Dynamic Mixture of Counter Network for Location-Agnostic Crowd Counting

Mingjie Wang
School of Science
Zhejiang Sci-Tech University
mwang20@uoguelph.ca

Hao Cai
School of Computer Science
Memorial University of Newfoundland
hc1864@mun.ca

Yong Dai
Tencent AI Lab
daiyongya@outlook.com

Minglun Gong*
School of Computer Science
University of Guelph
minglun@uoguelph.ca

Abstract

Crowd counting has attracted increasing attentions in recent years due to its challenges and wide societal applications. Despite persevering efforts made by the research community, most of existing methods require a large amount of location-level annotations. Collecting such type of fine-granularity supervisory signals is extremely time-consuming and labour-intensive, thereby hindering the well generalization of these location-adherent models. To shun this drawback, several pioneering studies open a promising research direction of location-agonistic crowd counting. Albeit the noticeable efforts, they somewhat ignore the merits of diverse learning paradigms and the issue of intractable density shift. To ameliorate these issues, in this paper, a novel Dynamic Mixture of Counter Network (DMCNet) is proposed for location-agnostic crowd counting. Specifically, our DMCNet inherits the hybrid advantages of CNNs (e.g. locality-oriented and pyramidal property) and MLP-based structure (e.g. global receptive fields and light weight). Particularly, the dynamic counter predictor and the mixture of counter heads are delicately designed to hammer at combating huge density shift and overfitting. Extensive experiments demonstrate that our DMCNet attains state-of-the-art performance against existing location-agnostic approaches and performs on par with many conventional location-adherent ones.

1. Introduction

During the past few years, counting problems (*e.g.* crowd [51], cells [17], fruits [42] and generalized object [43] counting) have drawn ever-increasing attention from the research community in the realm of computer vision, thanks to their far-ranging impacts on a train of societal applications, such as social distance monitoring [41], metropolis management [36], traffic controlling [64] and agriculture industry intelligentization [34], *etc*. The outbreak of COVID-19 pandemic has further stimulated the resurging of the crowd counting field which deserves to be dug deep into. Crowd counting task hammers at deriving single and unconstrained count values from the input still crowd scenes [24] or spatio-temporal video signals [35]. Inchoate approaches for crowd counting attach more priority to detect the body parts of crowd individuals dispersing across the whole image through heuristically-designed fea-

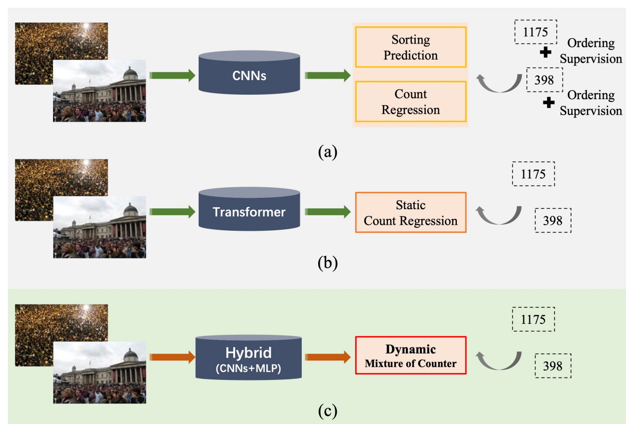

Figure 1. The categories of location-agnostic counting frameworks under count-level supervisions. (a) Learn from locality-oriented CNNs paradigm endowed with auxiliary sorting guidance [20, 62]. (b) Learn characteristics with global receptive fields via transformer-based structure [27]. (c) Our method inherits both merits of CNNs and MLP-based paradigms in a hybrid manner, and proposes a dynamic regression protocol.

ture engineering [66, 45, 23]. Albeit awesome accuracy improvements achieved, they are incompetent to produce satisfactory results when encountering highly-congested images with severe occlusions, large scale changes and density shifts, thereby hampering their generalization to wider scenarios. To surmount the barriers of those detection-based approaches, Lempitsky *et al*. [21] initiate the supervisory signal of density maps, and cast crowd counting problem into a new trend of density map regression.

More recently, the superb representational ability of Convolutional Neural Networks (CNNs) [46, 11] has ushered crowd counting in a booming era via a sequence of prevailing CNN-based models [65, 1, 4, 37, 55, 61, 25, 29]. The mainstream of existing methods take location-wise dot or density maps as the central supervisory signals, and therefore requires a large amount of location-level annotations. A series of crowd datasets in vogue (*e.g*. ShanghaiTech [65], UCF_QNRF [12], JHU-CROWD++ [50]) are perseveringly produced by manually marking dots around centroids of all peoples heads appearing across congested scenes, which is extremely time-consuming and labour-intensive. For example, 1.51 million dots were manually annotated for JHU-CROWD++ [50] whereas 1.25 million heads were labelled for UCF_QNRF [12].

Considering the arduous procedure of collecting samples with strong spatial hints, efforts to ease the dependency on location-wise annotations are well worth the trouble. In specific, L2R [32] and Sindagi *et al*. [49] endeavour to absorb a mass of unlabelled data from Internet through designing side sorting task and generating pseudo labels, respectively. AL-AC [67] strives to limit the use rate of labelled images, and attains the competitive results only using 10% annotated samples. To reduce the usage of ground-truth spatial regions, Xu *et al*. [61] excel in learning informative features from stochastically-predefined partial regions with smaller areas. In spite of great efforts, the requirement of gathering burdensome location-wise annotations still cannot be circumvented and these models fail to cope with the case where supervisions of pure crowd counts are solely available, as crowd counts can be easily inferred from ground-truth density maps but not *vice versa* [20].

To evade the burden of location annotations and narrow the gap between training and inference domains, the idea of location-agnostic crowd counting emerged [20, 62, 27]. In practice, large-scale datasets taking only single counts as annotations can be easily acquired in many target scenarios. For instance, once a ground-truth crowd count is collected and fixed for a venue with controlled access, *e.g*. bus station, the annotations for subsequent images can be quickly inferred by adding/subtracting numbers of objects entering/leaving [56]. Hence, weakly-supervised crowd counting has a vast prospect in expanding dataset scales, and enhancing the evolution of more generalized counting

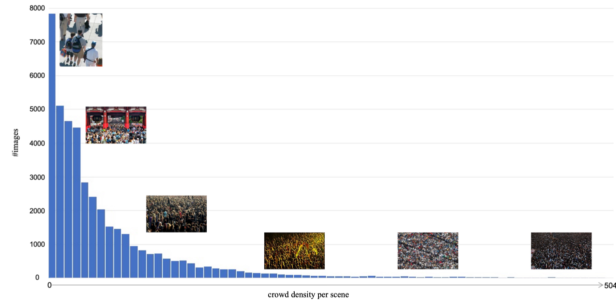

Figure 2. Long-tailed distribution of density values upon plenty of patches stochastically cropped from the dataset ShanghaiTech Part A [65]. Drastic and non-uniform density shifts existing in training set can be easily observed.

problem with multifarious objects. Location-free counting model aims at directly learning mapping functions from count-level supervisory signals, which is completely in accordance with the ultimate goal of crowd counting task.

Existing count-level approaches either resort to CNNs to capture feature vector [20, 62] (see Fig. 1 (a)), or devote to cutting-edge learning paradigms (*e.g*. Transformer [54, 7] (see Fig. 1 (b))) for explicitly capturing global receptive fields. Although the effectiveness of locality-oriented features and global receptive fields have been demonstrated by these individual approaches for weakly-supervised crowd counting, they somewhat neglect the collaborative impacts of meritorious learning paradigms (CNNs and Transformer). Recently, several eye-catching attempts have been made to delve into the hybrid combination to maximize the advantages of distinct paradigms in the field of image classification. For example, approaches [63, 6, 9] allow models to marry properties of CNNs and transformer, whereas Li *et al*. [22] try to seamlessly cascade the CNNs and MLP-based structure through a Hierarchical Convolutional MLPs.

Apart from the lack of complementarity between intrinsic merits from disparate learning paradigms, recent approaches ignore the issue of density shift (illustrated in Fig. 2) to some extent. Density map-based algorithms have extensively investigated the problem of density changes by presenting a pool of sophisticated techniques, such as multi-column [65, 1] and divide-and-conquer strategies [60]. On top of homogenous supervisory signal (*i.e*. count-level annotations), density shifts inevitably bring implicit ambiguity in the training procedure of location-agnostic models, resulting in unsatisfactory and overfitting-prone performance. Besides, the non-uniform density shift easily confuses the model on what distribution to learn. It is therefore more pregnant to suppress the negative influences of density change for weakly-supervised counting models than that for conventional ones, as strong location cues (density or dot maps) contribute to moderate this issue.

To ameliorate aforementioned challenges and further advance the blossom of location-agnostic counting protocols, a novel *Dynamic Mixture of Counter Network* (DMCNet) for location-agnostic crowd counting is presented in this paper, see Fig. 3. The proposed DMCNet features the seamless collaboration of locality-oriented CNNs and global MLP-based paradigms, dubbed as *Global Token Mixer* and *Pyramidal Feature Extractor* accordingly, and a dynamic counter condenser. Wherein, crude features are characterized at the first place by a pretrained VGG-16 [46] on ImageNet [44] before entering high-level transformations. The transfer of pretrained prior hammers at avoiding the collapsing of model trained from scratch. Then, MLP-based global token mixer is designed to proceed to extract multi-scale feature tokens with global receptive fields, while pyramidal feature extractor progressively enlarges receptive fields with the goals of hunting for spatial cues and steering the model towards learning dynamic weights. To better resist the huge density shift, we excavate a new dynamic scheme to dynamically choose the capability-sufficient and density-aware regression head instead of fixed counter head in existing work. Since the translation from the soft outputs of counter predictor to the discrete selection operation will hinder the back-propagation of gradient flow [68], here a principled reparameterization method Gumbel-Softmax [14] is delicately adopted to preserve end-to-end training. In short, the main contributions of this work are fourfold:

- A novel Dynamic Mixture of Counter Network (DMCNet) for location-agnostic crowd counting is proposed for boosting weakly-supervised crowd counting with count-level supervisory signals.

- Seamless collaboration between global token mixer and pyramidal feature extractor is dug into with the goal of sharing intrinsic merits of hybrid learning paradigms and enrich feature steering space.

- To combat the density shift and overfitting, a gumbel-softmax-based dynamic strategy is put forward towards dynamically and adaptively choosing the appropriate regression head for attaining an ensemble from a mixture of counter experts [13].

- Extensive experiments and ablation studies on prevailing benchmark datasets (*e.g*. ShanghaiTech Part A, Part B, UCF_QNRF and JHU-CROWD++) demonstrate the superiority of our proposed DMCNet over the state of the arts.

2. Related Work

Location-adherent Crowd Counting. During the recent few years, the supervisory signals of density or dot maps have been dominating the realm of crowd counting. The attention-getting challenges mainly include drastic scale variation, huge density shift and cluttered backgrounds. Multi-branch/column structures are explored in MCNN [65], Hydra-CNN [40], Switch-CNN [1], SANet [4], DSSINet [30], ASNet [15], and SASNet [52] to broaden the range of feature scales and cater for large scale variations. CSRNet [24], ADCrowdNet [31] and Adaptive Dilated Network [2] trigger a new line of explorations on expanding receptive fields via dilated and deformable convolutions. MBTTBF [48] devise a principled way of deriving pseudo scale supervision from density map for strengthening the scale awareness of features. Sindagi *et al*. [47] try to simultaneously predict global and local density levels, whereas DensityCNN [16] introduces an auxiliary density classifier for predicting global density. To filter out the background noises, Miao *et al*. [39] propose to reduce the false positive predictions by equipping attention mechanism in shallow layers, whereas Liu *et al*. [31] train an independent front-end network to estimate foreground crowd region maps imposed on original inputs.

Location-agnostic Crowd Counting. To get rid of intractable pixel-wise annotations, several location-agnostic counting pioneers lay the foundation of weakly-supervised crowd counting. Yang *et al*. [62] propose a soft-label sorting sub-network working with the counting backbone to explicitly mine the density-sensitivity ability. Although it tries to learn from sorting rather than location cues, the model is built upon CNNs and deliveries extremely limited receptive fields, thereby leading to very unsatisfactory and error-prone prediction. In addition, the soft target of auxiliary order matrix is heuristically-defined, which contributes little to the holistic performance. To promote the accuracy of count-level regressors, MATT [20] feeds few location-level annotations together with numerous count-level samples into the CNNs-based backbone at the same time. However, the usage of density maps still not be dispensed with. More recently, thanks to the widespread application of transformer in computer vision, Liang *et al*. [27] steer the approach to abstract features with global receptive fields through leaning upon transformer modules. Albeit intriguing improvements, it overlooks the locality-oriented representations from CNNs units, and introduces cumbrous and data-consuming self-attention condensers, especially in the case where spatial hints are removed and training set is insufficient. Besides, all above prior approaches are ill-considered in terms of ambiguity caused by drastic density shift. Hence, there is still large room for optimizing feature extractions and adapting them to location-agnostic models.

Dynamic Reparameterization Shemes. VAE [18] proposes to reparameterize internal random variable by decompose it into random (normal distribution) and certainty factors. Kusner *et al*. [19] utilize gumbel softmax to fit continuous distribution for GANs generating sequences of discrete elements. FBNet [59] deals with non-differentiable issue introduced by sampling operation via a gumbel-based differentiable neural architecture search. DRNet [68] determine the input resolution dynamically based on each input sample, resulting in a better trade-off between classification accuracy and computational overheads. Inspired by the effectiveness of these attempts in other application scenarios, we exploit dynamic reparameterization technique for crowd counting to approximate continues density distribution, thereby lowering the risks of overfitting and the sensitivity to unpredictable density shifts.

3. Dynamic Mixture of Counter Network

In this section, we elaborate the proposed Dynamic Mixture of Counter Network (DMCNet) for weakly-supervised crowd counting. Fig. 3 depicts the overall schema of our DMCNet. Following the common practice [24], the first ten layers (involving three max pooling layers) of a VGG-16 pretrained on ImageNet are incorporated as the frontend, with the purpose of preventing the model from seriously degenerating. After passing the raw crowd scene I through the frontend, a set of crude low-level features, denoted as F_l, are extracted and then fed into the subsequent high-level transformations consisting of global token mixer, pyramidal feature extractor, and dynamic counter predictor.

3.1. Global Token Mixer

Recently, the great potentials of global receptive fields have been exhibited by excavating cutting-edge learning paradigms, particularly transformer [7] and multi-layer perceptron (MLP [53, 26])-based methods, and stimulate a promising research direction in computer vision. TransCrowd [27] is the pioneering work that utilizes transformer to dig up clues with global receptive fields, and achieves impressive improvements for location-agnostic crowd counting. Nevertheless, self-attention condenser is data-consuming and makes the model prone to overfitting due to the insufficiency of training crowd samples, which is in line with the observations in several existing works [3, 5, 58]. Taking these drawbacks into consideration, here we choose the MLP-based paradigm to tokenize and optimize features for directly regressing total counts. Motivated by the fascinating performance and efficiency of MLPMixer [53] in image classification, three-level MLP transformations are designed to form the global token mixer module. As demonstrated in Fig. 4, the low-level features are split into a sequence of feature patches at three different granularities/resolutions, which is followed by linear projection operation and pivotal MLP transformations for global modelling.

In specific, individual MLP unit is comprised of two types of MLP layers, a token MLP and a shared channel MLP, which aim at mixing information along dimensions of spatial and channel. The zoomed-in view in Fig. 4 shows the details. Moreover, inspired by shake-shake regularization technique [10], a principled aggregation strategy is presented through summing multi-scale tokens to facilitate the communication across tokens at multiple scales. During the training phase, stochastic affine combination of levels are performed to avoid overfitting rather than directly summing. Apart from feature refinement, the multi-scale summation operations implicitly introduce residual learning [11] into the model learning simultaneously, which is conducive to expedite model's convergence.

3.2. Pyramidal Feature Extractor

Although global token mixer is capable of extracting multi-granularity tokens with global receptive fields, the pyramidal structure of features are completely discarded. The natural property of feature pyramid delivered by CNNs has been proven to be beneficial for enhancing capability of architectures [63, 6, 9, 22]. To preserve the hierarchical attribute inherited from CNNs-based frontend and mine global spatial cues, a stem of CNN-based pyramidal feature extractor is devised to steer the model's learning. Analogous to the frontend, *de facto* standard building units of 3×3 convolutions and max pooling layers are setup to gradually expand the receptive fields of spatial feature maps. The pyramidal feature extractor consists of a sequence of layers "conv$\rightarrow$pooling$\rightarrow$conv$\rightarrow$pooling$\rightarrow$conv$\rightarrow$pooling$\rightarrow$conv" followed by a global average pooling to generate the feature vector with high-level semantics. Batch normalization and ReLU function are leveraged to reduce internal covariate shift and add non-linearity of feature space. The introduction of pyramidal feature extraction succeeds in inheriting meritorious hierarchical features working with global-oriented tokens from global token mixer module.

3.3. Dynamic Mixture of Counter Experts

On top of high-level tokens and feature vector with pyramidal semantics provided by both MLP- and CNNs-based structures, the DMCNet proceeds to learn dynamic counter weights and regress final counts through Dynamic Counter Predictor and the Mixture of Counter Experts. Inspired by the effectiveness of dynamic resolution [68] and mixture of experts (MoE) [13, 8], a dynamic mixture of counter is proposed at the end of the network to dynamically and adaptively determine the suitable density-specific regression counter. As for the mixture of counters, it includes a group of MLP-based heads with varied model scales. Specifically, each regression head consists of operations (*e.g.* fully-connected layers, 1D batch normalization, 2D dropout and ReLU activate functions), and its width is heuristically-defined. To better combat density shift, a principled way is to assign small-scale heads to samples with lower density values (*e.g.* 5 people number), whereas the

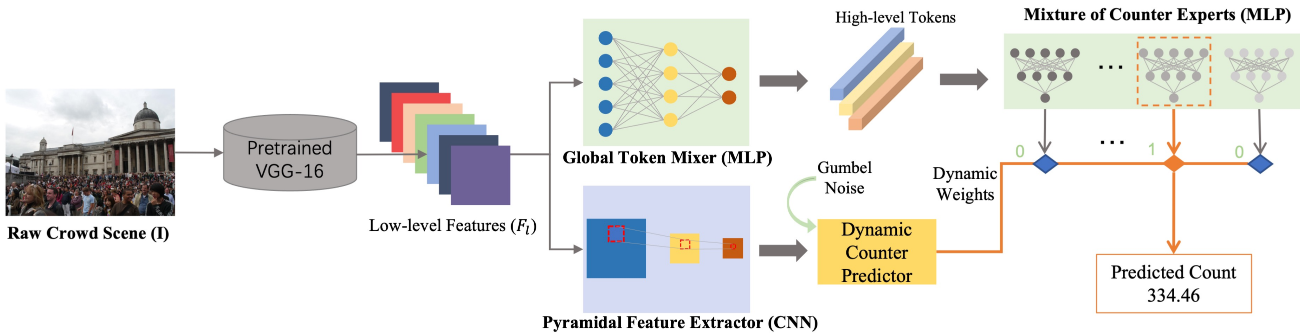

Figure 3. The overall schema for DMCNet architecture, which consists of a pretrained VGG-16 frontent, global token mixer, pyramidal feature extractor and dynamic mixture of counter experts. Crude low-level features are fed into two meritorious learning paradigms (MLP and CNN)-based modules for characterizing tokens embracing global receptive fields and features with pyramidal property, respectively. Finally, a dynamic scheme is delicately devised to dynamically and automatically determine the usage status of pre-designed mixture of counter experts and produce the predicted count values.

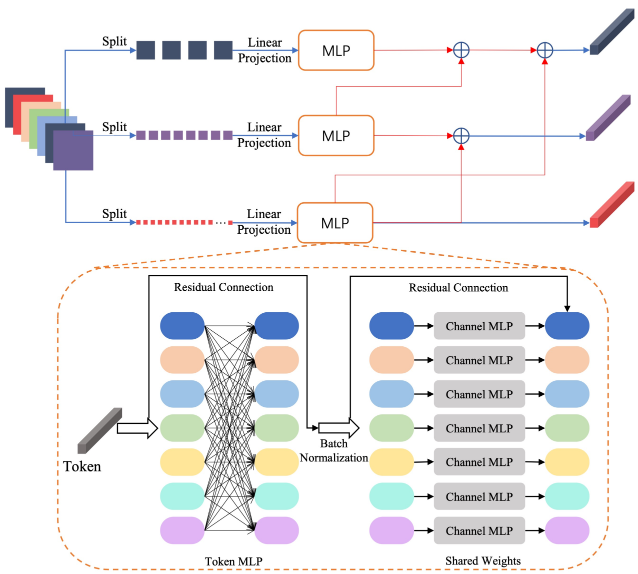

Figure 4. The details of the proposed global token mixer. To capture multi-scale tokens, three-granularity strategy is designed to split the low-level features into three sequences of non-overlapped patches at different resolutions. The cross-level tokens are integrated via point-wise summations.

crowd scenes with larger densities (*e.g.* 500 count) deserve to be tied to counters with higher complexity. An intuitive and natural solution for automatically selecting the corresponding counters is to train an attention condenser and softly recalibrate the outputs of all counter heads. Albeit feasibility, this scheme is of limited benefit as all counters are jointly trained, which is inclined to disturb each other's learning procedure and introduce ambiguity. Therefore, following the previous attempts [18, 13], we try to force the intermediate token to represent underlying distribution instead of specific features by dynamically resampling or reparameterizing rear structure of regression counters in a one-hot manner.

Hard resampling inevitably cuts the backpropagation flow and hurts the smooth process of end-to-end training. To address this issue, a gumbel softmax is adopted here to predict dynamic one-hot encodings and make the discrete selection differentiable during backpropagation phase. In our method, the dynamic counter predictor calculates a set of probabilities for the mixture of counters, denoted as $P_c = [p_{c1}, p_{c2}, ..., p_{cn}]$, where n is the total number of counter heads in the mixture. Given the probabilities, the discrete counter decisions (one-hot vector D) can be dynamically computed as:

$$D = onehot(argmax(log(p_{ci}) + G_i)),\ i = 1,\ 2,\ ...,\ n, \tag{1}$$

where $onehot$ means the function generating one-hot masks and G_i indicates the gumbel noise drawn from independent identically distribution U for each crowd scene:

$$G_i = -log(-log(x)), x \in U(0, 1). \tag{2}$$

During back propagation, considering the non-differentiability of argmax operation, the inference of the one-hot sampling can be approximated by following continuous and differentiable gumbel softmax:

$$\hat{D}_i = \frac{exp(log(p_i) + G_i)/r}{\sum_{j=1}^{n}(exp(log(p_j) + G_j)/r)},\ i = 1,\ 2,\ ...,\ n, \tag{3}$$

where r is temperature hyperparameter and set as 1 in our experiments. Through this straight-through trick of gumbel softmax, the gradient flows from discrete hard argmax are adjusted to be continuous and fluent, whereas the highest density entry of the original density distribution is not affected [68]. The probabilities for candidate counter heads are computed by our dynamic counter predictor including three linear layers with ReLU function and 2D dropout. Given the outputs of the mixture of counter experts $C = [C_1, C_2, ..., C_n]$, the final count prediction N_p of the DMC-Net is formulated as:

$$N_p = C \cdot D, \tag{4}$$

where $\cdot$ denotes inner product between candidate output vector and the one-hot selection weights.

3.4. Objective Function

Given the location-agnostic labels N_{gt} (*i.e.* only ground-truth counts), the primary supervisory signal L_reg is derived from the $L1$ distance between N_{gt} and N_p as follows:

$$L_{reg} = \sum_{i=1}^{B} |N_p - N_{gt}|, \tag{5}$$

where B is the batch size and this loss term is crucial for optimizing the model to predict accurate crowd counts. Besides, to steer the better learning of the dynamic counter predictor, a deeply-supervised loss term L_{cls} is designed and imposed on the intermediate logits P_c. L_{cls} is calculated via cross entropy between P_c and y_c as follows:

$$L_{cls} = -\sum_{i=1}^{B} \sum_{c=1}^{M} y_{i,c} log(p_{i,c}), \tag{6}$$

where B is the batch size and M means the total class number. Wherein, y_c is obtained through the function $m = Floor(N_{gt} \div T)$ and $(y_{i,m} = 1, y_{i,else} = 0)$, where T represents the heuristically-defined thresholds depending on the maximum values of crowd counts. By adopting this density classification constraint, the proposed dynamic counter predictor can be driven to provide density-aware features and infer more accurate dynamic selection weights for further easing ambiguity caused by severe density shifts. The overall objective function for optimizing our DMCNet is therefore formulated as:

$$L = L_{reg} + L_{cls}. \tag{7}$$

As two types of ground truths are homogenous, two optimization directions from L_{reg} and L_{cls} are in parallel with the learning target of the holistic model. Hence, heuristically-predefined hyperparameter is not introduced to balance the impacts of two loss terms, thereby mitigating the extra burden of manual fine-tuning.

4. Experiments

4.1. Implementation Details

Datasets and Evaluation Metrics. The ShanghaiTech benchmark [65] is formed by two parts of evaluation datasets: Part A and Part B. This dataset consists of 1,198 crowd scenes with a total number of 330,165 labelled people. Wherein, Part A contains 482 (300 for training and 182 for testing) congested images while Part B includes 716 images (400 for training and 316 for testing). More difficult UCF_QNRF [12] dataset is collected from the website and is comprised of 1,553 images with a total number of 1,252,642 people, in which 1201 images are taken for training the model and 334 samples for inference. To better verify the superiority of our DMCNet, a large-scale dataset JHU-CROWD++ [50] is also considered, which includes 4,372 images with a division of 2,722 images for training, 1,600 samples for inference, and 500 ones for validation. This dataset has the issue of huger density shift ranging from 0 to 25,791. Even though NWPU [57] is another alternative large-scale source, its ground truths are not released for testing. Therefore, we choose the JHU-CROWD++ as the representative large-scale dataset. For evaluation metrics, we choose Mean Absolute Error (MAE) to indicate the counting accuracy and Mean Square Error (MSE) to reflect the volatility of predicted results.

Implementation. To suppress the computational overheads caused by over-large resolutions, all raw samples are resized to 1024×768 or under. During the training phase, a batch of patches at the resolution of 256×256 are stochastically cropped online from the resized images. The ground truths only involve the single and unconstrained values of crowd counts related to patches in a batch. Random horizontal/vertical flip, random rotation and lighting are utilized to form the data augmentation. In our experiments, batch size is set to 24 for Part A and Part B, 36 for UCF-QNRF and JHU-CROWD++. The initial learning rate is 1e-5. Adam optimizer with momentum of 0.95 and weight decay of 5e-4 is leveraged to train our model. We allocate 45 counter heads in the mixture for all datasets. At the training stage, the gumbel-based hard mask is used to dynamically determine the candidate counter head, while soft weights from softmax function are calculated to dynamically and adaptively integrate the outputs of the mixture of counters during the inference. In Equ. 6, the threshold T is set to be 20 for Part A and Part B, 70 for UCF_QNRF and 220 for JHU-CROWD++ due to diverse ranges of density values.

4.2. Comparison with State-of-the-art

We compare our proposed DMCNet with state-of-the-art approaches under different supervisory signals of location-level and count-level annotations, as demonstrated in Table 1. Our method consistently and significantly outperforms the best location-agnostic TransCrowd by 11.55% MAE on Part A, 7.09% on Part B, 0.69% on UCF_QNRF and 7.02% on JHU-CROWD++, accordingly, and achieves the state-of-the-art performance on four benchmark datasets. The best results are shown

Methods	Location Agnostic	**Part A**		**Part B**		**UCF-QNRF**		**JHU-CROWD++**	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
ADCrowdNet [31]	×	63.2	98.9	7.6	13.9	-	-	-	-
MBTTBF [48]	×	60.2	94.1	8.0	15.5	97.5	165.2	81.8	**299.1**
DSSINet [30]	×	60.63	96.04	6.85	10.34	99.1	159.2	133.5	416.5
S-DCNet [60]	×	58.3	95.0	6.7	10.7	104.4	176.1	-	-
ASNet [15]	×	57.78	90.13	-	-	91.59	159.71	-	-
AMRNet [33]	×	61.59	98.36	7.02	11.00	86.6	152.2	-	-
S3 [28]	×	57.0	96.0	6.3	10.6	**80.6**	**139.8**	-	-
DM-Count [55]	×	59.7	95.7	7.4	11.8	85.6	148.3	-	-
BL [37]	×	62.8	101.8	7.7	12.7	88.7	154.8	**75.0**	299.9
UOT [38]	×	58.1	95.9	6.5	10.2	83.3	142.3	-	-
P2PNet [51]	×	**52.74**	**85.06**	**6.25**	**9.9**	85.32	154.5	-	-
MATT [20]	✓	80.1	129.4	11.7	17.5	-	-	-	-
Sorting [62]	✓	104.6	145.2	12.3	21.2	-	-	-	-
TransCrowd-T [27]	✓	69.0	116.5	10.6	19.7	98.9	176.1	76.4	319.8
TransCrowd-G [27]	✓	66.1	105.1	9.3	16.1	97.2	168.5	74.9	295.6
Our DMCNet	✓	**58.46**	**84.55**	**8.64**	**13.67**	**96.52**	**163.99**	**69.64**	**246.93**

Table 1. Experimental comparisons against existing state of the arts under two types of annotation configurations on four prevailing datasets. Best results for location-adherent and -agnostic are shown in boldface. Our approach consistently outperforms current location-agnostic methods and attains state-of-the-art accuracy as well as lowest volatility. It also performs on par with many conventional location-adherent approaches.

in boldface and demonstrate the superiority of the proposed model. Even though the labels adopted by our method are extremely weak and simple (*i.e*. only total crowd counts), DMCNet attains competitive accuracy against fully-supervised counterparts. For example, on Part A, our model produces the 58.46 MAE, which is on par with the location-demanding S-DCNet/UOT and outperforms other models in vogue, *e.g*. BL (62.8 MAE), AMRNet (61.59 MAE) and DM-count (59.7 MAE). More interestingly, our model achieves the best MSE value of 84.55 on Part A, which shows that location-agnostic DMCNet delivers great stability of predictions due to less ambiguity on point locations. On more large-scale and arduous datasets UCF_QNRF and JHU-CROWD++ with huger density shifts, our DMCNet still performs well. For dataset UCF_QNRF, our model obtains the best MAE, MSE of 96.52, 163.99 and even outperforms up-to-date conventional MBTTBF (97.5 MAE), DSSINet (99.1MAE) and S-DCNet (104.4 MAE). On more large-scale dataset JHU-CROWD++, our DMCNet outperforms all location-adherent and location-agnostic counting approaches for comparison, which illustrates the consistent superiority of the proposed method on simple or more complex datasets.

4.3. Ablation Study

To verify the impacts of the individual modules in our DMCNet (*e.g*. mixture of counter heads, auxiliary classification loss, pyramidal feature extractor and gumbel softmax), a series of experiments is conducted here to ablate these components. All ablation studies are carried out based on the dataset ShanghaiTech Part A.

Models	MAE	MSE	MAE Gains
Baseline	63.31	95.67	-
+ Mixture (w.o L_{cls})	60.86	86.98	2.45
+ L_{cls}	60.11	87.17	0.75
+ CNNs (w.o Gumbel)	59.86	88.62	0.24
+ Gumbel Noise	58.46	84.55	1.40

Table 2. Ablation study on different components. Baseline is constructed only using MLP-based global token mixer and a fixed regression head. Then a set of proposed individuals are plugged progressively to enrich the model until the final DMCNet. The MAE gains demonstrate the effectiveness of four internal elements.

Importance of different components. We first design a pool of experiments to incorporate each proposed component step by step and report the corresponding MAE, see Table 2. Wherein, the baseline represents the plain model without any bells and whistles, which is built by cascading the frontend, global token mixer and a fixed regression head. The first counter in the mixture of counters is selected as the fixed counter head in baseline. On top of the baseline, we involve the mixture of counters, auxiliary classification supervision L_{cls}, CNNs-based pyramidal feature extractor and gumbel softmax, respectively. The results in Table 2 demonstrates that the introduction of these fundamental mechanisms empowers the model and brings positive impacts on the accuracy of DMCNet. Particularly, the MAE

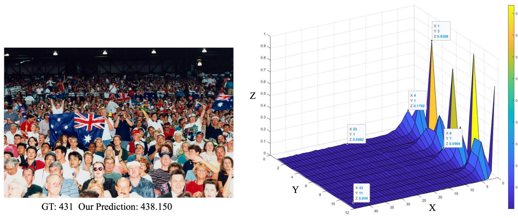

Figure 5. The visualization of learned dynamic weights (*right*) for an example (*left*) in ShanghaiTech Part A, which shows that our dynamic predictor indeed allows the model to generate diverse dynamic decision status for adapting to patches with density shifts. Here, X axis denotes the number of mixture of counters and Y axis is the sliding window index over input test scene, whereas Z axis represents the probability value.

reduction provided by the mixture of counter heads and gumbel softmax-based reparameterization are more prominent, *e.g*. 2.45 and 1.40 accordingly.

Models	MAE	MSE	Degradation
Holistic DMCNet	58.46	84.55	-
Gumbel Hard	58.75	83.99	0.49%
AVG Aggregation	70.18	95.68	20.04%
0^{th} Head	178.79	295.69	205.84%
20^{th} Head	67.35	95.87	15.20%
45^{th} Head	70.64	99.28	20.83%

Table 3. Ablation study and performance degradation under different testing schemes on ShanghaiTech Part A.

The impacts of dynamic mechanism. The dynamic selection mechanism in DMCNet includes the mixture of counters and the dynamic counter predictor. Their enhancements on accuracies have been investigated in Table 2. To better give insights into the behaviour of the dynamic mechanism, we investigate DMCNet under different configurations and examine the performance degradations compared with the results reported on Part A. Gumbel Hard means that the one-hot dynamic weights (completely same as the operation used in training phase) are generated during inference time. As the stochastic gumbel noises are introduced, we run ten times and report the best result. AVG Aggregation aims to execute recalibration by averaging outputs of mixture of counters, whereas i^{th} Head ($i = 0, 20, 45$) is implemented by only retaining i^{th} regression head with removal of all other counters. It can be observed from Table 3 that dynamically determining the usage of counter set contributes model's performance. Choosing distinct heads incurs performance degradations with considerable difference (from 67.35 to 178.79), which illustrates that heads with different complexities play diverse roles in our holistic DMCNet. To further visualize the discrepancy among learned dynamic weights over varying patches, the weight distributions provided by our dynamic counter predictor on a Part A example are depicted in Fig. 5. For different patches from the same crowd scene, the dynamic weights for the mixture of counters are adaptive. The confidence probabilities of front counter heads focusing on sparse densities are relatively higher than rare counters with huge densities. The small values of X axis correspond to larger variations (long-tailed distributions), which may be caused by the implicit imbalance in training data and massive samples with sparse densities bring less ambiguity (higher confidence).

The effects of auxiliary loss L_{cls}. To validate the effectiveness of the proposed auxiliary loss term L_{cls}, we conduct experiments to ablate this term in Table 4, which empirically demonstrates that the guidance of density classification is beneficial for combatting density shits and preventing the model from degenerating. When removing the supervision of L_{cls}, the performance of our DMCNet is heavily influenced by the number of heads in the mixture. As shown in Table 4, more heads are setup, worse accuracies are obtained by models without L_{cls}. This phenomenon may be caused by the fact that it is challenging for the model to adaptively assign many heads to corresponding density levels in a spontaneous manner. The counterpart with larger number of heads is prone to overfitting due to limited training samples.

Models	MAE	MSE
w.i. L_{cls}	58.46	84.55
w.o. L_{cls} and 45 Heads	71.61	100.31
w.o. L_{cls} and 25 Heads	63.96	95.71
w.o. L_{cls} and 15 Heads	60.86	86.98

Table 4. The impact of removing the auxiliary classification loss term under varied numbers of counters on ShanghaiTech Part A.

5. Conclusion

In this paper, we propose a novel Dynamic Mixture of Counter Network (DMCNet) for location-agnostic crowd counting to further enhance weakly-supervised counting protocols. Our DMCNet adopt the hybrid combination of pyramidal CNNs and MLP-based structure to inherit both meritorious learning paradigms. Wherein, multi-level MLP global token mixer hammers at capturing global receptive fields without resorting to cumbrous and data-consuming transformers, whereas the pyramidal feature module aims to preserve the property of hierarchical features. Besides, to ameliorate the issue of density shift, we propose a dynamic counter predictor and the mixture of counter with the goal of dynamically and automatically choosing appropriate fusion status of regression heads focusing on different density levels. Extensive experiments on several prevailing benchmark datasets demonstrate the superiority of our DMCNet.

References

[1] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5744–5752, 2017.

[2] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2020.

[3] Ankur Bapna, Mia Xu Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. Training deeper neural machine translation models with transparent attention. *EMNLP*, 2018.

[4] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[5] Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. Autoformer: Searching transformers for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12270–12280, 2021.

[6] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Zhipeng Du, Miaojing Shi, Jiankang Deng, and Stefanos Zafeiriou. Redesigning multi-scale neural network for crowd counting. *arXiv preprint arXiv:2208.02894*, 2022.

[9] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[10] Xavier Gastaldi. Shake-shake regularization. *International Conference on Learning Representations*, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the european conference on computer vision (ECCV)*, pages 532–546, 2018.

[13] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[14] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[15] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4706–4715, 2020.

[16] Xiaoheng Jiang, Li Zhang, Tianzhu Zhang, Pei Lv, Bing Zhou, Yanwei Pang, Mingliang Xu, and Changsheng Xu. Density-aware multi-task learning for crowd counting. *IEEE Transactions on Multimedia*, 2020.

[17] Aisha Khan, Stephen Gould, and Mathieu Salzmann. Deep convolutional neural networks for human embryonic cell counting. In *European conference on computer vision*, pages 339–348. Springer, 2016.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Matt J Kusner and José Miguel Hernández-Lobato. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*, 2016.

[20] Yinjie Lei, Yan Liu, et al. Towards using count-level weak supervision for crowd counting. *Pattern Recognition*, 2021.

[21] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in neural information processing systems*, 23, 2010.

[22] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. Convmlp: Hierarchical convolutional mlps for vision. *arXiv preprint arXiv:2109.04454*, 2021.

[23] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th international conference on pattern recognition*, pages 1–4. IEEE, 2008.

[24] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[25] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao. Locating and counting heads in crowds with a depth prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[26] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. As-mlp: An axial shifted mlp architecture for vision. In *International Conference on Learning Representations*, 2021.

[27] Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: Weakly-supervised crowd counting with transformer. *arXiv preprint arXiv:2104.09116*, 2021.

[28] Hui Lin, Xiaopeng Hong, Zhiheng Ma, Xing Wei, Yunfeng Qiu, Yaowei Wang, and Yihong Gong. Direct measure matching for crowd counting. *arXiv preprint arXiv:2107.01558*, 2021.

[29] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19628–19637, 2022.

[30] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the*

IEEE/CVF international conference on computer vision, pages 1774–1783, 2019.

[31] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3225–3234, 2019.

[32] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7661–7669, 2018.

[33] Xiyang Liu, Jie Yang, Wenrui Ding, Tieqiang Wang, Zhijin Wang, and Junjun Xiong. Adaptive mixture regression network with local counting map for crowd counting. In *European Conference on Computer Vision*, pages 241–257, 2020.

[34] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. Tasselnet: counting maize tassels in the wild via local counts regression network. *Plant methods*, 13(1):1–17, 2017.

[35] Yu-Jen Ma, Hong-Han Shuai, and Wen-Huang Cheng. Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. *IEEE Transactions on Multimedia*, 2021.

[36] Zhiheng Ma, Xiaopeng Hong, Xing Wei, Yunfeng Qiu, and Yihong Gong. Towards a universal model for cross-dataset crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3214, 2021.

[37] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6142–6151, 2019.

[38] Zhiheng Ma, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong. Learning to count via unbalanced optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2319–2327, 2021.

[39] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. Shallow feature based dense attention network for crowd counting. In *AAAI*, pages 11765–11772, 2020.

[40] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *European conference on computer vision*, pages 615–629. Springer, 2016.

[41] Narinder Singh Punn, Sanjay Kumar Sonbhadra, Sonali Agarwal, and Gaurav Rai. Monitoring covid-19 social distancing with person detection and tracking via finetuned yolo v3 and deepsort techniques. *arXiv preprint arXiv:2005.01385*, 2020.

[42] Maryam Rahnemoonfar and Clay Sheppard. Deep count: fruit counting based on deep simulated learning. *Sensors*, 17(4):905, 2017.

[43] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021.

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[45] Oliver Sidla, Yuriy Lypetskyy, Norbert Brandle, and Stefan Seer. Pedestrian detection and tracking for counting applications in crowded situations. In *2006 IEEE International Conference on Video and Signal Based Surveillance*, pages 70–70. IEEE, 2006.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1861–1870, 2017.

[48] Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *ICCV*, pages 1002–1012, 2019.

[49] Vishwanath A Sindagi, Rajeev Yasarla, Deepak Sam Babu, R Venkatesh Babu, and Vishal M Patel. Learning to count in the crowd from limited labeled data. In *European Conference on Computer Vision*, pages 212–229. Springer, 2020.

[50] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*, 2020.

[51] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021.

[52] Qingyu Song, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. To choose or to fuse? scale selection for crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2576–2583, 2021.

[53] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34, 2021.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[55] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33:1595–1607, 2020.

[56] Mingjie Wang, Jun Zhou, Hao Cai, and Minglun Gong. Crowdmlp: Weakly-supervised crowd counting via multi-granularity mlp. *arXiv preprint arXiv:2203.08219*, 2022.

[57] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020.

[58] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *ACL*, 2019.

[59] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019.

[60] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8362–8371, 2019.

[61] Yanyu Xu, Ziming Zhong, Dongze Lian, Jing Li, Zhengxin Li, Xinxing Xu, and Shenghua Gao. Crowd counting with partial annotations in an image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15570–15579, 2021.

[62] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.

[63] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021.

[64] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 3667–3676, 2017.

[65] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016.

[66] Tao Zhao and Ramakant Nevatia. Bayesian human segmentation in crowded situations. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–459. IEEE, 2003.

[67] Zhen Zhao, Miaojing Shi, Xiaoxiao Zhao, and Li Li. Active crowd counting with limited supervision. In *European Conference on Computer Vision*, pages 565–581. Springer, 2020.

[68] Mingjian Zhu, Kai Han, Enhua Wu, Qiulin Zhang, Ying Nie, Zhenzhong Lan, and Yunhe Wang. Dynamic resolution network. *Advances in Neural Information Processing Systems*, 34:27319–27330, 2021.