

An End-to-End Transformer Model for Crowd Localization

Dingkang Liang¹, Wei Xu², and Xiang Bai^{1,†}

¹ Huazhong University of Science and Technology, Wuhan 430074, China
{dkliang,xbai}@hust.edu.cn

² Beijing University of Posts and Telecommunications, Beijing 100876, China
xuwei2020@bupt.edu.cn

[†]Corresponding Author

Abstract. Crowd localization, predicting head positions, is a more practical and high-level task than simply counting. Existing methods employ pseudo-bounding boxes or pre-designed localization maps, relying on complex post-processing to obtain the head positions. In this paper, we propose an elegant, end-to-end **C**rowd **L**ocalization **T**Ransformer named CLTR that solves the task in the regression-based paradigm. The proposed method views the crowd localization as a direct set prediction problem, taking extracted features and trainable embeddings as input of the transformer-decoder. To reduce the ambiguous points and generate more reasonable matching results, we introduce a KMO-based Hungarian matcher, which adopts the nearby context as the auxiliary matching cost. Extensive experiments conducted on five datasets in various data settings show the effectiveness of our method. In particular, the proposed method achieves the best localization performance on the NWPU-Crowd, UCF-QNRF, and ShanghaiTech Part A datasets.

Keywords: Crowd localization, crowd counting, transformer

1 Introduction

Crowd localization, a fundamental subtask of crowd analysis, aims to provide the location of each instance. Here, the location means the center points of heads because annotating the bounding box for each head is expensive and laborious in dense scenes. Thus, most crowd datasets only provide point-level annotations. A powerful crowd localization algorithm can give great potential for similar tasks, *e.g.*, crowd tracking [45], object counting [16,5], and object localization [46,3].

The mainstream crowd localization methods can be generally categorized into detection-based (Fig. 1(a)) and map-based (Fig. 1(b)) methods. The detection-based methods [31,24] utilize nearest-neighbor head distances to initialize the pseudo ground truth (GT) bounding boxes. However, these detection-based

Project page at <https://dk-liang.github.io/CLTR/>.

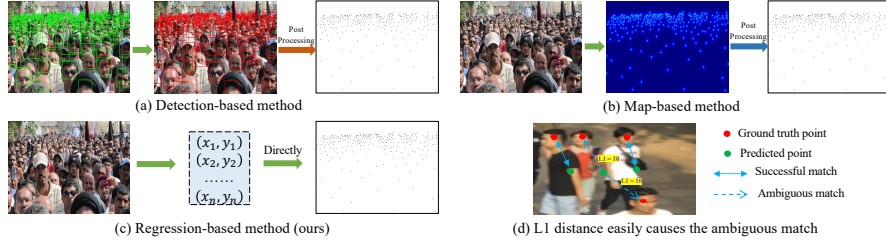


Fig. 1. (a) Detection-based methods, using the predefined pseudo GT bounding boxes. (b) Map-based methods, using high-resolution representation and complicated post-processing. (c) Regression-based methods, mapping the input images to the point coordinates directly. (d) Due to lack of context, the $L1$ distance easily causes the ambiguous match pair.

methods can not report satisfactory performance. Moreover, the heuristic non-maximum suppression (NMS) is used to remove the negative predictions. Most crowd localization methods [1,46] are map-based because it has relatively higher localization accuracy. Nevertheless, the map-based methods still suffer some inevitable problems. For instance, complex multi-scale representation is necessary to generate sharp maps. Besides, they adopt non-differentiable post-processing (*e.g.*, find-maxima) to extract the location, which precludes end-to-end training.

In contrast, the regression-based methods, directly regressing the coordinates, are more straightforward than the detection-based and map-based methods, as shown in Fig. 1(c). The benefits of regression-based can be summarized as two folds. (1) It is simple, elegant, and end-to-end trainable since it does not need pre-processing (*e.g.*, pseudo GT boxes or maps generation) and post-processing (*e.g.*, NMS or find-maxima). (2) It does not rely on complex multi-scale fusion mechanisms to generate high-quality feature maps.

Recently, we have witnessed the rise of Transformer [2] in computer vision. A pioneer is DETR [2], an end-to-end trainable detector without NMS, which models the relations of the object queries and context via Transformer and achieves competitive performance only using a single-level feature map. This simple and effective detection method gives rise to a question: *can crowd localization be solved with such a simple model as well?*

Our answer is: “Yes, such a framework can be applied to crowd localization.” Indeed, it is nothing special to directly apply the DETR-based pipeline in crowd localization. However, crowd localization is quite different from object detection. DETR shows terrible performance in the crowd localization task, attributed to the intrinsic limitation of the matcher. Specifically, the key component in DETR is the $L1$ -based Hungarian matcher, which measures $L1$ distance and bounding box IoU to match the prediction-GT bounding box pairs, showing superior performance in object detection. However, no bounding box is given in crowd datasets, and more importantly, $L1$ distance easily gives rise to ambiguous matching in the point-to-point pairs (*i.e.*, a point that can belong to

multiple *gts* simultaneously as shown in Fig. 1(d)). The main reasons are two-fold: (1) The $L1$ -based Hungarian is a local view without context. (2) Different from the object detection, the crowd images only contain one category (heads), and the dense heads usually have similar textures, reporting close confidence, confusing the matcher. To this end, we introduce the k -nearest neighbors (KNN) matching objective named KMO as an auxiliary matching cost. The KMO-based Hungarian considers the context from nearby heads, which helps to reduce the ambiguous points and generate more reasonable matching results.

In summary, the main contributions of this paper are two-fold: 1) We propose an end-to-end **C**rowd **L**ocalization **T**Ransformer framework named CLTR, which formulates the crowd localization as a point set prediction task. CLTR significantly simplifies the crowd localization pipeline by removing pre-processing and post-processing. 2) We introduce the KMO-based Hungarian bipartite matching, which takes the context from nearby heads as an auxiliary matching cost. As a result, the matcher can effectively reduce the ambiguous points and generate more reasonable matching results.

Extensive experiments are carried out on five challenge datasets, and significant improvements from KMO-based Hungarian matcher indicate its effectiveness. In particular, just with a single-scale and low-resolution ($\frac{1}{32}$ of input images) feature map, CLTR can achieve state-of-the-art or highly competitive localization performance.

2 Related Works

2.1 Detection-based methods

The detection-based methods [31,24,44] mainly follow the pipeline of Faster RCNN [29]. Specifically, PSDDN [24] utilizes the nearest neighbor distance to initialize the pseudo bounding boxes and update the pseudo boxes by choosing smaller predicted boxes in the training phase. LSC-CNN [31] also uses a similar mechanism to generate the pseudo bounding boxes and propose a new winner-take-all loss for better training at higher resolutions. These methods [31,24,44] usually use NMS to filter the predicted boxes, which is not end-to-end trainable.

2.2 Map-based methods

Map-based methods are the mainstream of the crowd localization task. Idress *et al.* [13], and Gao *et al.* [7] utilize small Gaussian kernel density maps, and the head locations are equal to the maxima of density maps. Even though using the small kernel can generate sharp density maps, it still exists overlaps in the extremely dense region, making the head location undistinguishable. To solve this, some methods [46,18,8,1] focus on designing new maps to handle the extremely dense region, such as the distance label map [46], Focal Inverse Distance Transform Map (FIDTM) [18] and Independent Instance Map (IIM) [8]. These methods can effectively avoid overlap in the dense region, but they need post-processing (“find-maxima”) to extract the instance location and rely on multi-scale feature maps, which is not simple and elegant.

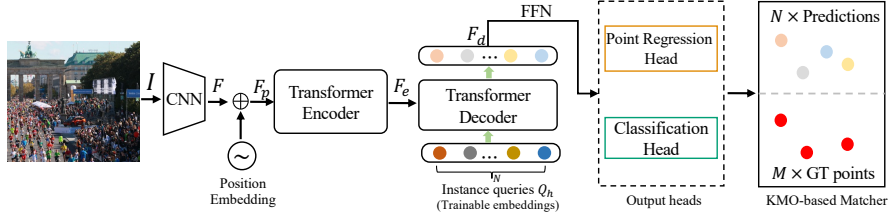


Fig. 2. The overview of our CLTR. First, the input image I is fed to the CNN-based backbone to extract the features F . Second, the features F are added position embedding, resulting in F_p , fed to the transformer-encoder layers, outputting F_e . Third, we define $N \times$ trainable embeddings Q_h as query, F_e as key, and transformer decoder takes the Q_h and F_e as input to generate the decoded feature F_d . Finally, the F_d can be decoupled to the point coordinate and corresponding confidence score.

2.3 Regression-based methods

Just a few research works focus on regression-based. We note a recent paper [34], P2PNet, also a regression-based framework for crowd localization. But this is a concurrent work that has appeared while this manuscript is under preparation. P2PNet [34] defines surrogate regression on a large set of proposals, and the model relies on pre-processing, such as producing $8 \times W \times H$ point proposals (anchors). In contrast, our method replaces massive fixed point proposals with a few trainable instance queries, which is more elegant and unified.

2.4 Visual transformer

Recently, visual transformers [4,37,23,2,28] have gone viral in computer vision. In particular, DETR [2] utilizes the Transformer-decoder to model object detection in an end-to-end pipeline, successfully removing the need for post-processing. Based on DETR [2], Conditional DETR [28] further adopts the spatial queries and keys to a band containing the object extremity or a region, accelerating the convergence of DETR [2]. In the crowd analysis, Liang *et al.* [17,26] propose TransCrowd, which reformulates the weakly-supervised counting problem from a sequence-to-count perspective. Several methods [36,35] demonstrate the power of transformers in point-supervised crowd counting setup. Method [6] adopts the IIM [8] in the swin transformer [25] to implement crowd localization.

3 Our Method

The overview of our method is shown in Fig. 2. The proposed method is an end-to-end network, directly predicting all instances at once without additional pre-processing and post-processing. The approach consists of a CNN-based backbone, a transformer encoder, a transformer decoder, and a KMO-based matcher. Given an image by $I \in \mathbb{R}^{H \times W \times 3}$, where H , W are the height and width of the image,

- The CNN-based backbone first extracts the feature maps $F \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C}$ from the input image I . To verify the effectiveness of our method, the F is only a single-scale feature map without feature aggregation.
- The feature maps F are then flattened into a 1D sequence with positional embedding, and the channel dimension is reduced from C to c , which results in $F_p \in \mathbb{R}^{\frac{HW}{32^2} \times c}$. The transformer-encoder layers take the F_p as input and output encoded features F_e .
- Next, the transformer-decoder layers take the trainable head queries Q_h and encoded features F_e as input and interact with each other via cross attention to generate the decoded embedding F_d , which contains the point (person’s head) and category information.
- Finally, the decoded embeddings F_d are decoupled to the point coordinates and confidence scores by a point regression head and a classification head, respectively.

3.1 Transformer Encoder

We use a 1×1 convolution to reduce the channel dimension of the extracted feature maps F from $\mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C}$ to $\mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times c}$ (c set as 256). Due to the transformer-encoder adopt a 1D sequence as input, we reshape the extracted features F and add position embedding, resulting in $F_p \in \mathbb{R}^{\frac{HW}{32^2} \times c}$. The F_p are then fed into the transformer-encoder layers to generate the encoded features F_e . Here the encoder contains many encoder layers, and each layer consists of a self-attention (SA) layer and a feed-forward (FC) layer. The SA consists of three inputs, including query (Q), key (K), and value (V), defined as follow:

$$SA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{c}})V, \quad (1)$$

where Q , K and V are obtained from the same input Z (e.g., $Q = ZW_Q$). In particular, we use the multi-head self-attention (MSA) to model the complex feature relation, which is an extension with several independent SA modules: $MSA = [SA_1; SA_2; \dots; SA_m]W_O$, where W_O is a re-projection matrix and m is the number of attention heads set as 8.

3.2 Transformer Decoder

The transformer-decoder consists of many decoder layers, and each layer is composed of three sub-layers: (1) a self-attention (SA) layer. (b) a cross attention (CA) layer. (3) a feed-forward (FC) layer. The SA and FC are the same as the encoder. The CA module takes two different embeddings as input instead of the same inputs in SA . Let us denote two embeddings as X and Z , and the CA can be written as $CA = SA(Q = XW_Q, K = ZW_K, V = ZW_V)$. Following [28], each decoder layer takes a set of trainable embedding ($Q_h \in \mathbb{R}^{N \times c}$) as query and the features from the last encoder layers as key and value. The decoder output the decoded features F_d , which are used to predict the point coordinates (regression head) and confidence scores (classification head).

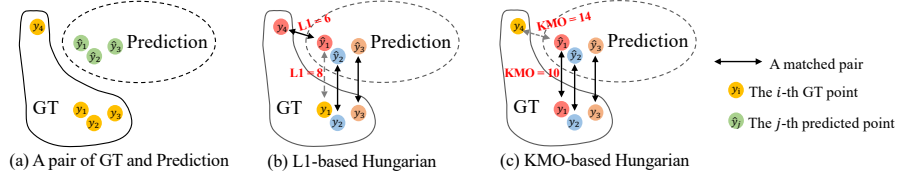


Fig. 3. (a) A pair of GT and predictions. (b) The $L1$ -based Hungarian generate unsatisfactory matching results. (c) The proposed KMO-based Hungarian models the context as the matching cost, generating more reasonable matching results.

3.3 KMO-based Matcher

To train the model, we need to match the predictions and GT by one-to-one, and the unmatched predicted points are considered to the “background” class.

Let us denote the predicted points set as $\hat{y} = \{\hat{y}_j\}_{j=1}^N$ and GT points set as $y = \{y_i\}_{i=1}^M$. N and M refer to the number of predicted heads and GT, respectively. N is larger than M to ensure each GT matches a prediction, and the rest of the predictions match failed can be classified as negative. Next, we need to find a bipartite matching between these two sets with the lowest cost. A straightforward way is to take the $L1$ distance and confidence as matching cost:

$$L_m(y_i, \hat{y}_j) = \|y_i^p - \hat{y}_j^p\|_1 - \hat{C}_j, i \in M, j \in N, \quad (2)$$

where $\|\cdot\|_1$ means the $L1$ distance and $\hat{C}_j$ is the confidence of the j -th predicted point. y_i^p is a vector that defines the i -th GT point coordinates. Accordingly, $\hat{y}_j^p$ is formed as the point coordinates of j -th predicted head. Based on the L_m , we can utilize the Hungarian [14] to implement one-to-one matching. However, we find that merely taking the $L1$ with confidence maybe generate unsatisfactory matching results (seen Fig. 1(d)). Another toy example is shown in Fig. 3, given a pair of GT and prediction set (Fig. 3(a)), from the whole perspective, the $\hat{y}_1$ should match the y_1 ideally (just like $\hat{y}_2$ match y_2). Using Eq. 2 for matching cost¹, it will match the $\hat{y}_1$ and y_4 since the $L1$ -based Hungarian lack of context information. Thus, we introduce the KMO-based Hungarian, which effectively utilizes the context as auxiliary matching cost, formulated as L_m^k :

$$L_m^k(y_i, \hat{y}_j) = \|y_i^p - \hat{y}_j^p\|_1 + \|y_i^k - \hat{y}_j^k\|_1 - \hat{C}_j, \quad (3)$$

$$y_i^k = \frac{1}{k} \sum_{k=1}^k d_i^k, \quad \hat{y}_j^k = \frac{1}{k} \sum_{k=1}^k \hat{d}_j^k,$$

where d_i^k means the distance between i -th GT point and its k -th neighbour. y_i^k refer to the average neighbour-distance of the i -th GT point. $\hat{d}_j^k$ and $\hat{y}_j^k$ have similar definitions as d_i^k and y_i^k , respectively. Taking inspirations from [24,31],

¹ Here, we ignore the $\hat{C}_j$ for simply illustrating since heads usually report similar confidence score.

in our experiments, $\hat{y}_j^k$ is predicted by the network. The proposed L_m^k , revisiting the label assignment from a context view, turns to find the whole optimum. As shown in Fig. 3(c), the proposed KMO-based Hungarian makes sure the $\hat{y}_1$ successfully matches the y_1 . Regarding the predictions as a point set containing the geometric relationships. Assignment on Fig. 3(b) is not wrong, but it is a local view without context, and it will break the internal geometric relationships of the point set. Assignment on Fig. 3(c) considers the context information from the nearby heads, pursuing the whole optimum and maintaining the geometric relationships of the point set, making the model easier to be optimized, which is more reasonable. For the case in Fig. 1(d), when multiple gts tend to match the same predicted point, the KMO-based Hungarian will resolve their conflicts by using the context information. Note that the matcher is just used in the training phase.

3.4 Loss function

After obtaining the one-to-one matching results, we calculate the loss for back-propagate. We make point predictions directly. The loss consists of point regression and classification. For the point regression, we employ the commonly-used L_1 loss, defined as:

$$L_{loc} = \|y_i^p - \hat{y}_{\sigma(i)}^p\|_1, \quad (4)$$

where $\hat{y}_{\sigma(i)}^p$ is the matched subset from y_i^p by using the proposed KMO-based Hungarian. It is noteworthy that we normalize all ground truth point range to $[0, 1]$ for scale invariance. We utilize the focal loss as the classification loss L_{cls} , and the final loss L is defined as:

$$L = L_{cls} + \lambda L_{loc}, \quad (5)$$

where λ is a balance weight, set as 2.5. These two losses are normalized by the number of instances inside the batch.

4 Experiments

4.1 Implementation details

We use the ResNet50 [10] as the backbone. The number of transformer encoder layers and decoder layers are both set to 6. The N is set to 500 (number of instance queries Q_h). We augment the training data using random cropping, random scaling, and horizontal flipping with a 0.5 probability. The crop size is set as 128×128 for ShanghaiTech Part A, 256×256 for the rest datasets. We use Adam with the learning rate of $1e-4$ to optimize the model. For the large-scale datasets (*i.e.*, UCF-QNRF, JHU-Crowd++, and NWPU-Crowd), we ensure the longer size is less than 2048, keeping the original aspect ratio. The batch size is set to 16. k is set as 4 for all datasets. During the testing phase, each image is split into non-overlapped patches (size same as training phase). Zero padding is adopted if a cropped patch is smaller than the predefined size. And a simple confidence threshold (set to 0.35) is used to filter the “background” class.

Table 1. Localization performance on NWPU-Crowd dataset. * means the methods rely on box-level instead of point-level annotations.

Method	Validation set			Test set		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Faster RCNN* [29]	96.4%	3.8%	7.3%	95.8%	3.5%	6.7%
TinyFaces* [11]	54.3%	66.6%	59.8%	52.9%	61.1%	56.7%
TopoCount* [1]	-	-	-	69.5%	68.7%	69.1%
GPR [7]	61.0%	52.2%	56.3%	55.8%	49.6%	52.5%
RAZ_Loc [19]	69.2%	56.9%	62.5%	66.6%	54.3%	59.8%
AutoScale_loc [46]	70.1%	63.8%	66.8%	67.3%	57.4%	62.0%
Crowd-SDNet [44]	-	-	-	65.1%	62.4%	63.7%
GL [39]	-	-	-	80.0%	56.2%	66.0%
CLTR (ours)	73.9%	71.3%	72.6%	69.4%	67.6%	68.5%

Table 2. Localization performance on the UCF-QNRF dataset. We report the Average Precision, Recall, and F1-measure at different thresholds σ : (1, 2, 3, ..., 100) pixels.

Method	Av.Precision	Av.Recall	F1-measure
CL [13]	75.80%	59.75%	66.82%
LCFCN [15]	77.89%	52.40%	62.65%
Method in [30]	75.46%	49.87%	60.05%
LSC-CNN[31]	75.84%	74.69%	75.26%
AutoScale_loc [46]	81.31%	75.75%	78.43%
GL [39]	78.20%	74.80%	76.40%
TopoCount [1]	81.77%	78.96%	80.34%
CLTR (ours)	82.22%	79.75%	80.97%

4.2 Dataset

We evaluate our method on five challenging public datasets, each being elaborated below.

NWPU-Crowd [42] is a large-scale dataset collected from various scenes, consisting of 5,109 images. The images are randomly split into training, validation, and test sets, which contain 3109, 500, and 1500 images, respectively. This dataset provides point-level and box-level annotations.

JHU-Crowd++ [33] is a challenging dataset containing 4372 crowd images. This dataset consists of 2272 training images, 500 validation images, and 1600 test images. And the total number of people in each image ranges from 0 to 25791.

UCF-QNRF [13], a dense dataset, contains 1535 images (1201 for training and 334 for testing) and about one million annotations. The average number of pedestrians per image is 815, and the maximum number reaches 12865.

ShanghaiTech [48] is divided into Part A and Part B. Part A consists of 300 training images and 182 test images. Part B consists of 400 training images and 316 test images.

4.3 Evaluation Metrics

This paper mainly focuses on crowd localization, and counting is an incidental task, *i.e.*, the total count is equal to the number of predicted points.

Table 3. Comparison of the localization performance on the Part A dataset.

Method	$\sigma = 4$			$\sigma = 8$		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
LCFCN[15]	43.3%	26.0%	32.5%	75.1%	45.1%	56.3%
Method in [30]	34.9%	20.7%	25.9%	67.7%	44.8%	53.9%
LSC-CNN [31]	33.4%	31.9%	32.6%	63.9%	61.0%	62.4%
TopoCount [1]	41.7%	40.6%	41.1%	74.6%	72.7%	73.6%
CLTR (ours)	43.6%	42.7%	43.2%	74.9%	73.5%	74.2%

Localization Metrics. In this work, we utilize the Precision, Recall, and F1-measure as the localization metrics, following [42,13]. If the distance between a predicted point and GT point is less than the predefined distance threshold σ , this predicted point will be treated as True Positive (TP). For the NWPU-Crowd dataset [42], containing the box-level annotations, we set σ to $\sqrt{w^2 + h^2}/2$, where w and h are the width and height of each head, respectively. For the ShanghaiTech dataset, we utilize two fixed thresholds, including $\sigma = 4$ and $\sigma = 8$. For the UCF-QNRF, we use various threshold ranges from [1, 100], following CL [13].

Counting Metrics. The Mean Absolute Error (MAE) and Mean Square Error (MSE) are used as counting metrics, defined as: $MAE = \frac{1}{N_c} \sum_{i=1}^{N_c} |P_i - G_i|$, $MSE = \sqrt{\frac{1}{N_c} \sum_{i=1}^{N_c} |P_i - G_i|^2}$, where N_c is the total number of images, P_i and G_i are the predicted and GT count of the i -th image, respectively.

5 Results and Analysis

5.1 Crowd Localization

We first evaluate the localization performance with some state-of-the-art localization methods [39,1,46], as shown in Table 1, Table 2, and Table 3. For the NWPU-Crowd (see Table 1), a large-scale dataset, our CLTR outperforms GL [39] and AutoScale [46] at least 5.8% (*resp.* 2.5%) for F1-measure on the validation set (*resp.* test set). It is noteworthy that this dataset provides precise box-level annotations, and the TopoCount [1] relies on the labeled box in the training phase instead of using the point-level annotations. Even though our method is just based on point-annotation, a more weak label mechanism, it can still achieve competitive performance against the TopoCount [1] on the NWPU-Crowd (test set). For the dense dataset, UCF-QNRF (see Table 2), our method achieves the best Average Precision, Average Recall and F1-measure. For the ShanghaiTech Part A (see Table 3), a sparse dataset, our CLTR outperforms the state-of-the-art method TopoCount [1] by 2.1% F1-measure for the strict setting ($\sigma = 4$), and still get ahead for the less strict settings ($\sigma = 8$). These results demonstrate that the proposed method can cope with various scenes, including large-scale, dense and sparse scenes. Note that all of other localization methods [46,39,1] adopt multi-scale or higher-resolution feature that potentially

Table 4. Counting results of various methods on the NWPU validation and test sets.

Method	Output Position Information	Validation set		Test set	
		MAE	MSE	MAE	MSE
MCNN [48]	✗	218.5	700.6	232.5	714.6
CSRNet [16]	✗	104.8	433.4	121.3	387.8
CAN [22]	✗	93.5	489.9	106.3	386.5
SCAR [9]	✗	81.5	397.9	110.0	495.3
BL [27]	✗	93.6	470.3	105.4	454.2
SFCN [43]	✗	95.4	608.3	105.4	424.1
DM-Count [41]	✗	70.5	357.6	88.4	388.6
RAZ.loc [19]	✓	128.7	665.4	151.4	634.6
AutoScale.loc [46]	✓	97.3	571.2	123.9	515.5
TopoCount [1]	✓	-	-	107.8	438.5
GL [39]	✓	-	-	79.3	346.1
CLTR (ours)	✓	61.9	246.3	74.3	333.8

Table 5. Comparison of the counting performance on the UCF-QNRF, ShanghaiTech Part A, and Part B datasets.

Method	Output Position Information	QNRF		Part A		Part B	
		MAE	MSE	MAE	MSE	MAE	MSE
CSRNet [16]	✗	-	-	68.2	115.0	10.6	16.0
L2SM [47]	✗	104.7	173.6	64.2	98.4	7.2	11.1
DSSI-Net [21]	✗	99.1	159.2	60.6	96.0	6.9	10.3
MBTTBF [32]	✗	97.5	165.2	60.2	94.1	8.0	15.5
BL [27]	✗	88.7	154.8	62.8	101.8	7.7	12.7
AMSNet [12]	✗	101.8	163.2	56.7	93.4	6.7	10.2
LibraNet [20]	✗	88.1	143.7	55.9	97.1	7.3	11.3
KDMG [40]	✗	99.5	173.0	63.8	99.2	7.8	12.7
NoisyCC [38]	✗	85.8	150.6	61.9	99.6	7.4	11.3
DM-Count [41]	✗	85.6	148.3	59.7	95.7	7.4	11.8
CL [13]	✓	132.0	191.0	-	-	-	-
RAZ.loc+ [19]	✓	118.0	198.0	71.6	120.1	9.9	15.6
PSDDN [24]	✓	-	-	65.9	112.3	9.1	14.2
LSC-CNN [31]	✓	120.5	218.2	66.4	117.0	8.1	12.7
TopoCount [1]	✓	89.0	159.0	61.2	104.6	7.8	13.7
AutoScale.loc [46]	✓	104.4	174.2	65.8	112.1	8.6	13.9
GL [39]	✓	84.3	147.5	61.3	95.4	7.3	11.7
CLTR (ours)	✓	85.8	141.3	56.9	95.2	6.5	10.6

benefit our approach, which is currently not our focus and left as our future work.

5.2 Crowd Counting

In this section, we compare the counting performance with various methods (including density map regression-based and localization-based), as shown in Table 4, Table 5 and Table 6. Although our method only uses a single-scale and low-resolution ($\frac{1}{32}$ of input image) feature map, it can achieve state-of-the-art or highly competitive performance in all experiments. Specifically, our method achieves the first MAE and MSE on the NWPU-Crowd test set (see Table 4). Compared with the localization-based counting methods (the bottom part of

Table 6. Categorical counting results on JHU-Crowd++ dataset. Low, Medium, and High respectively indicate three categories based on different ranges: $[0, 50]$, $(50, 500]$, and $(500, +\infty)$. Weather means the degraded images (*e.g.*, haze, snow, rain).

Methods	Output Position Information	Low		Medium		High		Weather		Overall	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [48]	✗	97.1	192.3	121.4	191.3	618.6	1,166.7	330.6	852.1	188.9	483.4
CSRNET [16]	✗	27.1	64.9	43.9	71.2	356.2	784.4	141.4	640.1	85.9	309.2
JHU++ [33]	✗	14.0	42.8	35.0	53.7	314.7	712.3	120.0	580.8	71.0	278.6
LSC-CNN [31]	✗	10.6	31.8	34.9	55.6	601.9	1,172.2	178.0	744.3	112.7	454.4
BL [27]	✗	10.1	32.7	34.2	54.5	352.0	768.7	140.1	675.7	75.0	299.9
AutoScale_loc [46]	✓	13.2	30.2	32.3	52.8	425.6	916.5	-	-	85.6	356.1
TopoCount [1]	✓	8.2	20.5	28.9	50.0	282.0	685.8	120.4	635.1	60.9	267.4
GL [39]	✓	-	-	-	-	-	-	-	-	59.9	259.5
CLTR (ours)	✓	8.3	21.8	30.7	53.8	265.2	614.0	109.5	568.5	59.5	240.6

Table 7. Effect of transformer size (the number of layers and the number of trainable queries Q_h) on UCF-QNRF dataset.

Layers	N (Queries number)	Localization			Counting	
		P (%)	R (%)	F (%)	MAE	MSE
3	500	80.60%	79.44%	80.02%	88.4	149.9
6	500	82.22%	79.75%	80.97%	85.8	141.3
12	500	80.82%	79.41%	80.11%	87.7	150.3
6	300	80.61%	79.18%	79.89%	89.9	153.6
6	700	81.32%	79.38%	80.34%	86.8	146.4

Table 5), which can provide the position information, our method achieves the best counting performance in MAE and MSE on ShanghaiTech Part A and Part B datasets. On the UCF-QNRF dataset, our method achieves the best MSE and reports comparable MAE. On the JHU-Crowd++ dataset (Table 6), our method outperforms the state-of-the-art method GL [39] by a significant margin of 18.9 MSE. Furthermore, CLTR has superior performance on the extremely dense (the “High” part) and degraded set (the “Weather” part).

5.3 Visualizations

We further give some qualitative visualizations to analyze the effectiveness of our method, as shown in Fig. 4. The samples are selected from some typical scenes on the NWPU-Crowd dataset (validation set), including negative, sparse, extremely dense and dark scenes. In the first row, CLTR shows a strong robust on the negative sample (“dense fake humans”). CLTR performs well in different congested scenes, such as the sparse scene (row 2) and extremely dense scene (row 3). Additionally, we find that CLTR can also make promising localization results in dark scenes (row 4). These impressive visualizations demonstrate the effectiveness of our method in crowd localization and counting.

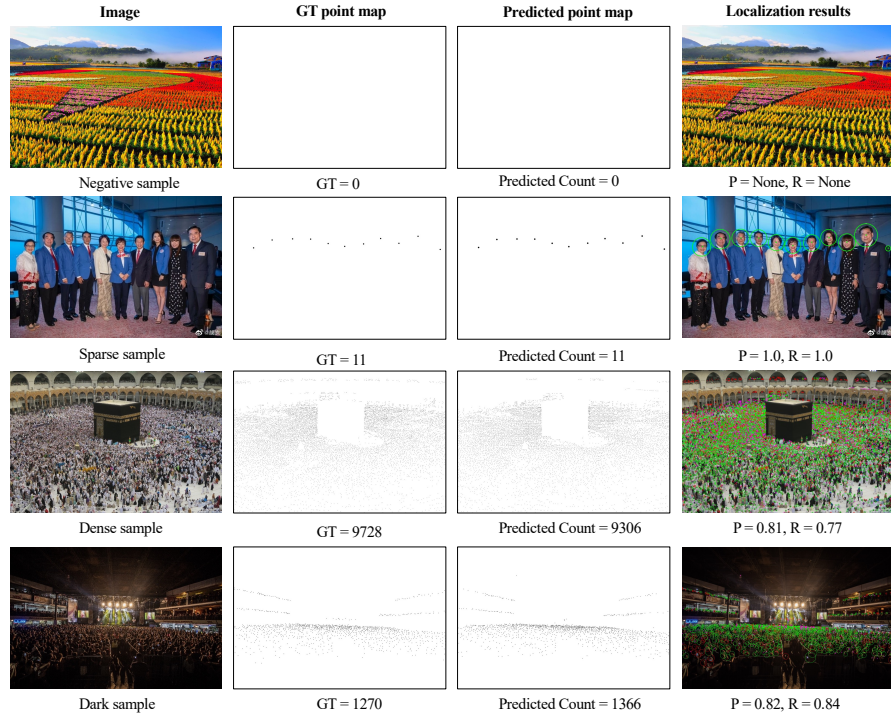


Fig. 4. Some examples from the NWPU-Crowd dataset (validation set). From left to right, there are images, GT points, predicted points, and localization results. Row 1, row 2, row 3, and row 4 refer to the negative, sparse, dense and dark samples, respectively. In the last column, P and R are the Precision and Recall, respectively. The green, red and magenta points denote true positive (TP), false negative (FN) and false positive (FP), respectively. It can be seen that the proposed method can effectively handle various scenes.

5.4 Ablation studies

The ablation studies are carried out on the UCF-QNRF dataset, a large and dense dataset, which can effectively avoid overfitting.

Effect of transformer. We first study the influence by changing the size of the transformer, including the number of encoder/decoder layers and trainable instance queries. As listed in Table 7, we find that when the layer and queries number are set to 6 and 500, the CLTR achieves the best performance. When the number of queries changes to 700 (resp. 300), the performance of MAE drops from 85.8 to 86.8 (resp. 89.9). We hypothesize that, by using a small number of queries, CLTR may lose potential heads, while using a large number of queries, CLTR may generate massive negative samples. We empirically find that all the pre-defined non-overlap patches contain less than 500 persons. The following ablation studies are organized using 6 transformer layers and 500 queries.

Table 8. The effectiveness of the proposed KMO-based Hungarian on UCF-QNRF dataset. L_m only adopts the $L1$ distance with confidence as matching cost, and L_m^k contains the proposed KMO.

Matching cost	Localization			Counting	
	Av.Precision(%)	Av.Recall(%)	F1-measure(%)	MAE	MSE
L_m	80.89%	79.17%	80.02%	91.3	157.4
L_m^k (ours)	82.22%	79.75%	80.97%	85.8	141.3

Table 9. The influence of using different numbers of nearest-neighbour on the UCF-QNRF dataset.

k	Localization			Counting	
	Av.Precision(%)	Av.Recall(%)	F1-measure(%)	MAE	MSE
3	81.46%	79.19%	80.31%	87.1	146.8
4	82.22%	79.75%	80.97%	85.8	141.3
5	81.52%	79.34%	80.42%	86.9	148.1

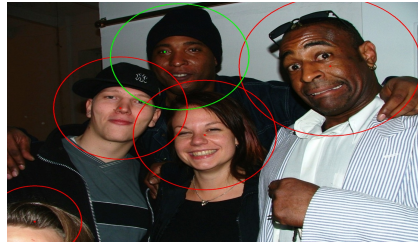
Effect of matching cost. We next study the impact of the proposed KMO, as shown in Table 8. When removing the KMO, we observe a significant performance drop for the counting (MAE from 85.8 to 91.3) and localization as well. We hypothesize that the $L1$ with classification can not provide a strong matching indicator, while the proposed KMO gives a direct signal to achieve great one-to-one matching based on whole-optimal.

Effect of K . We then study the effect of using different k (the number of nearest-neighbor), listed in Table 9. The proposed CLTR with different k consistently achieves improvement compared with the baseline, demonstrating the proposed KMO-based Hungarian’s effectiveness. When the k is set to 4, we find that the result achieves the best on the UCF-QNRF dataset. We then set the same k in all datasets without further fine-tuning, which works well. We also try to use a fixed radius around each point and take as many NN as they fall within that circle. However, the training time is unacceptable because calculating dynamic KNN in each circle is time-consuming.

The computational statistics. Finally, we report the Multiply-Accumulate Operations (MACs) and parameters, as listed in Table 10. Although the proposed method has the largest parameters (mainly from the transformer part), it still reports the smallest MACs. Speeding up our model is a future work that is worthy of being studied.

5.5 Limitations

Our method has some limitations. For instance, due to the CLTR crop a fixed-size (*i.e.*, 256×256) sub-image for training and testing, it may fail on extremely large heads, which are significantly larger than the crop size, as shown



(a) Large scale head, image size: 2048×1536



(b) “Fake” people in the background

Fig. 5. Failure cases analysis. (a) Large-scale head, significantly larger than the crop-size. (b) Confusing regions that do not need to localize. The green, red, and magenta denote true positive (TP), false negative (FN), and false positive (FP), respectively.

Table 10. The comparisons of complexity. The experiments are conducted on a 3090 GPU, and the size of the input image is 1024×768 .

Method	Parameters (M)	MACs (G)
LSC-CNN [31]	35.0	1244.3
AutoScale [46]	24.9	1074.6
TopoCount [1]	25.8	797.2
GL [39]	21.5	324.6
CLTR (ours)	43.4	157.2

in Fig. 5(a). This problem can be solved by resizing the image into a small resolution. Another case of unsatisfied localization is shown in Fig. 5(b), where there are some confused background regions (containing “fake” people that do not need localization). This failure case can be solved using more modalities, such as thermal images.

6 Conclusion

In this work, we propose an end-to-end crowd localization framework named CLTR, solving the task in the regression-based paradigm. The proposed method follows a one-to-one matching mechanism during the training phase. To achieve a good matching result, we propose the KMO-based Hungarian matcher, using the context information as an auxiliary matching cost. Our approach is simple yet effective. Experiments on five challenge datasets demonstrate the effectiveness of our methods. We hope our method can provide a new perspective for the crowd localization task.

Acknowledgment

This work was supported by National Key R&D Program of China (Grant No. 2018YFB1004602).

References

1. Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. In: Proc. of the AAAI Conf. on Artificial Intelligence (2021)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proc. of European Conference on Computer Vision. pp. 213–229. Springer (2020)
3. Chen, Y., Liang, D., Bai, X., Xu, Y., Yang, X.: Cell localization and counting using direction field map. *IEEE Journal of Biomedical and Health Informatics* **26**(1), 359–368 (2021)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Proc. of International Conference on Learning Representations (2020)
5. Du, D., Wen, L., Zhu, P., Fan, H., Hu, Q., Ling, H., Shah, M., Pan, J., Al-Ali, A., Mohamed, A., et al.: Visdrone-cc2020: The vision meets drone crowd counting challenge results. In: Proc. of European Conference on Computer Vision. pp. 675–691. Springer (2020)
6. Gao, J., Gong, M., Li, X.: Congested crowd instance localization with dilated convolutional swin transformer. arXiv preprint arXiv:2108.00584 (2021)
7. Gao, J., Han, T., Wang, Q., Yuan, Y.: Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. arXiv preprint arXiv:1912.03677 (2019)
8. Gao, J., Han, T., Yuan, Y., Wang, Q.: Learning independent instance maps for crowd localization. arXiv preprint arXiv:2012.04164 (2020)
9. Gao, J., Wang, Q., Yuan, Y.: Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing* (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2016)
11. Hu, P., Ramanan, D.: Finding tiny faces. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2017)
12. Hu, Y., Jiang, X., Liu, X., Zhang, B., Han, J., Cao, X., Doermann, D.: Nas-count: Counting-by-density with neural architecture search. In: Proc. of European Conference on Computer Vision (2020)
13. 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: Proc. of European Conference on Computer Vision (2018)
14. Kuhn, H.W.: The hungarian method for the assignment problem. *Naval research logistics quarterly* **2**(1-2), 83–97 (1955)
15. Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M.: Where are the blobs: Counting by localization with point supervision. In: Proc. of European Conference on Computer Vision (2018)
16. Li, Y., Zhang, X., Chen, D.: CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2018)
17. Liang, D., Chen, X., Xu, W., Zhou, Y., Bai, X.: Transcrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences* **65**(6), 1–14 (2022)
18. Liang, D., Xu, W., Zhu, Y., Zhou, Y.: Focal inverse distance transform maps for crowd localization and counting in dense crowd. arXiv preprint arXiv:2102.07925 (2021)

19. Liu, C., Weng, X., Mu, Y.: Recurrent attentive zooming for joint crowd counting and precise localization. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2019)
20. Liu, L., Lu, H., Zou, H., Xiong, H., Cao, Z., Shen, C.: Weighing counts: Sequential crowd counting by reinforcement learning. In: Proc. of European Conference on Computer Vision (2020)
21. Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L.: Crowd counting with deep structured scale integration network. In: Proc. of IEEE Intl. Conf. on Computer Vision (2019)
22. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2019)
23. Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S., Bai, X.: End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing* (2022)
24. Liu, Y., Shi, M., Zhao, Q., Wang, X.: Point in, box out: Beyond counting persons in crowds. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2019)
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. of IEEE Intl. Conf. on Computer Vision. pp. 10012–10022 (2021)
26. Liu, Z., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J., et al.: Visdrone-cc2021: the vision meets drone crowd counting challenge results. In: Proc. of IEEE Intl. Conf. on Computer Vision. pp. 2830–2838 (2021)
27. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proc. of IEEE Intl. Conf. on Computer Vision (2019)
28. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proc. of IEEE Intl. Conf. on Computer Vision. pp. 3651–3660 (2021)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proc. of Advances in Neural Information Processing Systems (2015)
30. Ribera, J., Güera, D., Chen, Y., Delp, E.J.: Locating objects without bounding boxes. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2019)
31. Sam, D.B., Peri, S.V., Sundararaman, M.N., Kamath, A., Radhakrishnan, V.B.: Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020)
32. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proc. of IEEE Intl. Conf. on Computer Vision (2019)
33. Sindagi, V.A., Yasarla, R., Patel, V.M.: Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020)
34. Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y.: Rethinking counting and localization in crowds: A purely point-based framework. In: Proc. of IEEE Intl. Conf. on Computer Vision. pp. 3365–3374 (2021)
35. Sun, G., Liu, Y., Probst, T., Paudel, D.P., Popovic, N., Van Gool, L.: Boosting crowd counting with transformers. *arXiv preprint arXiv:2105.10926* (2021)
36. Tian, Y., Chu, X., Wang, H.: Cctrans: Simplifying and improving crowd counting with transformer. *arXiv preprint arXiv:2109.14483* (2021)

37. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proc. of Intl. Conf. on Machine Learning. pp. 10347–10357. PMLR (2021)
38. Wan, J., Chan, A.: Modeling noisy annotations for crowd counting. *Advances in Neural Information Processing Systems* (2020)
39. Wan, J., Liu, Z., Chan, A.B.: A generalized loss function for crowd counting and localization. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 1974–1983 (2021)
40. Wan, J., Wang, Q., Chan, A.B.: Kernel-based density map generation for dense object counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020)
41. Wang, B., Liu, H., Samaras, D., Hoai, M.: Distribution matching for crowd counting. In: Proc. of Advances in Neural Information Processing Systems (2020)
42. Wang, Q., Gao, J., Lin, W., Li, X.: Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020)
43. Wang, Q., Gao, J., Lin, W., Yuan, Y.: Learning from synthetic data for crowd counting in the wild. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2019)
44. Wang, Y., Hou, J., Hou, X., Chau, L.P.: A self-training approach for point-supervised object detection and counting in crowds. *IEEE Transactions on Image Processing* **30**, 2876–2887 (2021)
45. Wen, L., Du, D., Zhu, P., Hu, Q., Wang, Q., Bo, L., Lyu, S.: Detection, tracking, and counting meets drones in crowds: A benchmark. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 7812–7821 (2021)
46. Xu, C., Liang, D., Xu, Y., Bai, S., Zhan, W., Bai, X., Tomizuka, M.: Autoscale: Learning to scale for crowd counting. *International Journal of Computer Vision* pp. 1–30 (2022)
47. Xu, C., Qiu, K., Fu, J., Bai, S., Xu, Y., Bai, X.: Learn to scale: Generating multipolar normalized density map for crowd counting. In: Proc. of IEEE Intl. Conf. on Computer Vision (2019)
48. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (2016)