Dynamic Mixture of Counter Network for Location-Agnostic Crowd Counting

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

Yong Dai  
Tencent AI Lab  
daiyongya@outlook.com

Hao Cai  
School of Computer Science  
Memorial University of Newfoundland  
hcl864@mun.ca

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 vi-sion, 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-
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 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 blossoming 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 cumbersome 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 Schemes. 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 dis-
crete 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 \( F_i \), 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 \( 3 \times 3 \) 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 “\( \text{conv}\rightarrow\text{pooling}\rightarrow\text{conv} \)” 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 frontend, 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:

\[
D = \text{onehot}(\arg\max_i (\log(p_{ci}) + G_i)), i = 1, 2, ..., n, \quad (1)
\]

where \( \text{onehot} \) means the function generating one-hot masks and \( G_i \) indicates the gumbel noise drawn from independent identically distribution \( U \) for each crowd scene:

\[
G_i = -\log(-\log(x)), x \in U(0,1). \quad (2)
\]

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:

\[
\hat{D}_i = \frac{\exp(\log(p_{ci}) + G_i)/r}{\sum_{j=1}^{n}\exp(\log(p_{cj}) + G_j)/r}, i = 1, 2, ..., n, \quad (3)
\]

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 \(C = [C_1, C_2, \ldots, C_n]\), the final count prediction \(N_p\) of the DMCNet is formulated as:

\[
N_p = C \cdot D,
\]

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:

\[
L_{reg} = \sum_{i=1}^{B} |N_p - N_{gt}|,
\]

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:

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

where \(B\) is the batch size and \(M\) means the total class number. Wherein, \(y_c\) is obtained through the function \(m = \text{Floor}(N_{gt} \div T)\) 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:

\[
L = L_{reg} + L_{cls}.
\]

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 \(1024 \times 768\) or under. During the training phase, a batch of patches at the resolution of \(256 \times 256\) 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
### 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.

<table>
<thead>
<tr>
<th>Models</th>
<th>MAE</th>
<th>MSE</th>
<th>MAE Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>63.31</td>
<td>95.67</td>
<td>-</td>
</tr>
<tr>
<td>+ Mixture (w.o $L_{cls}$)</td>
<td>60.86</td>
<td>86.98</td>
<td>2.45</td>
</tr>
<tr>
<td>+ $L_{cls}$</td>
<td>60.11</td>
<td>87.17</td>
<td>0.75</td>
</tr>
<tr>
<td>+ CNNs (w.o Gumbel)</td>
<td>59.86</td>
<td>88.62</td>
<td>0.24</td>
</tr>
<tr>
<td>+ Gumbel Noise</td>
<td>58.46</td>
<td>84.55</td>
<td>1.40</td>
</tr>
</tbody>
</table>

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

---

### 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Location Agnostic</th>
<th>Part A</th>
<th>Part B</th>
<th>UCF-QNRF</th>
<th>JHU-CROWD++</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>ADCrowdNet [31]</td>
<td>×</td>
<td>63.2</td>
<td>98.9</td>
<td>7.6</td>
<td>13.9</td>
</tr>
<tr>
<td>MBTTBF [48]</td>
<td>×</td>
<td>60.2</td>
<td>94.1</td>
<td>8.0</td>
<td>15.5</td>
</tr>
<tr>
<td>DSSINet [30]</td>
<td>×</td>
<td>60.63</td>
<td>96.04</td>
<td>6.85</td>
<td>10.34</td>
</tr>
<tr>
<td>S-DCNet [60]</td>
<td>×</td>
<td>58.3</td>
<td>95.0</td>
<td>6.7</td>
<td>10.7</td>
</tr>
<tr>
<td>ASNet [15]</td>
<td>×</td>
<td>57.78</td>
<td>90.13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AMRNet [33]</td>
<td>×</td>
<td>61.59</td>
<td>98.36</td>
<td>7.02</td>
<td>11.00</td>
</tr>
<tr>
<td>S3 [28]</td>
<td>×</td>
<td>57.0</td>
<td>96.0</td>
<td>6.3</td>
<td>10.6</td>
</tr>
<tr>
<td>DM-Count [55]</td>
<td>×</td>
<td>59.7</td>
<td>95.7</td>
<td>7.4</td>
<td>11.8</td>
</tr>
<tr>
<td>BL [37]</td>
<td>×</td>
<td>62.8</td>
<td>101.8</td>
<td>7.7</td>
<td>12.7</td>
</tr>
<tr>
<td>UOT [38]</td>
<td>×</td>
<td>58.1</td>
<td>95.9</td>
<td>6.5</td>
<td>10.2</td>
</tr>
<tr>
<td>P2PNet [51]</td>
<td>×</td>
<td>52.74</td>
<td>85.06</td>
<td>6.25</td>
<td>9.9</td>
</tr>
<tr>
<td>MATT [20]</td>
<td>✓</td>
<td>80.1</td>
<td>129.4</td>
<td>11.7</td>
<td>17.5</td>
</tr>
<tr>
<td>Sorting [62]</td>
<td>✓</td>
<td>104.6</td>
<td>145.2</td>
<td>12.3</td>
<td>21.2</td>
</tr>
<tr>
<td>TransCrowd-T [27]</td>
<td>✓</td>
<td>69.0</td>
<td>116.5</td>
<td>10.6</td>
<td>19.7</td>
</tr>
<tr>
<td>TransCrowd-G [27]</td>
<td>✓</td>
<td>66.1</td>
<td>105.1</td>
<td>9.3</td>
<td>16.1</td>
</tr>
<tr>
<td>Our DMCNet</td>
<td>✓</td>
<td>58.46</td>
<td>84.55</td>
<td>8.64</td>
<td>13.67</td>
</tr>
</tbody>
</table>

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.

### 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.

<table>
<thead>
<tr>
<th>Models</th>
<th>MAE</th>
<th>MSE</th>
<th>MAE Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>63.31</td>
<td>95.67</td>
<td>-</td>
</tr>
<tr>
<td>+ Mixture (w.o $L_{cls}$)</td>
<td>60.86</td>
<td>86.98</td>
<td>2.45</td>
</tr>
<tr>
<td>+ $L_{cls}$</td>
<td>60.11</td>
<td>87.17</td>
<td>0.75</td>
</tr>
<tr>
<td>+ CNNs (w.o Gumbel)</td>
<td>59.86</td>
<td>88.62</td>
<td>0.24</td>
</tr>
<tr>
<td>+ Gumbel Noise</td>
<td>58.46</td>
<td>84.55</td>
<td>1.40</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Models</th>
<th>MAE</th>
<th>MSE</th>
<th>Degradation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic DMCNet</td>
<td>58.46</td>
<td>84.55</td>
<td>-</td>
</tr>
<tr>
<td>Gumbel Hard</td>
<td>58.75</td>
<td>83.99</td>
<td>0.49%</td>
</tr>
<tr>
<td>AVG Aggregation</td>
<td>70.18</td>
<td>95.68</td>
<td>20.04%</td>
</tr>
<tr>
<td>0\textsuperscript{th} Head</td>
<td>178.79</td>
<td>295.69</td>
<td>205.84%</td>
</tr>
<tr>
<td>20\textsuperscript{th} Head</td>
<td>67.35</td>
<td>95.87</td>
<td>15.20%</td>
</tr>
<tr>
<td>45\textsuperscript{th} Head</td>
<td>70.64</td>
<td>99.28</td>
<td>20.83%</td>
</tr>
</tbody>
</table>

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 \( i \textsuperscript{th} \) Head (\( i = 0, 20, 45 \)) is implemented by only retaining \( i \textsuperscript{th} \) 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 combating density shifts 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.

<table>
<thead>
<tr>
<th>Models</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>w.l. ( L_{cls} )</td>
<td>58.46</td>
<td>84.55</td>
</tr>
<tr>
<td>w.o. ( L_{cls} ) and 45 Heads</td>
<td>71.61</td>
<td>100.31</td>
</tr>
<tr>
<td>w.o. ( L_{cls} ) and 25 Heads</td>
<td>63.96</td>
<td>95.71</td>
</tr>
<tr>
<td>w.o. ( L_{cls} ) and 15 Heads</td>
<td>60.86</td>
<td>86.98</td>
</tr>
</tbody>
</table>

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 cumbersome 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


[29] Lingbo Liu, Zhihui Qiu, Guanlin Bi, 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.


