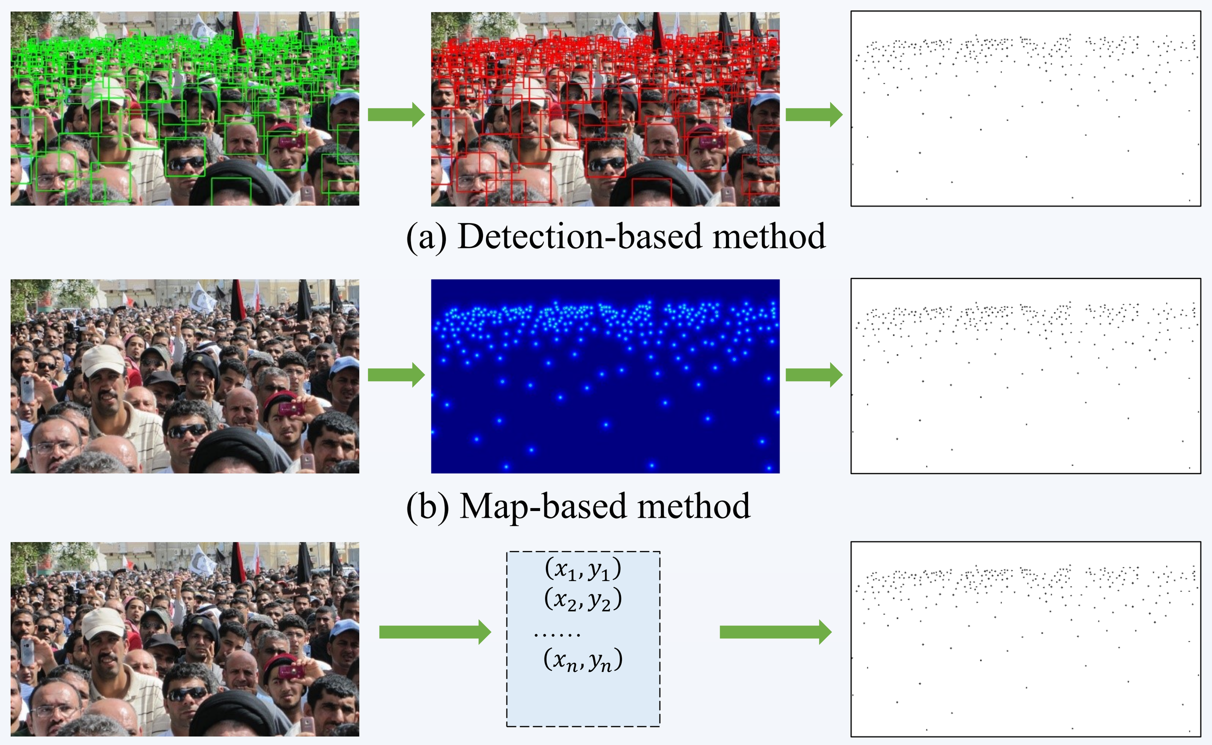
An End-to-End Transformer Model for Crowd Localization

Dingkang Liang[1], Wei Xu[2], Xiang Bai[1]

[1] Huazhong University of Science and Technology; [2] Beijing University of Posts and Telecommunications

MOTIVATION



(a) Detection-based method

(b) Map-based method

(c) Regression-based method

➢Crowd localization aims to provide the location of each instance.

➢The regression-based methods, directly regressing the coordinates, are more straightforward than the detection-based and map-based methods.

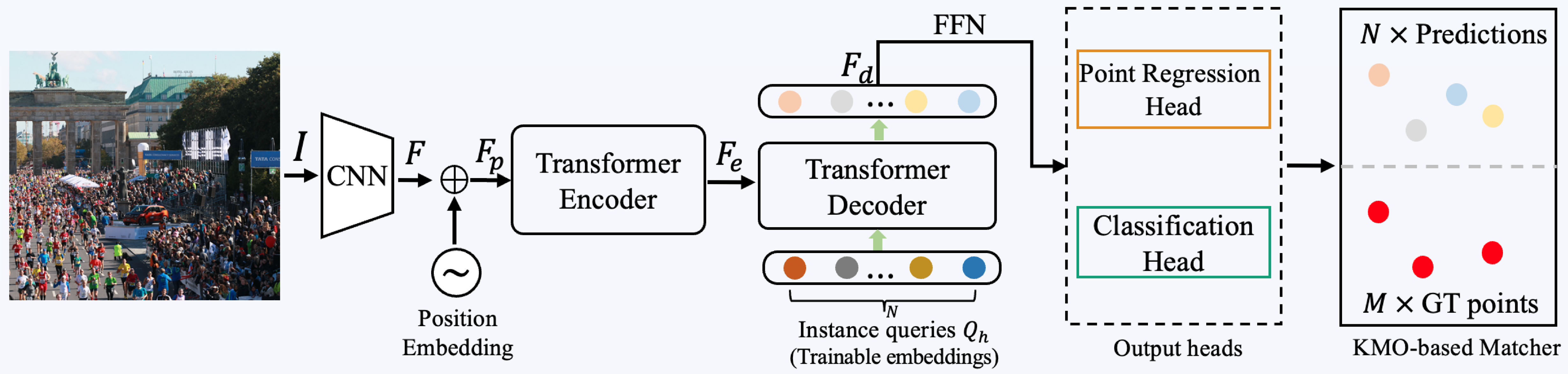
➢DETR shows terrible performance in the crowd localization task, attributed to the intrinsic limitation of the matcher. Due to lack of context, the $L1$ distance easily causes the ambiguous match pair.

CONTRIBUTION

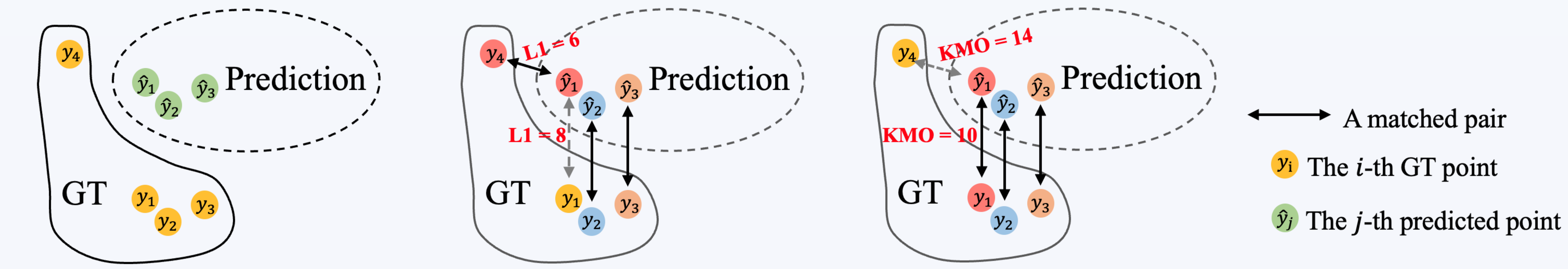
➢We propose an end-to-end Crowd Localization TRansformer framework named CLTR, which formulates the crowd localization as a point set prediction task.

➢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.

METHOD



➢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.



(a) A pair of GT and Prediction (b) L1-based Hungarian (c) KMO-based Hungarian

➢(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.

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

➢Merely taking the $L1$ with confidence will generate unsatisfactory matching results on specific cases.

$$L_m^k(y_i, \hat{y}_j) = \left\| y_i^p - \hat{y}_j^p \right\|_1 + \left\| y_i^k - \hat{y}_j^k \right\|_1 - \hat{C}_j,$$
$$y_i^k = \frac{1}{k}\sum_{k=1}^{k} d_i^k, \quad \hat{y}_j^k = \frac{1}{k}\sum_{k=1}^{k} \hat{d}_j^k,$$

➢The proposed KMO-based matcher, revisiting the label assignment from a context view, turns to find the whole optimum.

RESULTS

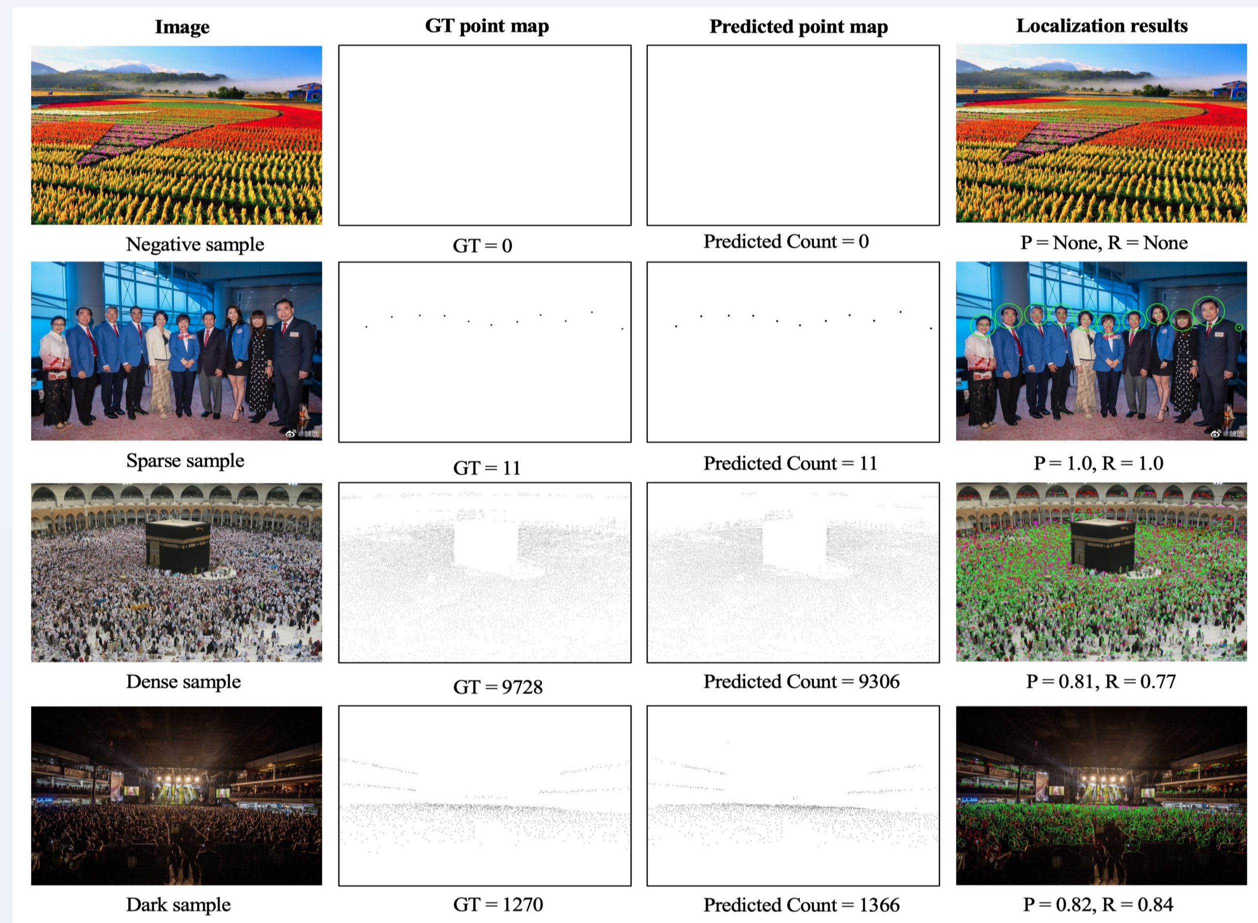
Method	Validation set P(%)	Validation set R(%)	Validation set F(%)	Test set P(%)	Test set R(%)	Test set F(%)
Faster RCNN* [29]	**96.4%**	3.8%	7.3%	**95.8%**	3.5%	6.7%
TinyFaces* [11]	54.3%	**66.6%**	**59.8%**	52.9%	61.1%	56.7%
TopoCount* [1]	-	-	-	69.5%	**68.7%**	**69.1%**
GPR [7]	61.0%	52.2%	56.3%	55.8%	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%**

Localization performance on NWPU dataset.

Method	Output Position Information	Validation set MAE	Validation set MSE	Test set MAE	Test set 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**

Counting performance on NWPU dataset.

VISUALIZATIONS



Some examples from the NWPU dataset. From left to right, there are images, GT points, predicted points, and localization results.

ACKNOWLEDGEMENT

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