65.1	104.4	7.8	12.6
AutoScale [27]	65.8	112.1	8.6	13.9
TopoCount [1]	61.2	104.6	7.8	13.7
FIDT [5]	57.0	103.4	6.9	11.8
Ours	53.4	93.5	6.4	10.7

Figure 5 (panels 1 and 2) demonstrates our model's superior performance in densely crowded scenes, offering finer granularity compared to the FIDT model, which tends to produce results with significant clumping. In scenarios with even higher density, as shown in panel 3, the FIDT model sometimes resulted in prediction voids. In contrast, our model effectively prevented such issues through the implementation of multi-scale connections. Furthermore, Figure 6 (panels 1 and 2) identifies minor missed detections in the FIDT model, and panel 3 shows that the FIDT had numerous erroneous localizations in its lower right segment due to errors in mapping from single-layer feature maps.

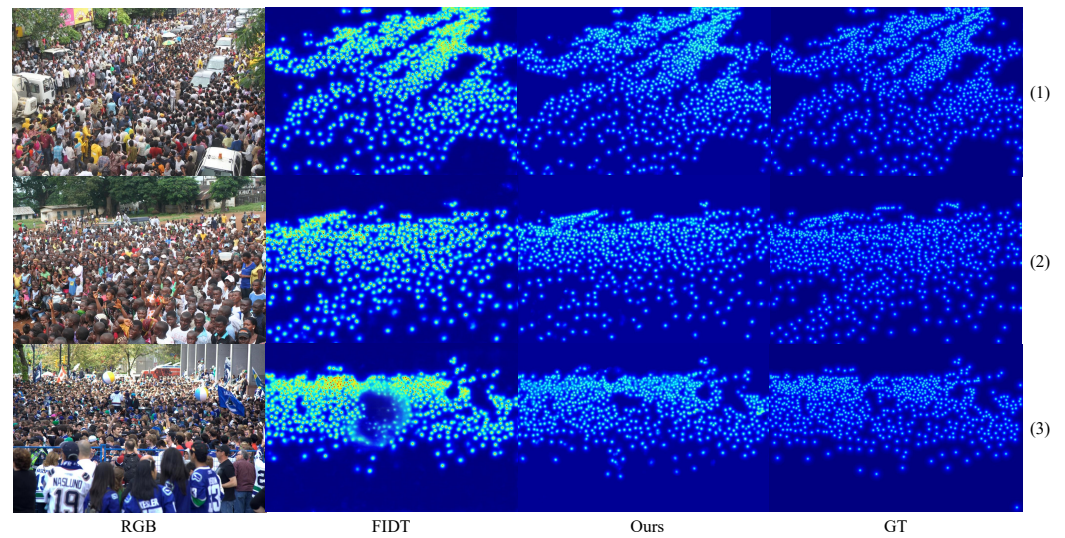


Figure 5. Qualitative results on ShanghaiTech Part A. In comparison to the FIDT, our method achieves more accurate and refined localization.

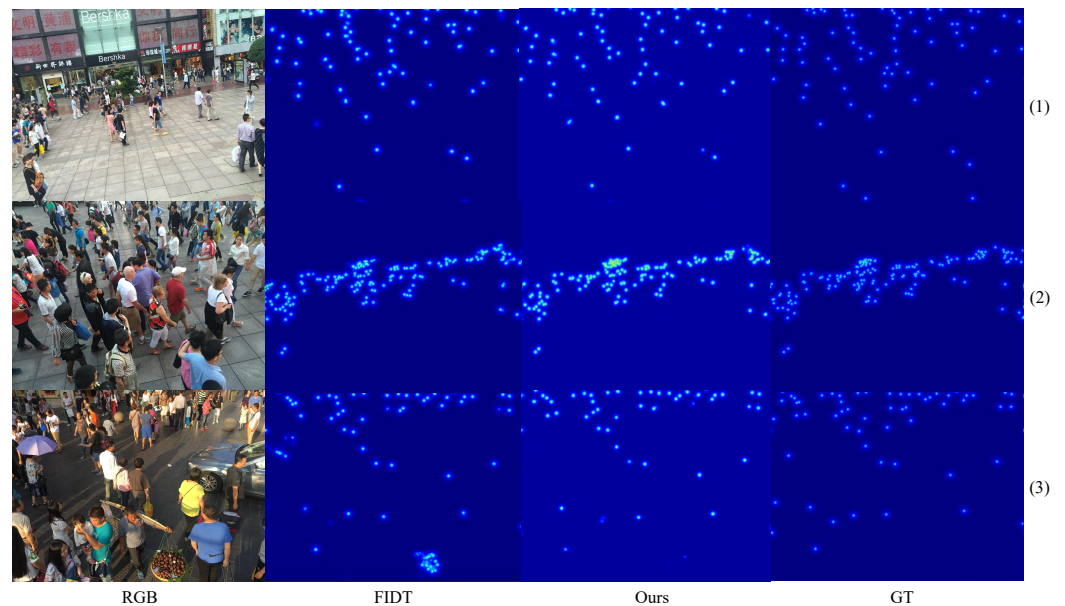


Figure 6. Qualitative results on ShanghaiTech Part B. In comparison to the FIDT, our method achieves more accurate and refined localization.

5. Conclusions

Our proposed Hierarchical Inverted Distance Transformer (HIDT) represents an advanced model. Unlike baseline models that utilize fully convolutional networks, our encoder employs global attention computation based on Transformers, while the decoder integrates dense skip connections with a Laplacian Pyramid Structure. This integration facilitates the capture and synthesis of multi-scale contextual information, significantly enhancing the feature extraction process. When evaluated on standard datasets, our model demonstrated substantial improvements in accuracy compared to other methods, highlighting its efficiency and potential for reliable crowd localization in complex environments. Looking ahead, we intend to further our research by integrating techniques akin to SAM [28]. Specifically, we aim to employ extensive pre-training on large datasets to enhance the accuracy of crowd localization data, thereby improving the practical utility and applicability of the model in real-world scenarios.

Author Contributions: S.C. and X.Q. were responsible for the methodology research, while S.C. and J.Y. prepared the manuscript. J.S. oversaw the strategic planning of the project and contributed to the critical review and refinement of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work received funding from several sources: the Natural Science Foundation of Xiamen, China (3502ZZ202373036); the National Natural Science Foundation of China (62006096, 42371457); and the Natural Science Foundation of Fujian Province (2022J01337, 2022J01819).

Data Availability Statement: This paper utilizes the ShanghaiTech_Crowd and JHU-CROWD++ datasets, which can be accessed and downloaded via www.datafountain.cn/datasets/5670 (accessed on 21 April 2024) and www.crowd-counting.com (accessed on 21 April 2024), respectively.

Conflicts of Interest: Author Xiangfeng Qiu was employed by Xiamen Kingtop Information Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Abousamra, S.; Hoai, M.; Samaras, D.; Chen, C. Localization in the crowd with topological constraints. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 872–881.
2. Liu, Y.; Shi, M.; Zhao, Q.; Wang, X. Point in, box out: Beyond counting persons in crowds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 6469–6478.
3. Sam, D.B.; Peri, S.V.; Sundararaman, M.N.; Kamath, A.; Babu, R.V. Locate, size, and count: Accurately resolving people in dense crowds via detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2739–2751.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
5. Liang, D.; Xu, W.; Zhu, Y.; Zhou, Y. Focal inverse distance transform maps for crowd localization. *IEEE Trans. Multimed.* **2022**, *25*, 6040–6052. [[CrossRef](#)]
6. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 5693–5703.
7. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [[CrossRef](#)]
8. Sindagi, V.A.; Yasarla, R.; Patel, V.M. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2594–2609. [[CrossRef](#)] [[PubMed](#)]
9. Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N.; Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 532–546.
10. Khan, A.; Ali Shah, J.; Kadir, K.; Albattah, W.; Khan, F. Crowd monitoring and localization using deep convolutional neural network: A review. *Appl. Sci.* **2020**, *10*, 4781. [[CrossRef](#)]
11. Hassen, K.B.A.; Machado, J.J.; Tavares, J.M.R. Convolutional neural networks and heuristic methods for crowd counting: A systematic review. *Sensors* **2022**, *22*, 5286. [[CrossRef](#)] [[PubMed](#)]
12. Ma, Y.; Sanchez, V.; Guha, T. CLIP-EBC: CLIP Can Count Accurately through Enhanced Blockwise Classification. *arXiv* **2024**, arXiv:2403.09281.
13. Xu, Y.; Zhong, Z.; Lian, D.; Li, J.; Li, Z.; Xu, X.; Gao, S. Crowd counting with partial annotations in an image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15570–15579.
14. Yang, Y.; Li, G.; Wu, Z.; Su, L.; Huang, Q.; Sebe, N. Weakly-supervised crowd counting learns from sorting rather than locations. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–17.
15. Li, Y.; Zhang, X.; Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1091–1100.
16. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Learning from synthetic data for crowd counting in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8198–8207.
17. Xu, C.; Qiu, K.; Fu, J.; Bai, S.; Xu, Y.; Bai, X. Learn to scale: Generating multipolar normalized density maps for crowd counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8382–8390.
18. Sindagi, V.A.; Yasarla, R.; Patel, V.M. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1221–1231.

19. Olmschenk, G.; Tang, H.; Zhu, Z. Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling. *arXiv* **2019**, arXiv:1902.05379.
20. Liu, L.; Qiu, Z.; Li, G.; Liu, S.; Ouyang, W.; Lin, L. Crowd counting with deep structured scale integration network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1774–1783.
21. Sindagi, V.A.; Patel, V.M. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1002–1012.
22. Ma, Z.; Wei, X.; Hong, X.; Gong, Y. Bayesian loss for crowd count estimation with point supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6142–6151.
23. Wan, J.; Wang, Q.; Chan, A.B. Kernel-based density map generation for dense object counting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1357–1370. [[CrossRef](#)] [[PubMed](#)]
24. Wan, J.; Chan, A. Modeling noisy annotations for crowd counting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3386–3396.
25. Wang, B.; Liu, H.; Samaras, D.; Nguyen, M.H. Distribution matching for crowd counting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1595–1607.
26. Wang, Y.; Hou, J.; Hou, X.; Chau, L.P. A self-training approach for point-supervised object detection and counting in crowds. *IEEE Trans. Image Process.* **2021**, *30*, 2876–2887. [[CrossRef](#)] [[PubMed](#)]
27. Xu, C.; Liang, D.; Xu, Y.; Bai, S.; Zhan, W.; Bai, X.; Tomizuka, M. Autoscale: Learning to scale for crowd counting. *Int. J. Comput. Vis.* **2022**, *130*, 405–434. [[CrossRef](#)]
28. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.