Effective Uni-Modal to Multi-Modal Crowd Estimation based on Deep Neural Networks

Usman Sajid

Submitted to the graduate degree program in Department of Electrical Engineering and Computer Science and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Date defended: December 10, 2021

The Dissertation Committee for Usman Sajid certifies
that this is the approved version of the following dissertation :

Effective Uni-Modal to Multi-Modal Crowd Estimation based on Deep Neural Networks

Date approved: December 10, 2021

Abstract

Crowd estimation is a vital component of crowd analysis. It finds many applications in real-world scenarios, e.g. huge gatherings management like Hajj, sporting and musical events, or political rallies. Automated crowd counting facilitates better and effective management of such events and consequently prevents any undesired situation. This is a very challenging problem in practice since there exists a significant difference in the crowd number in and across different images, varying image resolution, large perspective, severe occlusions, and dense crowd-like cluttered background regions. Current approaches do not handle huge crowd diversity well and thus perform poorly in cases ranging from extreme low to high crowd-density, thus, yielding huge crowd underestimation or overestimation. Also, manual crowd counting proves to be infeasible due to very slow and inaccurate results. To address these major crowd counting issues and challenges, we investigate two different types of input data: uni-modal (image) and multi-modal (image and audio).

In the uni-modal setting, we propose and analyze four novel end-to-end crowd counting networks, ranging from multi-scale fusion-based models to uni-scale one-pass and two-pass multi-task networks. The multi-scale networks employ the attention mechanism to enhance the model efficacy. On the other hand, the uni-scale models are well-equipped with novel and simple-yet-effective patch re-scaling module (PRM) that functions identical but is more lightweight than multi-scale approaches. Experimental evaluation demonstrates that the proposed networks outperform the state-of-the-art in majority cases on four different benchmark datasets with up to 12.6% improvement for the RMSE evaluation metric. The better cross-dataset performance also validates the better generalization ability of our schemes. For the multi-modal input, effective feature-extraction (FE) and strong information fusion between two modalities remain a big challenge. Thus, the multi-modal novel network design focuses on investigating different features fusion techniques amid improving the FE. Based on the comprehensive experimental evaluation, the proposed multi-modal network increases the performance under all standard evaluation criteria with up to 33.8% improvement in comparison to the state-of-the-art. The application of multi-scale uni-modal attention networks also proves more effective in other deep learning domains, as demonstrated successfully on seven different scene-text recognition task datasets with better performance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the problem and motivation, background and challenges, our major contributions, and related work. First, we discuss the motivation and challenges related to the crowd estimation task. Next, we briefly explain different types of methods and approaches we have proposed and designed for effective crowd estimation. At the end of this chapter, we list down our major contributions, related work and the organization of this work. Most part of this chapter comes from introduction and related work sections of our work [79, 80, 81, 82, 83, 84].

1.1 Motivation

Deep learning has achieved significant progress in many computer vision applications, like image classification [9, 3], object detection [22, 18], face recognition [4], depth estimation [12, 13], image translation [40, 39], and crowd counting [82, 6, 85]. Crowd estimation aims to count the total number of people in the given image (image-only) or combination of different input modalities (e.g. Image+Audio). This dissertation mainly focuses on designing effective and independent crowd counting networks for two different types of inputs: Image-only (uni-modal) and Audio+Visual (multi-modal). Crowd estimation finds a very important and integral place in the crowd analysis paradigm. Crowd gatherings are ubiquitous and bound to happen frequently at sports, musical, political, and other social events. Accurate and automated crowd estimation can help in effectively organizing large crowd gatherings. Whereas, manual human-based crowd counting process is unreliable and ineffective due to the tedious and time-consuming nature of this task. In addition to the crowd estimation, the same methods can also be applied to other fields like the counting of animals, crops, and microscopic organisms [1, 28].

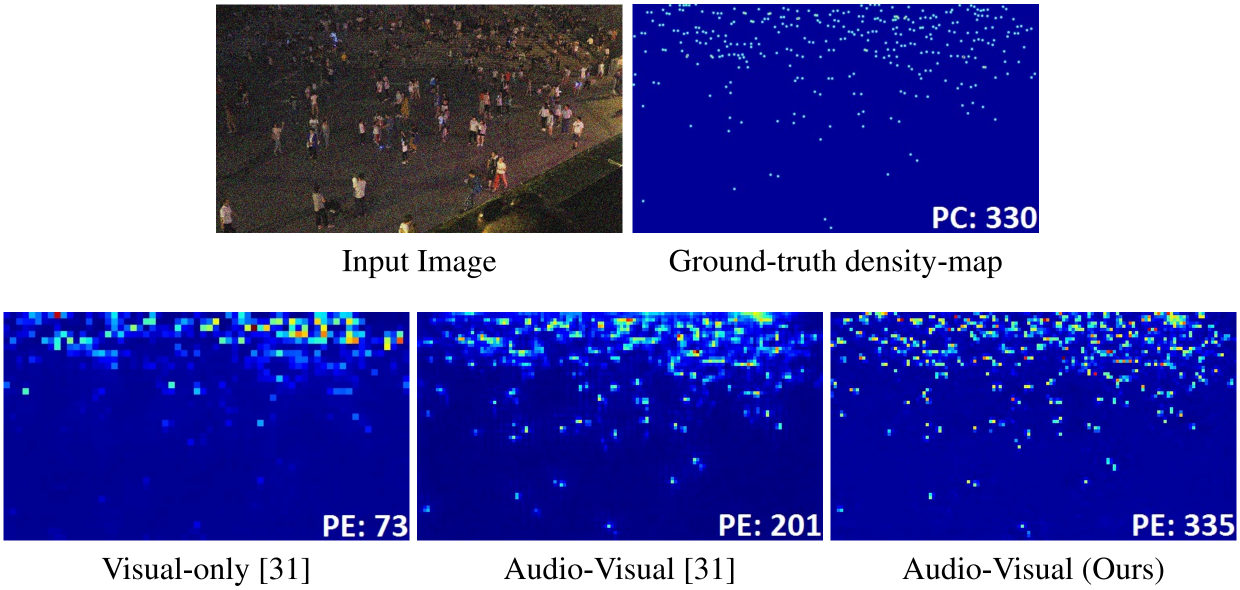
1.2 Background and Challenges

In recent years, computer vision has witnessed great developments in several sub-areas, such as image classification [40], object detection [20], image translation [42], and face recognition [2], with the introduction of convolution neural networks (CNNs). Inevitably, recent state-of-the-art crowd counting methods are overwhelmingly dominated by the CNN based approaches, which generally belong to either direct regression (DR) [7], [27], [36] based or density-map estimation (DME) [18], [24], [29], [33], [35], [41], [45] based architectures. Here, we separately discuss the background and challenges associated with uni-modal (Image-only) and multi-modal (Audio-Visual) crowd counting.

1.2.1 Uni-Modal Crowd Estimation

Automated uni-modal crowd counting comes up with different challenges including large perspective, huge crowd diversity across different images, severe occlusion, and dense crowd-like complex background patterns. The direct regression (DR) based methods directly regress or estimate the crowd number from the input image or patch. These methods alone do not prove effective for crowd counting due to huge crowd diversity and multi-scale variation in and across different images. The DME based methods perform crowd counting by estimating the crowd-density value per pixel. This type of approach, in general, also tends to struggle against the above-stated major issues and challenges.

Multi-column or multi-regressor CNN based architectures [24], [29], [33], [45] have proved to be very effective for crowd counting tasks. MCNN [45] is a state-of-the-art three column density-map estimation based end-to-end crowd counting network, where each CNN based column specializes in handling the specific crowd-density level. At the end of this network, all columns are merged together to yield the crowd estimate after the remaining processing. Similarly, multi-



Input Image Ground-truth density-map

Visual-only [31] Audio-Visual [31] Audio-Visual (Ours)

Figure 1.1: For the low-quality input image with severe conditions such as noise, low-illumination, or low-resolution, the proposed audio-visual model yields the best and more fine-grained people estimate (PE) as evaluated using the ground-truth density-map and people count (PC).

column based architectures [18], [29] utilize multiple specialized crowd count regressors to cope with multiple crowd-density scales separately. For example, Switch-CNN [29], a density map estimation based network, consists of a CNN based switch classifier that routes the input image or patch to one of three crowd count regressors, where each regressor deals with a specific crowd level. In addition, many single-column or single-regressor based architectures [17], [30] have also been proposed to address the crowd counting issues and challenges. These methods produce promising results, but still lack the generalization ability for crowd estimation, ranging from low to high crowd density.

1.2.2 Multi-modal (Audio-Visual) Crowd Estimation

Many deep learning-based image-only schemes [81, 85, 83, 35, 83, 40, 127, 60] have been proposed to date, ranging from single and multi-branch networks [127, 81, 83], multi-regressors [85] based to trellis networks [40]. Although they show reasonable performance in regular images, they fail to generalize well in many practical scenarios such as low illumination and lighting con-

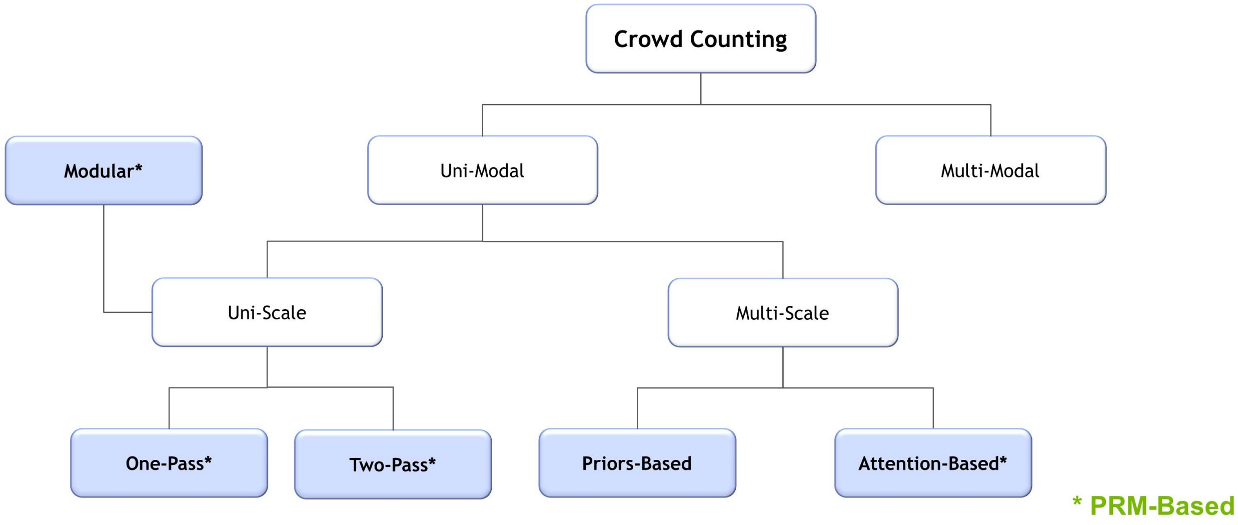


Figure 1.2: Different Branches of Our Crowd Estimation Work. Our research diversely ranges from uni-modal to multi-modal crowd counting.

ditions, noise, severe occlusion, and low-resolution images, where visual information is scarce. Consequently, they give huge crowd under-estimation as shown in Fig. 1.1. Lack of visual clues may also invoke highly sensitized behavior in these models towards different image regions, resulting in large over-estimation. Moreover, in the case of regular images (uni-modal), sub-optimal capabilities of these state-of-the-arts implicate that there is a lot of room for further improvement.

1.3 Our Work Overview

To address the above mentioned crowd estimation challenges, we investigate two different types of input modalities: uni-modal (Image) and multi-modal (Image and Audio) as shown in Fig. 1.2. In the uni-modal setting, we design five (four end-to-end and one modular) novel crowd counting networks with different architectures. Three of them are uni-scale channel resolution networks, where we deploy the novel and simple-yet-effective patch re-scaling module (PRM) to balance it to multi-scale capability. The remaining two belong to multi-scale fusion-based networks that focus on multiple channels with different resolutions. The uni-scale networks consume the input either once (one-pass) or twice (two-pass) based on the architectural design. We discuss all five uni-modal models individually in detail in the next chapters. In the multi-modal environment, we investigate the methods for effective integration and fusion of the Image and Audio inputs amid extracting the valuable and strong features for both modalities. We also apply our multi-scale technique on the scene-text recognition task with promising and effective results. All the uni-modal and multi-modal networks have been comprehensively and separately explained in the next chapters.

1.4 Contributions of Dissertation

The major contributions of our work are listed as follows:

1.4.1 Uni-Modal Crowd Counting Networks

- We extensively study the crowd estimation problem for the uni-modal (image-only) input and design several novel modules and networks to address key issues and challenges in this domain amid mitigating major SOTA limitations.

- We propose a conceptually simple yet effective and plug-and-play based patch rescaling module (PRM) to address the major huge crowd diversity issue in crowd counting problems.

- We present three new uni-modal and uni-branch crowd counting multi-task frameworks that utilize the lightweight PRM module instead of computationally expensive recent multi-column or multi-regressor based architectures.

- We devise a new multi-resolution and multi-branch feature-level fusion based end-to-end crowd counting approach for still images that effectively deals with significant variations of crowd-density, lighting conditions, and large perspective.

- We also propose an alternative to the patch rescaling module by more effectively using the input priors. Unlike the PRM, the proposed module fully utilizes all three crowd density levels without requiring any compromising or additional crowd-density classification process.

- We also present a novel multi-resolution and multi-task PRM based visually attended crowd-counting network for static images that effectively address major crowd counting challenges, including the issues of crowd-like background regions and huge crowd-variation.

- We deploy the plug-and-play PRM module so as to further push its boundaries and utilize it more effectively as compared to its previous deployments.

- We employ the visual attention mechanism in a unique and effective way on early-stage feature-maps that facilitate the later-stage channels to better understand the foreground regions.

- Experimental evaluation demonstrates that the proposed networks outperform the state-of-the-art methods in majority cases on four different benchmark datasets with up to 12.6% improvement in terms of the RMSE evaluation metric. Better cross-dataset performance also validates the better generalization ability of our schemes.

1.4.2 Multi-Modal Crowd Estimation

- We propose a novel audio-visual multi-task crowd counting network for effective estimation in both regular and severe conditions. To the best of our knowledge, this is the first attempt to use the transformer-style mechanism for this task.

- We introduce the notion of auxiliary PIR and PCE information, and empirically show that it is beneficial for better modalities association and extracting rich visual features without

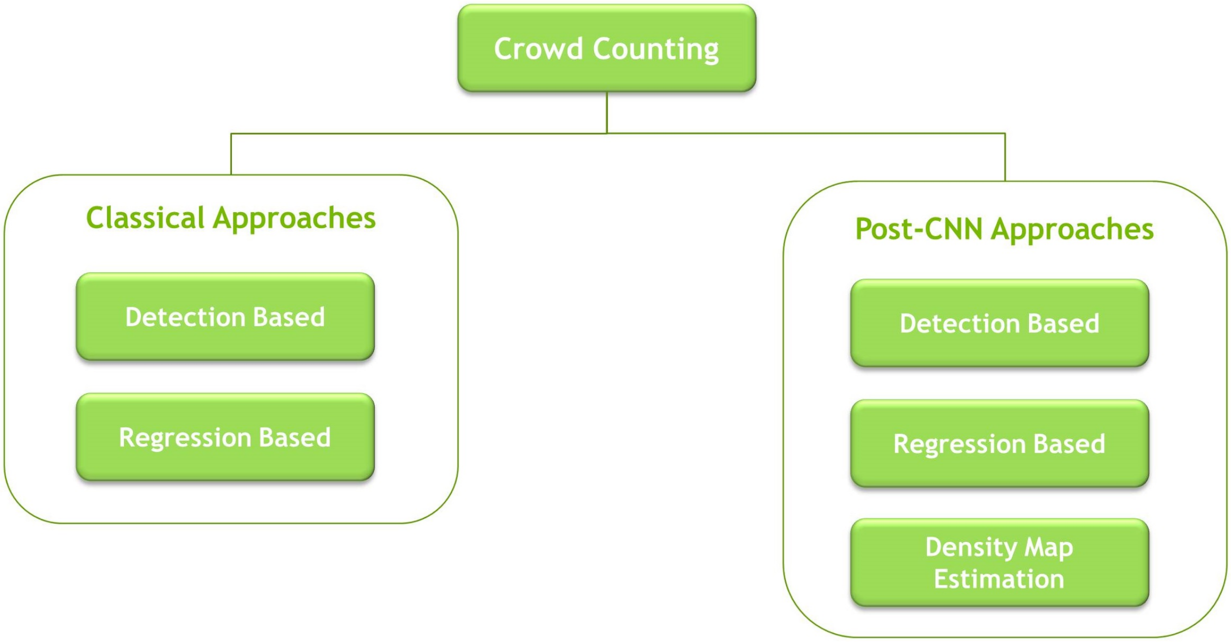


Figure 1.3: Crowd Counting methods category-wise. Most methods nowadays belong to either regression based or density map estimation based CNN methods.

requiring any extra ground-truth annotation process.

- We also design an image-only variant of our model. Extensive experimental evaluations on benchmark datasets indicate that the proposed networks significantly outperform the state-of-the-art.

1.5 Related Work

Different crowd estimation methods have been proposed over the time to address the key problems like huge crowd diversity, severe occlusions, cluttered crowd-like background regions, and large perspective changes. Uni-modal Crowd estimation research can be divided into pre deep-learning (classical) and post deep-learning era because of superior performance of deep-learning based techniques including convolution neural networks (CNNs) as shown in Fig. 1.3.

1.5.1 Uni-Modal Classical Techniques

Classical approaches can be categorized into two classes: Counting by detection and counting by regression. Count by detection classical methods first detect individual persons in the image [111, 107, 22, 52] using handcrafted features. Then, the final image count is obtained by the summation of all detections. These detectors fail quite drastically in the case of high-density crowd because of few pixels per person. Classical regression based methods [10, 12, 78] learn to directly map the image local patch to the crowd count. They yield better performance, however, they still suffer from lack of generalization on reasonable crowd diversity range.

1.5.2 Uni-Modal CNN-based (deep-learning) Techniques

CNN based methods have been widely used nowadays for crowd counting. They are of three types: Counting by detection, by regression, and density-map estimation based methods.

1.5.2.1 Counting by Detection

In detection based methods, CNN detectors (like YOLO, Faster-RCNN [23, 76]) detect each person in the image, followed by the sum of all detections to yield final crowd estimate. These methods fail in the case of high occlusion and crowd.

1.5.2.2 Counting by Regression

Regression based methods [102, 104] learn to directly regress the crowd count. Wang *et. al* [104] used the AlexNet [49] based scheme to directly regress and estimate the people number. These regressors alone do not generalize well for the huge crowd diversity range.

1.5.2.3 Density Map Estimation based Crowd Counting

The density-map estimation methods [127, 85, 94, 126, 59, 103, 91, 114, 40] estimate crowd density value per pixel. The final count is obtained from the sum of all pixel density values. These

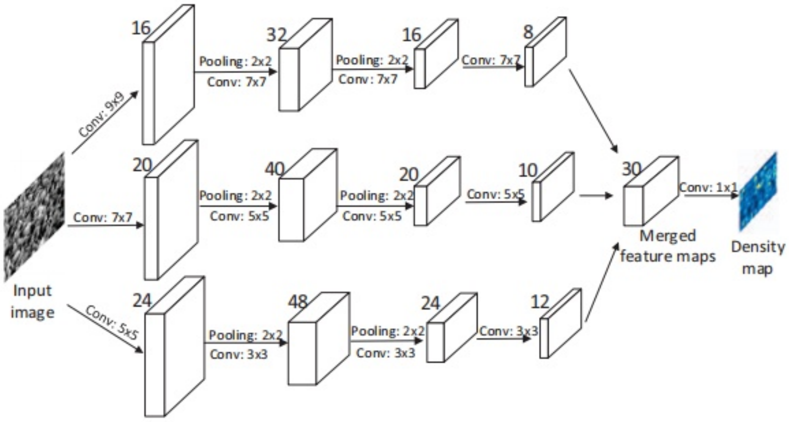


Figure 1.4: Multi-column convolutional neural network (MCNN) based on density map estimation method [127].

methods are widely used recently with better performance and focus on multi-scale or diverse crowd-density range handling. Zhang *et al.* [127] proposed a three-column architecture, where each column caters respective scale using different filter sizes, followed by a fusion of the columns to produce final crowd density-map as shown in Fig. 1.4. Similar to this idea, Switch-CNN [85] utilized three specialized count regressors to cater three different scales. Each input image routes to one of these regressors using a switch CNN-based classifier. Cascaded-mtl [94] generated density-map by first classifying image 10-way prior to actual density-map production. Recently, *Zhang et al.* [126] proposed a scale-adaptive network, which employs a single backbone structure and the same filter size and combines different layers feature maps to handle the multi-scale variation. Ranjan *et al.* [75] proposed a two-column network where first low-resolution based column feature maps are combined with high-resolution input column to generate final density map. Recently, Wan *et al.* [103] used support image density-map to predict the input image density-map by the residual regression based difference between the two density-maps. Xu *et al.* [115] first grouped patch-level density-maps into several density levels, followed by the automatic normalization via an online learning strategy with a multipolar center loss. Liu *et al.* [58] proposed a hybrid approach that coupled both detection and density-map estimation techniques, and used the appropriate counting mode based on the crowd-density.

1.5.3 Multi-Modal CNN-based (deep-learning) Networks

Like the uni-modal networks, multi-modal paradigm also focuses on the density-map estimation based methods. However, the multi-modal dimension stands at the early stage of research with just one noticeable work [31], where the authors combine image and audio to estimate the crowd density-map. Additionally, uni-modal work contains lots of standard datasets for analysis and comparison, whereas the multi-modal domain just published its first audio-visual crowd estimation dataset [31].

1.5.3.1 Audio-Visual Learning

Audio-visual representation learning aims to aid the visual modality with audio or vice-versa. Early speech perception research [66] demonstrates that the visual information can change what people hear, i.e., McGurk Effect. Since then, vision and audio modalities are widely explored in speech recognition[123, 25, 68, 96], video classification [112], emotion recognition [46] and video description [41]. Multiple kernels are broadly implemented as the fusion module by feeding the kernels with data from different modalities [11, 92, 110]. Another fusion method is based on graphical models considering its advantages in temporal related tasks [25, 28]. Besides, neural networks raise more attention in fusion especially since the appearance of RNN and LSTM [72, 96]. More recently, transformer-based [101] fusion raises growing attention [1, 98, 74, 34, 42], especially after its application in vision [19]. In addition to that, there are also some model-agnostic fusion methods, including the simple concatenation [50, 14, 122] and element-wise operation [21, 100].

1.5.3.2 Crowd Counting

The research of people count mainly focuses on image-only crowd estimation, and targets several issues such as varying crowd-density and scale, large perspective and heavy occlusions. They are of three categories: Count-by-detection (Det), by-direct-regression (DReg), and by-density-map (DMap). The Det methods [86, 53] detect each person via some standard object detectors (e.g.

Faster-RCNN [23], YOLO [76]). These methods give unsatisfactory results in the high-density crowd scenarios. The DReg models [81, 83, 104, 82, 104] directly regress the crowd number using CNN-based structures. Wang *et al.* [104] deployed the AlexNet [49] variant for direct crowd regression. Recently, Sajid *et al.* designed two different types of direct-regression counting methods [83, 82] that use the patch-rescaling module (PRM) and branch structure to deal with varying crowd levels. But these models fail to utilize the valuable density-map based computation. The DMap methods [35, 127, 85, 60, 95] estimate crowd density-map, where each pixel indicates crowd-density. The pixel values are then summed up to obtain the final count. Switch-CNN [85] uses CCN-based switch that routes the image to one of three specialized regressors, each dealing with a specific crowd-density. Li *et al.* [54] employed dilation layers for better contextual information retrieval. Liu *et al.* [60] used rank-based system for unsupervised learning. Idrees *et al.* [35] deployed composition loss to jointly learn the count, localization and density-map. HA-CCN [95] utilized global and spatial attention to enhance useful features. However, these schemes prove inadequate to handle extreme conditions such as noise, low illumination and resolution images.

In the audio-visual domain, Hu *et al.* [31] recently introduced the first-ever audio-visual crowd dataset, DISCO, making this type of crowd counting possible. For the audio-visual people count, how to constructively extract the audio-visual features and how to effectively fuse them together present the key challenges.

1.5.4 Comments on Existing Methods

These methods still struggle to handle the huge crowd diversity and thus do not generalize well. They also overestimate the cluttered background regions in the images. One major issue with the density-map estimation methods is to find the optimal Gaussian kernel size, which depends on many related factors. They also do not generalize well on the huge crowd-variation challenge. On the other hand, recent multi-modal models only account for the parametric influence of audio features on the visual ones without making full use of the audio-visual information. Therefore, we need to address the major crowd estimation issues and challenges in this work.

1.6 Organization

The rest of the dissertation is organized as follows. Next, we discuss five novel and different uni-modal crowd counting architectures separately in the next three chapters in detail. Following that, we detail our novel work on the multi-modal model design for people count in Chapter 5. Then, Chapter 6 discusses the application of our novel crowd counting technique to the scene-text recognition problem, followed by the discussion on the uni- and multi-modal crowd-counting based future work and our work conclusion in Chapter 7. Moreover, most content of this dissertation comes from our work as given in [79, 80, 81, 82, 83, 84].

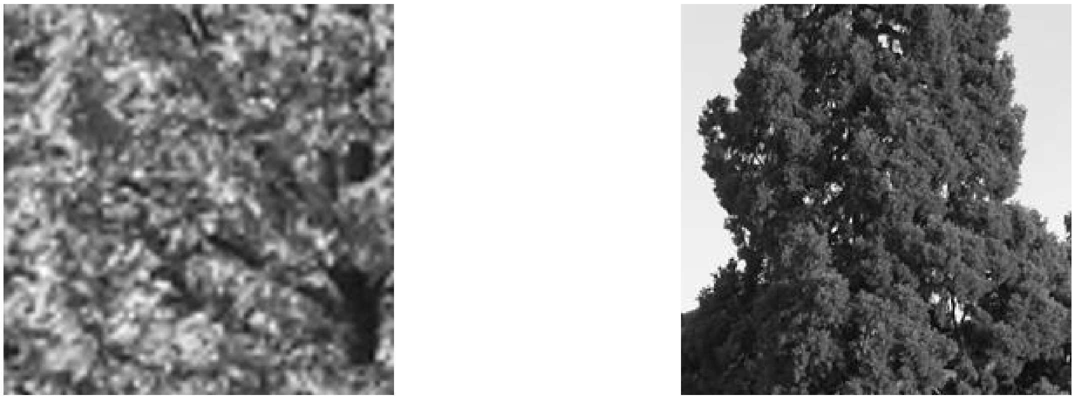
Chapter 2

Uni-modal and Uni-Scale PRM-based Multi-task Crowd Counting Networks

This chapter deals with the extensive discussion related to the novel patch re-scaling module (PRM) and three new uni-scale PRM-based crowd counting networks. Next, we state our main contributions, followed by a discussion regarding quantitative and qualitative evaluation on standard benchmark datasets and compare the results with the state-of-the-art methods. Cross-dataset evaluation demonstrates the generalization capability of the proposed scheme. Most of this chapter content has been taken from our work [82, 83].

2.1 Introduction to PRM module and Uni-scale Networks

Automated crowd counting comes up with different challenges including large perspective, huge crowd diversity across different images, severe occlusion, and dense crowd-like complex background patterns. Recent methods mostly employ deep convolutional neural networks (CNNs) to automate the crowd counting process. These approaches can be categorized as Detection based, Direct regression based, and Density map estimation based methods. Detection based methods use CNN based object detectors (e.g. Faster-RCNN [23], YOLO [76]) to detect each person in the image. The final crowd count is the sum of all detections. This idea does not generalize well for high-density crowd images, where detection fails due to very few pixels per head or person. Direct regression based methods [33, 82] learn to directly regress crowd count from the input image. These methods alone cannot handle huge crowd diversity and thus lack generalization. Density map estimation based methods [85, 94, 126, 59, 103, 91, 114, 40] estimate crowd density value



Ours=0, DR=61, DME=52

Ours=0, DR=27, DME=53

Figure 2.1: Current Direct Regression (DR) and Density-map estimation (DME) [35] based methods overestimate in case of even very small (224×224) size cluttered crowd-like background regions in images, as they face difficulty in recognizing and discarding such complex patterns.

per pixel instead of the whole input image. Most current state-of-the-art methods belong to this category due to their better and effective performance, however, limitations related to density map estimation per pixel pose a huge challenge [75] due to large variations in crowd number across different images.

Multi-column CNN (MCNN) model [127] is a three-column density-map estimation based network, that uses different filter sizes in each of its columns to account for multiple scales. Each branch is specialized in handling the respective scale. These columns are eventually concatenated at the end to output the final crowd estimate. Similarly, another state-of-the-art model, named Switch-CNN [85], deploys a hard switch to select one of three specialized crowd count regressors accordingly for the input image. Each count regressor is specialized to handle and focus on respective crowd density. Non-switch and single-column based models [126] are also being designed to solve the counting issues, but they lack the ability to generalize well on huge crowd diversity across different images, and thus, result in either high over-estimation or under-estimation.

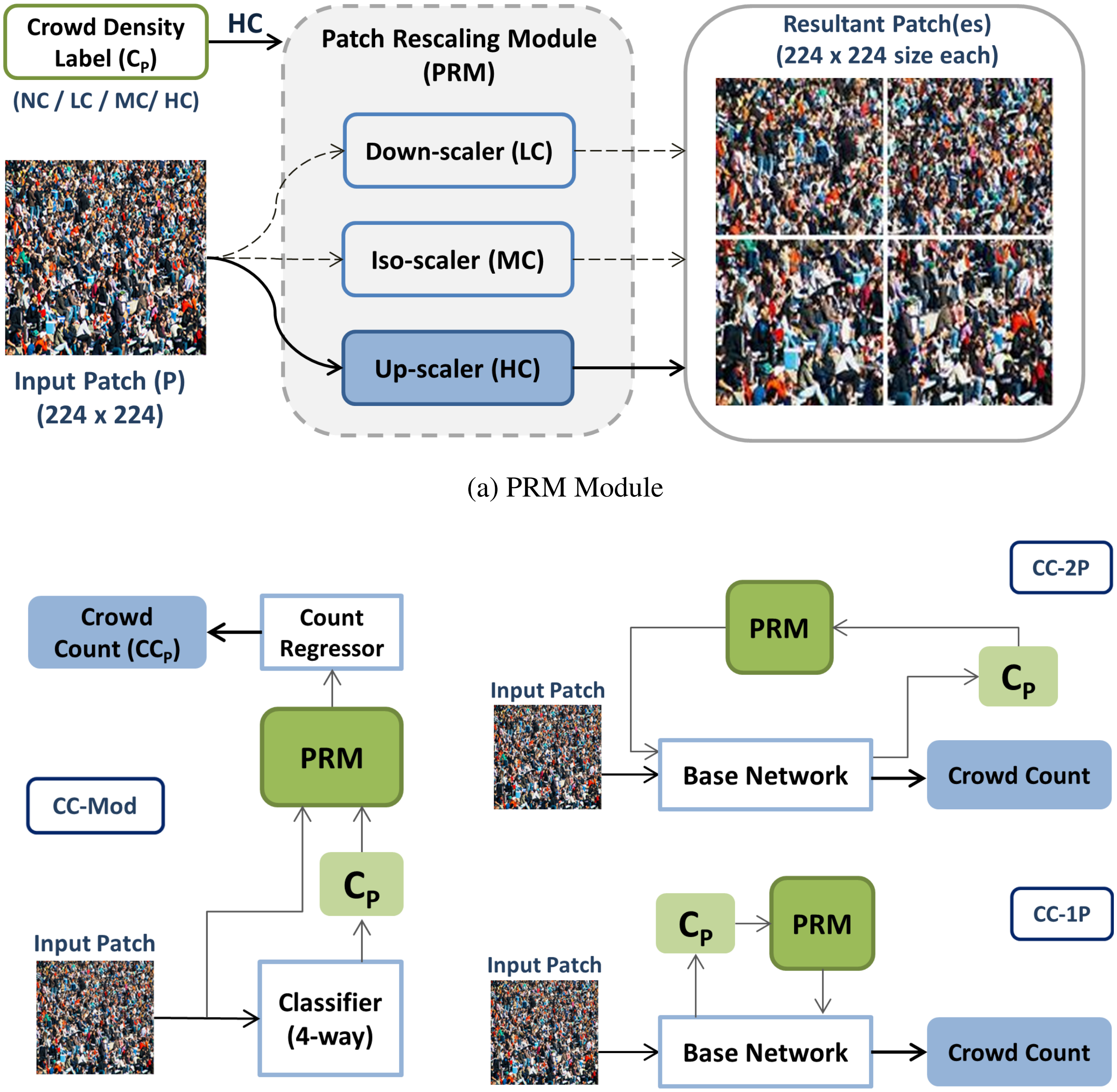
Another key issue with these methods is their noticeable inability and lack of focus towards detecting and discarding any cluttered crowd-like background region or patch in the image that may cause huge crowd over-estimation. As shown in Fig. 2.1, current methods do not detect these 224×224 cluttered crowd-like background regions in the images and thus result in crowd over-estimation. This problem would scale up quickly with more such regions occurring regularly in the images.

To address the aforementioned major issues, we propose a simple yet effective image patch rescaling module (PRM) and three new crowd counting frameworks employing the plug-and-play PRM module. These frameworks range from a modular approach to multi-task end-to-end networks. The lightweight PRM module addresses the huge crowd diversity issue efficiently and effectively, and also appears as a better alternative to the recent computationally heavy multi-column or multiple specialized count regressors based architectures. In the proposed frameworks, high-frequency crowd-like background regions also get discarded that may cause huge crowd over-estimation otherwise.

2.2 Our Main Contributions

The main contributions of this work are as follows:

- We propose a conceptually simple yet effective and plug-and-play based patch rescaling module (PRM) to address the major huge crowd diversity issue in crowd counting problems.
- We also propose three new and independent crowd counting frameworks that utilize the lightweight PRM module instead of computationally expensive recent multi-column or multi-regressor based architectures.
- Extensive experiments on three benchmark datasets show that our approach outperforms the state-of-the-art methods in terms of RMSE evaluation metric with the improvement up to 10.4%. Cross-dataset evaluation also demonstrates the better generalization ability of the proposed PRM module and crowd counting schemes relative to similar state-of-the-art methods.



(a) PRM Module

(b) Modular Scheme (CC-Mod)

(c) End-to-End Networks (CC-1P, CC-2P)

Figure 2.2: (a) **PRM Module.** Based on the prior estimated crowd-density class (C_P), the PRM module rescales the input patch P (when $C_P = HC\ or\ LC$) using one of its rescaling operations (*Up-scaler or Down-scaler*) and generates 4 or 1 new rescaled patch(es) respecively. The *MC* labeled patch bypasses any rescaling (*Iso-scaler*). (b) **CC-Mod**. In the modular crowd counting scheme, the input patch is first classified 4-way (NC, LC, MC, HC), followed by passing through the PRM and then through the regressor for final patch crowd count (CC_P). (c) **CC-1P, CC-2P.** These crowd counting networks couple the PRM module with the base network to address the huge crowd diversity issue amid giving better performance (Architectures detailed in Figs. 2.3 and 2.4).

2.3 Proposed Method

Our method focuses on addressing huge crowd diversity within as well as across different images and the presence of cluttered crowd-like background regions in these images. As shown in Fig. 2.2(a), the proposed Patch Rescaling Module (PRM) comprises of three straightforward operations i.e. *Up-scaler, Down-scaler and Iso-scaler*. The input image patch uses one of these operations to adjust its scaling accordingly, depending on its crowd density level. This lightweight rescaling process helps in addressing the crowd diversity issue efficiently and effectively. Next, we propose three new and different crowd counting schemes that employ the plug-and-play PRM module as shown in Fig. 2.2. These frameworks include a modular approach (Fig. 2.2b) and two end-to-end (Fig. 2.2c) networks. The modular framework uses the PRM block in between the independent classification and regression modules, while end-to-end multi-task networks utilize the PRM to facilitate the base network for better and efficient performance. Both PRM and the proposed frameworks are detailed in the following text.

2.3.1 Patch Rescaling Module (PRM)

The PRM module, as shown in Fig. 2.2a, is used to rescale the input patch by utilizing one of the two rescaling operations, namely *Up-scaler* and *Down-scaler*. PRM module selects the appropriate rescaling operation based on the crowd density level (C_P), which is computed prior to the PRM module usage by the 4-way classification (*no-crowd (NC), low-crowd (LC), medium-crowd (MC), high-crowd (HC)*). Crowd patches, classified as *LC or HC*, pass through the *Down-scaler* or *Up-scaler* operation, respectively. The *MC* labeled input patch bypasses the PRM without any rescaling (denoted by *Iso-scaler*). *NC* labeled input patch is automatically discarded without any PRM processing as it is a background region with zero people count. Every patch, coming out of the PRM module, will have the same fixed 224×224 size. By using the right scale for each input patch, the straightforward PRM module addresses the huge crowd diversity challenge and has been used as a plug-and-play block in different new crowd counting schemes given in Sec.

2.3.2. Each rescaling operation is detailed below.

Upscaling Operation (Up-scaler). The upscaling option is applied to the patches with high-crowd (HC) crowd. Up-scaler divides the input patch into four 112×112 patches, followed by upscaling of each new patch to 224×224 size. Intuitively, this simple operation facilitates the counting process by further dividing and zooming-in into each sub-divided part of the highly dense crowd patches separately. Consequently, it avoids the overestimation that occurs in complex multi-column architectures and multiple specialized count regressors based methods. Thus, this operation outputs four rescaled patches from the input patch.

Downscaling Operation (Down-scaler). The patches that are classified as low-crowd (LC) label are subjected to downscaling operation, where the patches are first down-scaled by $2\times$ and then zero-padded to 224×224 before proceeding for further processing. Primarily, this operation helps in avoiding underestimation by using smaller area for the input patch and achieves better results without the need for any specialized or complex additional component.

Iso-scaling block. The image patches that are labeled as medium-density (MC) class do not require any special attention as given to LC or HC based patches, because the deep CNN based crowd counting models can handle these cases effectively without using any upscaling or downscaling operation. Thus, they are directly forwarded to the next stage for crowd estimation.

2.3.2 PRM based Crowd Counting Frameworks

In this section, we discuss three independent proposed crowd counting schemes, ranging from a modular framework to two end-to-end multi-task networks. These methods address the huge crowd diversity using the PRM module as well as discard any cluttered background regions in the images. In each scheme, the input image is divided into 224×224 non-overlapping patches. Each patch then passes through that specific scheme for patch crowd count estimate. The final crowd count of the image is obtained by summing all its patches count. Each crowd counting scheme is discussed in the following subsections.

2.3.2.1 Modular Crowd Counting Scheme (CC-Mod)

The modular crowd counting framework (CC-Mod), as shown in Fig. 2.2b, consists of three main components, namely Deep CNN based 4-way classifier, PRM module, and crowd count regressor. Input image gets divided into 224 x 224 size non-overlapping patches. Each patch is then fed to a 4-way classifier that categorizes the input patch to its appropriate crowd-density label (*NC, LC, MC, HC*). Based on the assigned class label, each patch is rescaled accordingly using the PRM module before proceeding to the count regressor for the patch-wise crowd estimate. Image crowd count is finally obtained by summing all its crowd patches count. Each component has been detailed as follows.

Crowd Density Classifier. The goal of this module is to classify the input (224×224) image patch into one of the four crowd density labels, namely no-crowd (NC), low-crowd (LC), medium-crowd (MC), and high-crowd (HC) crowd. The definitions of these labels are given in the next paragraph. Based on the assigned class label, each patch will then be routed to the PRB module for the rescaling operation. The *NC* labeled patches are completely discarded without any further processing. Thus, using this specialized deep CNN classifier, we identify and discard the cluttered crowd-like background patches, which may result in huge accumulated crowd overestimation otherwise.

Crowd-density class labels definitions. Crowd density classifier requires the definitions of four class labels (NC, LC, MC, HC) to train and learn the 4-way classification. Since each labeled benchmark dataset contains the people localization information in the form of (x, y) as the center of each person's head, we utilize this information to define the class labels and generate training patches for each class label. The ground truth crowd-density label ($C_{P(gt)}$) for the 224×224 training image patch P is assigned as follows:

$$C_{P(gt)} = \begin{cases} NC & c_{gt} = 0 \\ LC & 0 < c_{gt} \leq 0.05 * c_{max} \\ MC & 0.05 * c_{max} < c_{gt} \leq 0.2 * c_{max} \\ HC & 0.2 * c_{max} < c_{gt} \end{cases} \tag{2.1}$$

where c_{gt} denotes the ground truth people count for the image patch X, c_{max} stands for the possible maximum ground truth people count in any 224×224 image patch of this benchmark dataset. Image patch, containing at most 5% of the maximum ground truth people count (and non-zero) is assigned with low-crowd (LC) crowd label. Similarly, a patch with actual count between 5 to 20% (including 20%) is classified with MC label, whereas patches containing more than 20% of the maximum possible crowd count or no crowd at all are given HC or NC labels respectively. In this way, a total of $90,000$ patches ($22,500$ per class) are being generated for the classifier training during each dataset experiment separately. In all proposed schemes, we use the same class definitions for the 4-way classification.

Classifier and Crowd Count Regressor details. We use customized DenseNet-121 [33] based architecture as the 4-way crowd-density classifier. Empirically, we found that using only the first three dense blocks give almost the same and consistent performance for this 4-way classification task, instead of using default four dense blocks in DenseNet-121. Consequently, this reduces the learnable parameters by a huge margin ($6.95M$ to $4.80M$). At the end of the third dense block, the classification layer is composed of 7×7 global average pooling followed by the 4D fully connected layer with a softmax 4-way classification (NC, LC, MC, HC) and cross-entropy loss.

The regressor is similar to the original DenseNet-121 architecture except that it has $\{6, 12, 18, 12\}$ sets in four dense blocks respectively instead of $\{6, 12, 24, 16\}$. This customization significantly reduces the model parameters ($6.95M$ to $5.05M$), while performing reasonably well. In addition, the classification layer has been replaced with a single neuron to directly regress the crowd count. The mean Squared Error (MSE) has been used as the loss function for the count regressor cr,

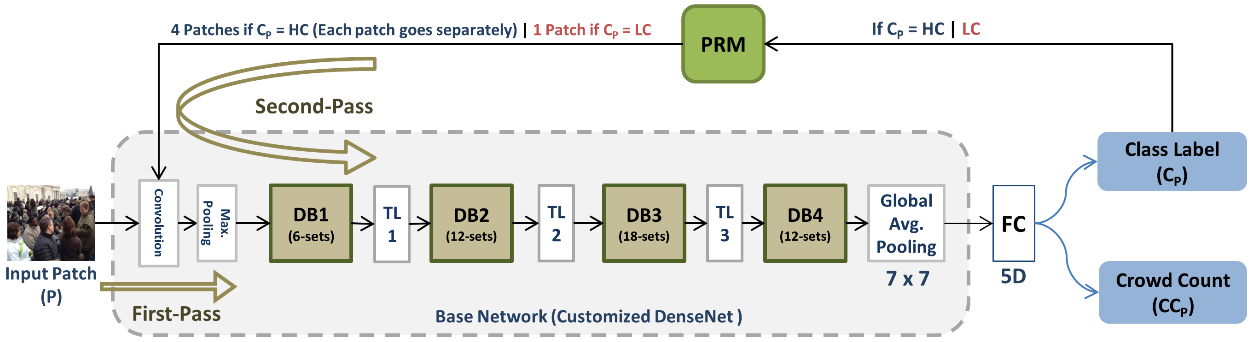


Figure 2.3: The proposed CC-2P architecture. The input patch *P*, classified as either *HC* or *LC* during the first-pass, passes through the base network again (second-pass) after the required PRM rescaling operation. Final patch count (CC_P) is the average of both passes crowd estimates. *MC* labeled input patch skips the second-pass without any PRM usage, and outputs the final first-pass Crowd Count (CC_P).

defined as:

$$Loss_{cr} = \frac{1}{n}\sum_{k=1}^{n}(F(x_k,\Theta) - y_k)^2 \tag{2.2}$$

where n denotes the training patches per batch, y_k is the actual crowd count for the input image patch x_k, and F is the mapping function that learns the input patch x_k mapping to the crowd count with weights parameters Θ.

2.3.2.2 Two-Pass Crowd Counting Network (CC-2P)

CC-2P, as shown in Fig. 2.3, is a multi-task 4-way classification and regression based network, that employs the PRM module. The input patch goes through the base network, consisting of four dense blocks (*DB1, DB2, DB3, DB4*), in the first pass to yield the crowd density class label as well as crowd number estimate. The patches, labeled as LC or HC label, proceed to the PRM module for required rescaling operation. The resultant new patch(es) then go through the base network for crowd count estimate in the second pass. Let P be the original input patch and first-pass class label (C_P) as LC or HC, then the final crowd count (CC_P) estimate for P is the average of the first-pass (fp) and the second-pass (sp) crowd estimates as follows.

$$CC_p = \begin{cases} \frac{cc_{fp} \ + \ cc_{sp}}{2} & C_p = LC \\ \frac{cc_{fp}+(cc_{u1}+cc_{u2}+cc_{u3}+cc_{u4})_{sp}}{2} & C_p = HC \end{cases} \tag{2.3}$$

Since the PRM produces four new upscaled patches $(u1, u2, u3, u4)$ for the input patch P when $C_p = HC$, therefore second-pass crowd count is the sum of these patches for this case. These four patches go through the network one by one during the second-pass to estimate their corresponding crowd counts. Input patch P that is labeled as MC in the first-pass, skips the second-pass as the PRM module has no effect on such patches (no rescaling). Also, NC labeled input patch is discarded without any further processing irrespective of their crowd estimate.

Network Details. We use customized DenseNet-121 [33] as our base network. Empirically, it has been observed that fewer sets of (1×1) and (3×3) layers in the Densenet-121 deeper dense blocks ($DB3$ and $DB4$) give almost the same and consistent performance for this problem amid reducing model parameters by a significant margin ($6.95M$ to $5.05M$). Consequently, we use {6,12,18,12} sets instead of {6,12,24,16} in the four dense blocks respectively, which reduces the 121 layers deep Densenet to 101 layers. Transition layers $(TL1, TL2, TL3)$ connect the dense blocks and adjust the feature maps size for the next dense block accordingly, as given in [33]. At the end of the base network, the final fully connected (FC) layer outputs the softmax based 4-way classification and regression based crowd count value. Multi-task loss ($Loss_{total}$) of CC-2P is defined as follows.

$$Loss_{total} = L_{reg} \ + \ L_{class} \tag{2.4}$$

where L_{reg} is the MSE loss as defined in Eq. 2.2, L_{class} is the cross-entropy loss for the softmax based 4-way labeling.

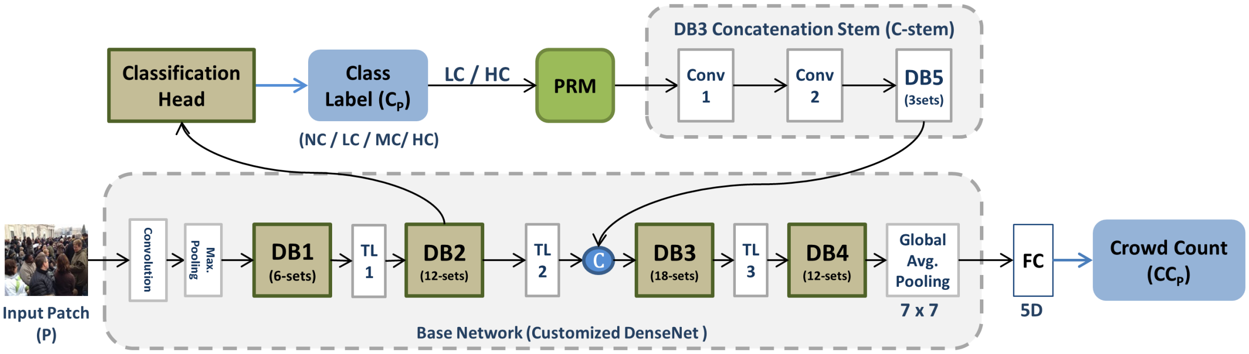


Figure 2.4: The proposed CC-1P architecture branches-out the dense block ($DB2$) output for the 4-way classification (C_P). The input patch P then passes through the PRM for any required rescaling. Resultant patch(es) then go though the *C-stem* block, followed by the channel- wise concatenation with the transition layer ($TL2$) output channels. Remaining processing finally yields the patch crowd count (CC_P).

2.3.2.3 Single-Pass Crowd Counting Network (CC-1P)

The multi-task single-pass network, as shown in Fig. 2.4, branches out the *dense block 2* (DB2) output for the 4-way classification (*NC, LC, MC, HC*) of the input patch. Based on the assigned class label, the input patch P passes through the PRM module for any required rescaling. Patch(es), coming out of the PRM, proceed to the DB3 concatenation stem (C-stem) for the extraction of their initial feature maps that are eventually concatenated with the *second transition layer* (TL2) output feature maps to serve as the input to the DB3 Finally, the global average pooling is being done on DB4 output channels followed by a single neuron to directly regress the input patch crowd count. The configurations of classification head and C-stem are shown in Table 2.1. Base network is the same as used in CC-2P except that the *compression factor* (θ) for *second transition layer* (TL2) has been set to 0.25 (instead of standard DenseNet-121 value of 0.5) to yield the same number of channels (256) after the concatenation process. Similar to the CC-2P scheme, the PRM generated four patches (when $C_p = HC$) go through the C-stem and subsequent blocks one by one to yield their corresponding crowd counts that are summed to output the final crowd estimate for the input patch P in this case. Empirically, it has been observed that the branching-out of the classification head after the DB2 achieves better results as compared to the same branching being deployed after

Layer	Output Size	Filters (F)
Classification Head		
	$512 \times 28 \times 28$	
	$64 \times 28 \times 28$	(1×1) conv, 64F
	$64 \times 14 \times 14$	(2×2) Avg Pooling, stride 2
	$32 \times 7 \times 7$	(3×3) conv, stride 2, padding 1, 32F
	4D FC, softmax	-
DB3 Concatenation stem block (C-stem)		
	$1 \times 224 \times 224$	
conv1	$64 \times 112 \times 112$	(3×3) conv, stride 2, padding 1, 64F
conv2	$32 \times 56 \times 56$	(3×3) conv, stride 2, padding 1, 64F
	$32 \times 28 \times 28$	(2×2) Avg Pooling, stride 2
DB3	$128 \times 28 \times 28$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 3$
	$128 \times 14 \times 14$	(2×2) Avg Pooling, stride 2

Table 2.1: Configurations of the CC-1P Classification Head and the C-stem block. Each conv represents the BN-ReLU-Convolution sequence [33].

other dense blocks as detailed in Sec. 2.5.4.

Proposed Approach and the Switch-CNN comparison. Switch-CNN [85], as detailed in Sec. 2.1, also classifies the input patch into the appropriate density level, followed by the crowd estimation using one of three specialized regressors. However, we approach this task in a totally different way by just using the straightforward plug-and-play PRM module with no learnable parameters and employing only one regressor or the base network. Whereas, the Switch-CNN uses complex coupling of the classifier with three specialized regressors. Consequently, the proposed frameworks (CC-Mod, CC-1P, CC-2P) have fewer model parameters ($9.85M, 6.7M, 5.05M$, respectively) as compared to the Switch-CNN ($15.1M$), and achieve better performance (see Sec. 2.5).

2.4 Evaluation and Training Details

As per the standard crowd counting evaluation criteria, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics have been used:

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|C_k - \hat{C_k}|, RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(C_k - \hat{C_k})^2} \tag{2.5}$$

where N indicates the total number of test images in the given benchmark dataset, and C_k, $\hat{C}_k$ represent the actual and the estimated counts respectively for test image k.

Training Details. In the end-to-end networks, the modular classifier and the crowd count regressor were trained separately using 90,000 patches each with mixed crowd numbers and original patch sizes of 112 × 112, 224 × 224, and 448 × 448. We used batch size of 16, stochastic gradient descent (SGD) as the optimizer and trains for 75 epochs with multi-step learning rate that starts at 0.001 and decreases to half each time after 25 and 50 epochs. Other parameters remain the same as for orignial DenseNet [33]. As per standard, 10% training data has been used for model validation.

2.5 Experimental Results

In this section, we report results obtained from extensive experiments on three diverse benchmark datasets: ShanghaiTech [127], UCF-QNRF [35], and AHU-Crowd [32]. These datasets vary drastically from each other in terms of crowd diversity range, image resolution, and complex cluttered background patterns. First, we analyze standard quantitative experimental results and ablation study on these datasets, followed by the cross-dataset evaluation. In the end, we analyze some qualitative results.

2.5.1 ShanghaiTech Dataset Experiments

The ShanghaiTech Part-A [127] dataset consists of diverse 482 images, with a predefined split of 300 training and 182 testing images. The proposed PRM based schemes are compared with the

	ShanghaiTech		UCF-QNRF	
Method	MAE	RMSE	MAE	RMSE
MCNN [127]	110.2	173.2	277	426
Cascaded-MTL [94]	101.3	152.4	252	514
Switch-CNN [85]	90.4	135.0	228	445
SaCNN [126]	86.8	139.2	-	-
IG-CNN [3]	72.5	118.2	-	-
ACSCP [87]	75.7	102.7	-	-
CSRNet [54]	68.2	115.0	-	-
CL[35]	-	-	132	191
CFF [91]	65.2	109.4	**93.8**	146.5
RRSP [103]	63.1	96.2	-	-
CAN [59]	**62.3**	100.0	107	183
TEDNet [40]	64.2	109.1	113	188
L2SM [115]	64.2	98.4	104.7	173.6
Densenet121[33]	93	139	167	229
CC-Mod (**ours**)	73.8	113.2	107.8	171.2
CC-2P (**ours**)	67.8	**86.2**	94.5	**141.9**
CC-1P (**ours**)	69.1	109.5	97.3	153
(CC-Mod/CC-2P/CC-1P) w/o PRM	93.8	139.2	168	230

Table 2.2: ShanghaiTech [127] and UCF-QNRF [35] datasets experiments and ablation study. Our PRM based approach (CC-2P) outperforms the state-of-the-art methods under the RMSE metric while giving competitive performance on MAE. Other PRM based proposed methods (CC-Mod and CC-1P) also give reasonable results. During the ablation study (last row), all proposed schemes give worse results after removing the PRM module, thus, indicating the quantitative importance of the proposed PRM.

state-of-the-art methods as shown in Table 2.2, where our approach (CC-2P) outperforms others under the *RMSE* evaluation metric with a significant improvement of 10.40% (96.2 to 86.2) and also give reasonable performance on MAE. The smallest RMSE value also indicates the lowest variance of our approach as compared with the other methods. Other proposed schemes (CC-Mod, CC-1P) also give comparable results in comparison.

To further evaluate the proposed methods, we removed the PRM module from each proposed scheme separately during the ablation study. After the PRM module removal, all three proposed schemes just become the same customized DenseNet based crowd count regressor (the base network), thus giving the same ablation performance as indicated by the last row of Table 2.2. The

Method	MAE	RMSE
Haar Wavelet [71]	409.0	-
DPM [20]	395.4	-
BOW–SVM [18]	218.8	-
Ridge Regression [12]	207.4	-
Hu et al. [32]	137	-
DSRM [120]	81	129
Densenet121[33]	88.2	126.1
CC-Mod (**ours**)	75.1	121.2
CC-2P (**ours**)	**66.6**	**101.9**
CC-1P (**ours**)	70.3	107.2
(CC-Mod / CC-2P / CC-1P) w/o PRM	89.9	127

Table 2.3: Our approach outperforms other models under all evaluation metrics on AHU-Crowd dataset. Ablation study (last row) also demonstrates the quantitative importance of the PRM module.

ablation results show that the performance decreases dramatically (MAE: 27.71%, RMSE: 38.07% error increase for CC-2P) without the PRM module, hence, quantifying the importance and effectiveness of the proposed PRM module.

2.5.2 UCF-QNRF Dataset Experiments

UCF-QNRF [35] dataset is the most diverse and challenging crowd counting benchmark dataset to date due to higher image resolutions, huge crowd diversity across different images and complex cluttered regions. It consists of 1,535 images with 1,251,642 annotations in total and a predefined training/testing split of 1,201/334 images, respectively. Also, the image resolution varies greatly from as low as 300×377 to as high as 6666×9999. As compared with the state-of-the-art models, our *CC-2P* approach outperforms them under RMSE evaluation metric while performing reasonably closer in terms of RMSE, as shown in Table 2.2. Our method shows a significant improvement as RMSE drops by 5.12% (146.5 to 139). The ablation study (last row of Table 2.2), same as in Sec. 2.5.1, quantifies the importance of the PRM module (MAE: 43.75%, RMSE: 38.3% error increase for CC-2P) after removing it from the proposed schemes.

	ShanghaiTech		UCF-QNRF	
Method	MAE	RMSE	MAE	RMSE
CC-Mod				
Using VGG-16	79.3	125.6	128	181
Using VGG-19	78.9	124.1	122	179
Using ResNet-50	77.2	121.2	120	177
Using ResNet-101	77.0	120.5	121	176
Customized DenseNet-121 (ours)	**73.8**	**113.2**	**107.8**	**171.2**
CC-1P Branching-out				
After DB1	78.3	123.2	128	181
After DB2 (ours)	**69.1**	**109.5**	**97.3**	**153**
After DB3	73.2	116.3	120	177
After DB4	79.1	124.2	121	176

Table 2.4: Ablation Study on the CC-Mod architecture choice and the CC-1P Branching-out effect. The results justify our use of customized DenseNet-121 architecture as the 4-way classifier and the count regressor in the CC-Mod framework, and also our DB2 based branching-out selection of the Classification-Head in the CC-1P model.

2.5.3 AHU-Crowd Dataset Experiments

AHU-Crowd dataset [32] consists of 107 images and as per standard convention, we perform 5-fold cross-validation. As shown in Table 2.3, comparison based results show that our methods outperform other state-of-the-arts in terms of all evaluation criteria. Ablation study (last row of Table 2.3), same as in Sec. 2.5.1, demonstrates the importance and effectiveness of the PRM module.

2.5.4 Ablation Study on CC-Mod Architecture Choice and CC-1P Branching-out Effect

In this experiment, we first explore different state-of-the-art architectures that can be used as the CC-mod classifier and the regressor. As shown in Table 2.4, our customized DenseNet-121 choice performs the best in all cases. All other architectures are adapted for this ablation study as described in Sec. 5.6 of [85]. Next, we analyze the possible branching-out of the classification-head after each dense block (DB1, DB2, DB3, DB4) separately in CC-1P. Again, Table 2.4 justifies our

Method	MAE	RMSE
Cascaded-mtl [94]	308	478
Switch-CNN [85]	301	457
CC-Mod (**ours**)	251	333
CC-2P (**ours**)	**219**	**305**
CC-1P (**ours**)	227	318

Table 2.5: Cross-dataset evaluation. Models are trained using ShanghaiTech Part A images patches and tested on the UCF-QNRF dataset. Results show the generalization ability of the proposed method.

DB2 based branching-out in CC-1P with least error.

2.5.5 Cross-dataset Evaluation

We perform cross-dataset evaluation and compare the results with the state-of-the-art models. Each model is trained on the ShanghaiTech part-A training dataset and evaluated on UCF-QNRF dataset testing images. Results are shown in Table 2.5, where we compare our method with two state-of-the-art models. These results demonstrate that our approach is much more reliable with better generalization ability as it yields the best performance with a decrease in MAE (from 301 to 219) and RMSE (from 457 to 305).

2.5.6 PRM Rescaling Operations Usage and Background Detection Analysis

Here, we make a quantitative analysis of each PRM rescaling option usage as well as the amount of background (NC) patches being detected and discarded by the proposed scheme (CC-2P) during the test on three benchmark datasets. In all benchmark evaluations, either PRM rescaling option (*HC* or *LC*) has been used for at least 14.25% and as high as 29.4% of the test images patches as shown in Fig. 2.5. Thus, the PRM have been utilized quite frequently and played an imperative role in enhancing the overall performance. Similarly, 36.2%, 30.9% and 32% (on average) of test image patches in ShanghaiTech, UCF-QNRF and AHU-Crowd datasets, respectively, have been detected as no-crowd (NC) and discarded after classification. These background patches could

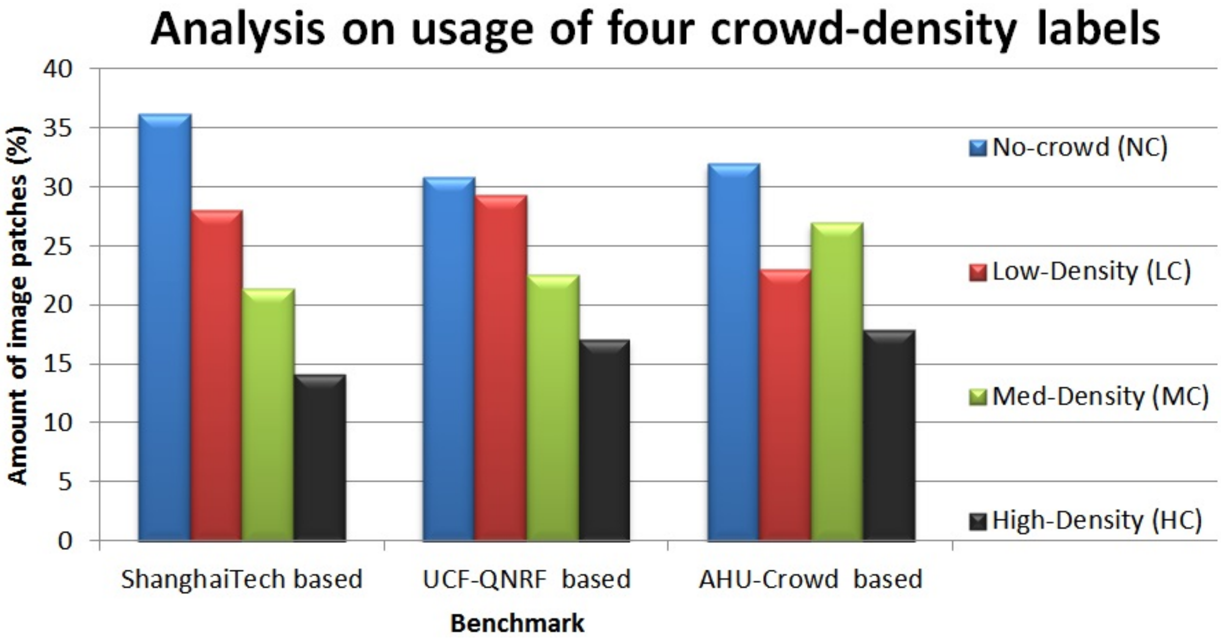
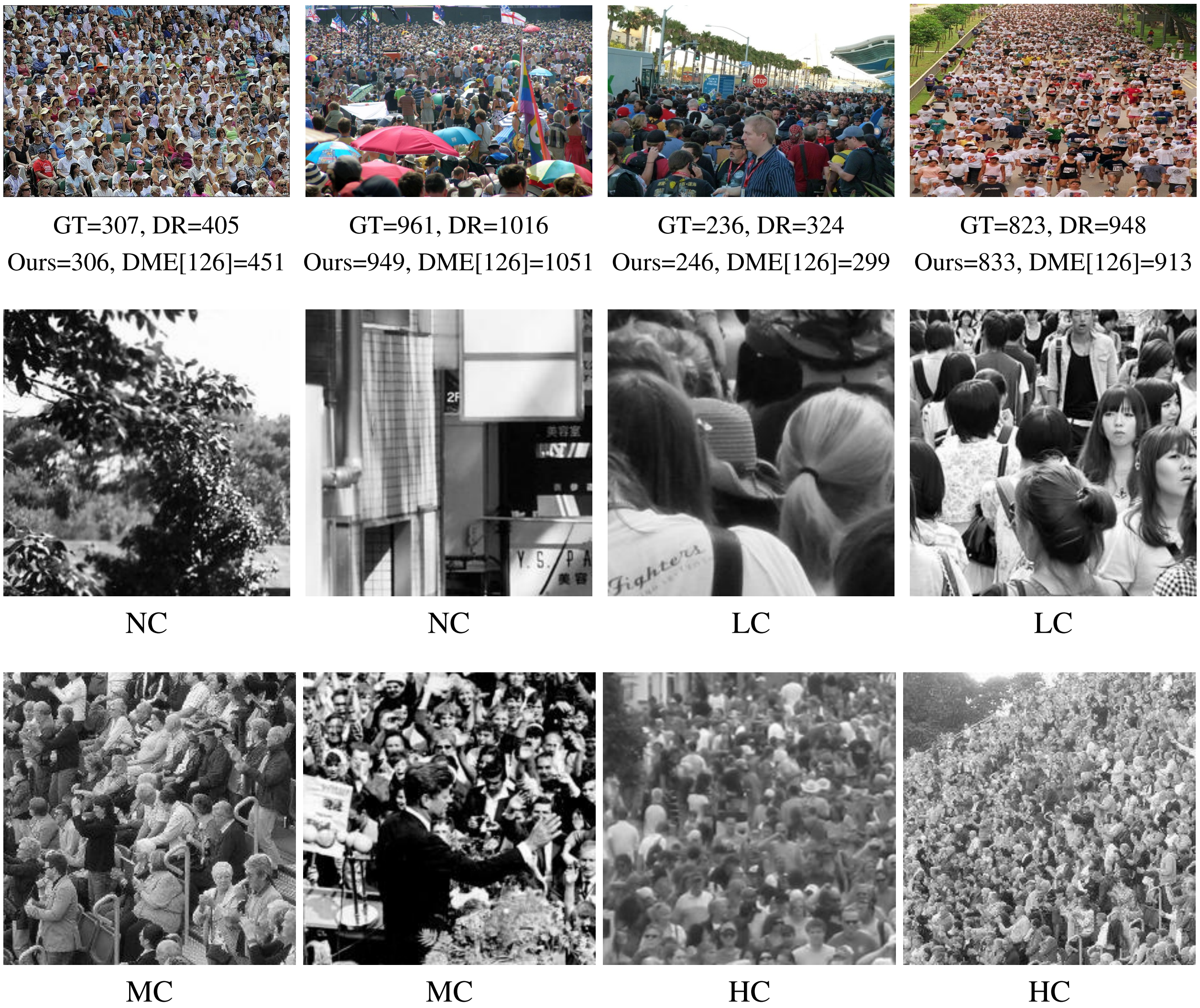


Figure 2.5: 4-way Classification (C_P) results on each benchmark dataset reveal the frequency and importance of the PRM rescaling operations (as applied on LC and HC labeled patches). It also indicates that a large number of patches have been classified as no-crowd (NC) and thus discarded to avoid overestimation.

have created a great crowd overestimation otherwise as described in Sec. 2.1.

2.5.7 Qualitative Results

Some qualitative results have been shown in Fig. 2.6, where the first row demonstrates the crowd count results on actual test images. As compared to Direct regression (DR) [33] and Density map estimation (DME) based methods, it is evident that our approach yields more accurate and reliable results. Sample crowd-density classification results, as shown in the second and third rows of the same figure, demonstrate the effectiveness of the 4-way classification, which is crucial in routing the test patch to the correct PRM rescaling operation as well as in detecting any background image patch.



GT=307, DR=405
Ours=306, DME[126]=451

GT=961, DR=1016
Ours=949, DME[126]=1051

GT=236, DR=324
Ours=246, DME[126]=299

GT=823, DR=948
Ours=833, DME[126]=913

NC

NC

LC

LC

MC

MC

HC

HC

Figure 2.6: Qualitative results. First row shows the visual results from the actual test images. As compared with Direct Regression (DR) [33] and density-map estimation (DME) methods, our approach yields closer to the ground truth (GT) results. Second and third rows show our 4-way classification results, where it labels these complex patches correctly, thus, helping in routing the patches to the correct PRM rescaling option and also discards any no-crowd patch.

2.6 Conclusion

In this chapter, we have presented an effective PRM module and three independent crowd counting frameworks. The proposed frameworks employ straightforward PRM rescaling operations instead of complex multi-column or multiple specialized crowd count regressors based architectures. The experimental results show that the proposed approach outperforms the state-of-the-art methods in terms of the RMSE metric and achieves competing performance in the MAE metric. The cross-dataset examination also indicates the great generalization ability of our method.

Chapter 3

Multi-Resolution Fusion and Multi-scale Input Priors Based Crowd Counting (Uni-Modal)

In this chapter, we discuss the first multi-scale crowd counting network belonging to the uni-modal category. The state-of-the-art patch rescaling module (PRM) based approaches prove to be very effective in improving the crowd counting performance. However, the PRM module requires an additional and compromising crowd-density classification process. To address these issues and challenges, this work proposes a new multi-resolution fusion based end-to-end crowd counting network. It employs three deep-layers based columns/branches, each catering the respective crowd-density scale. These columns regularly fuse (share) the information with each other. The network is divided into three phases with each phase containing one or more columns. Three input priors are introduced to serve as an efficient and effective alternative to the PRM module, without requiring any additional classification operations. Along with the final crowd count regression head, the network also contains three auxiliary crowd estimation regression heads, which are strategically placed at each phase end to boost the overall performance. Comprehensive experiments on three benchmark datasets demonstrate that the proposed approach outperforms all the state-of-the-art models under the RMSE evaluation metric. The proposed approach also has better generalization capability with the best results during the cross-dataset experiments. First, we discuss the introduction and motivation, proposed network details, followed by experimental evaluation including the cross-dataset evaluation. Most content of this chapter comes from our paper [81].

3.1 Introduction

Multi-column or multi-regressor CNN based architectures [127, 85, 75, 94] have proved to be very effective for crowd counting tasks. MCNN [127] is a state-of-the-art three-column density-map estimation based end-to-end crowd counting network, where each CNN based column specializes in handling the specific crowd-density level. At the end of this network, all columns are merged together to yield the crowd estimate after the remaining processing. Similarly, multi-column based architectures [85, 58] utilize multiple specialized crowd count regressors to cope with multiple crowd-density scales separately. For example, Switch-CNN [85], a density-map estimation based network, consists of a CNN based switch classifier that routes the input image or patch to one of three crowd count regressors, where each regressor deals with a specific crowd level. In addition, many single-column or single-regressor based architectures [54, 86] have also been proposed to address the crowd counting issues and challenges. These methods produce promising results, but still lack the generalization ability for crowd estimation, ranging from low to high crowd-density.

Recently, Sajid et al. [82, 83] observed that suitable rescaling (down-, no-, or up-scaling) of the input image or patch, according to its crowd density level (low-, medium-, or high-crowd), gives more effective results as compared to the multi-column or multi-regressor based methods. Based on this observation, they also designed a patch rescaling module (PRM) [83] that rescales the input image or patch accordingly based on its crowd-density class label. Although the PRM based single-column proposed schemes [83] empirically prove their observation to be imperative and effective, the PRM module does not fully capitalize on it and thus limits the efficacy of this observation. First, it requires the crowd-density classification label of the original input patch. This additional classification process comes up with its own inaccuracies [82, 83] that compromises the subsequent crowd counting process. Second, the PRM module selects only one of three available recalling operations (down-, no-, or up-scaling) for any given input patch. This limits the overall effectiveness and improvement of the PRM module and only utilizes the deployed observation partially. Contrary to only using the single rescaling for the input patch, we empirically observed that using all three rescaled versions of the input patch with feature-level fusion or sharing gives much better performance. Consequently, it also eliminates the need for any crowd-density classification process for the original input patch. To this end, we aim to achieve the following two objectives in this work:

- Better generalization ability: Design a multi-column crowd counting method with better generalization ability towards huge crowd variations.

- Effective input priors: Utilize the input patch rescaling based effective observation [82, 83] (as discussed above) without performing any expensive and compromising crowd-density classification process, and also use all three crowd-density levels (low-, medium, and high-crowd) in a more effective manner than the PRM module [83].

Thus, we propose a new multi-resolution feature-level fusion based end-to-end crowd counting network to achieve the above objectives amid addressing the major crowd counting challenges. The proposed approach works at multiple scales via multi-columns, where each column primarily focuses on the respective scale (low-, medium-, or high-crowd), as shown in Fig. 3.1. Unlike other state-of-the-art multi-scale or multi-column based methods, the columns also fuse and share the information with each other at a regular basis after every few deep layers (phase). Each column also takes the suitably rescaled version of the original input patch as its input prior without any classification process. Inspired by the success of high-resolution networks [99, 105], each column also serves as a high-resolution sub-network, where the resolution is maintained the same as its input throughout the column. These repetitive multi-scale fusions, coupled with column-wise rescaled input priors and high-resolution maintenance, prove to be more effective in generalizing towards huge crowd variation issue (Objective # 1) in comparison to recent state-of-the-art crowd counting methods as shown in the experiments section 3.5. In addition, the simple yet effective column-wise input priors inclusion fulfills our objective # 2 without using any compromising and extra crowd-density classification process.

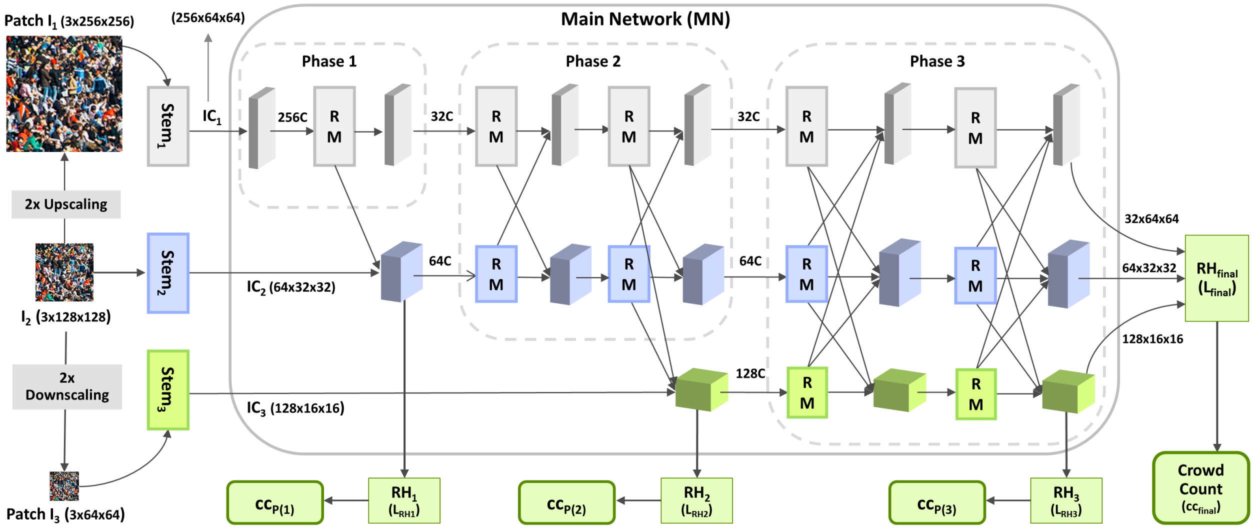


Figure 3.1: The proposed network. The original 128×128 input patch (I_2 or P) is used to produce the new up-scaled (I_1) and down-scaled (I_3) input priors, which go through their respective stems ($stem_1, stem_2, stem_3$). The resultant initial channels (IC_1, IC_2, IC_3) then pass through the phase-based main network, containing three deep columns/branches with the residual modules (RM). Multi-resolution fusion regularly occurs between these columns, followed by passing through the auxiliary (RH_1, RH_2, RH_3) and the final (RH_{final}) crowd regression heads to yield the respective crowd counts ($cc_{p(1)}, cc_{p(2)}, cc_{p(3)}$, and cc_{final}). The final crowd count for the input patch (I_2) is the weighted average of these crowd estimates. The MN maintains the channels (C) resolution throughout each column. (In this chapter, we used both terms (I_2 and P) interchangeably for the same original input patch. Similarly, multi-scale and multi-resolution fusion are interchangeable here.)

3.2 Our Contributions

The contributions of this chapter mainly include:

- We propose a new multi-resolution feature-level fusion based end-to-end crowd counting approach for still images that effectively deals with significant variations of crowd-density, lighting conditions, and large perspective.
- We propose an alternative patch rescaling module by more effectively using the input priors. Unlike the PRM [83], the proposed module fully utilizes all three crowd-density levels without requiring any compromising or additional crowd-density classification process.
- Quantitative experiments demonstrate that the proposed approach outperforms the state-of-the-art methods, including the PRM based schemes, by a large margin with up to 10% improvements.

3.3 Proposed Approach

The chapter proposes a multi-column and multi-resolution fusion based end-to-end crowd counting network to achieve the two set objectives in Sec. 3.1, amid addressing the major crowd counting challenges including huge crowd variation in and across different images, large perspective, and severe occlusions. The proposed scheme is shown in Fig. 3.1, where the input image is first divided into 128×128 non-overlapping patches. Each resultant patch then goes through the proposed network for the patch-wise crowd count. Finally, the image crowd estimate is computed by the sum of the crowd count of all patches. The 128×128 input patch is used to generate the new 256×256 and 64×64 size input priors by $2\times$ times up- and down-scaling, respectively. These multi-scale input priors pass through the respective stems ($Stem_1, Stem_2, Stem_3$) to generate three separate initial channels (IC_1, IC_2, IC_3), which act as the corresponding input to three columns/branches in the main network (MN). The MN regularly fuses feature maps in between these branches. At the

Name	Output size	Filters (F) Operation
$Stem_1$		
I_1	$3 \times 256 \times 256$	
	$64 \times 128 \times 128$	(3×3) conv, stride 2, padding 1, 64F
	$64 \times 64 \times 64$	(3×3) conv, stride 2, padding 1, 64F
IC_1	$256 \times 64 \times 64$	(1×1) conv, stride 1, padding 0, 256F
$Stem_2$		
I_2	$3 \times 128 \times 128$	
	$64 \times 64 \times 64$	(3×3) conv, stride 2, padding 1, 64F
IC_2	$64 \times 32 \times 32$	(3×3) conv, stride 2, padding 1, 64F
$Stem_3$		
I_3	$3 \times 64 \times 64$	
	$64 \times 32 \times 32$	(3×3) conv, stride 2, padding 1, 64F
IC_3	$128 \times 16 \times 16$	(3×3) conv, stride 2, padding 1, 128F

Table 3.1: Configurations of the Stems. Each conv operation denotes the Convolution-BN-ReLU series.

end of the main network, the resultant feature maps from three branches pass through the final regression head (RH_{final}) to yield the input patch crowd estimate. The MN also outputs into three auxiliary crowd estimating regression heads (RH_1, RH_2, RH_3) that helps in improving the input patch final crowd count. In the following, we will discuss three main components in detail.

3.3.1 Input priors and respective stems

We up- and down-scale the original 128×128 size input patch (I_2 or P) by $2\times$ to generate its rescaled versions (256×256 and 64×64 respectively). These input priors (I_1, I_2, I_3) pass through their respective stems ($Stem_1$, $Stem_2$, $Stem_3$) to produce initial feature channels (IC_1, IC_2, IC_3). These stems, as shown in Table 3.1, also decrease the input priors resolution to $1/4$, and the resultant initial feature maps resolution becomes half in the subsequent lower column. The upscaled input prior (I_1) helps in handling highly dense crowd regions by zooming in and observing the original input (I_2) in detail to avoid huge crowd under-estimation. Similarly, the input prior (I_3) uses a smaller scale, especially helpful for the low-crowd regions in the images that may otherwise cause significant crowd over-estimation. Empirically, it has been observed that coupling these simple yet

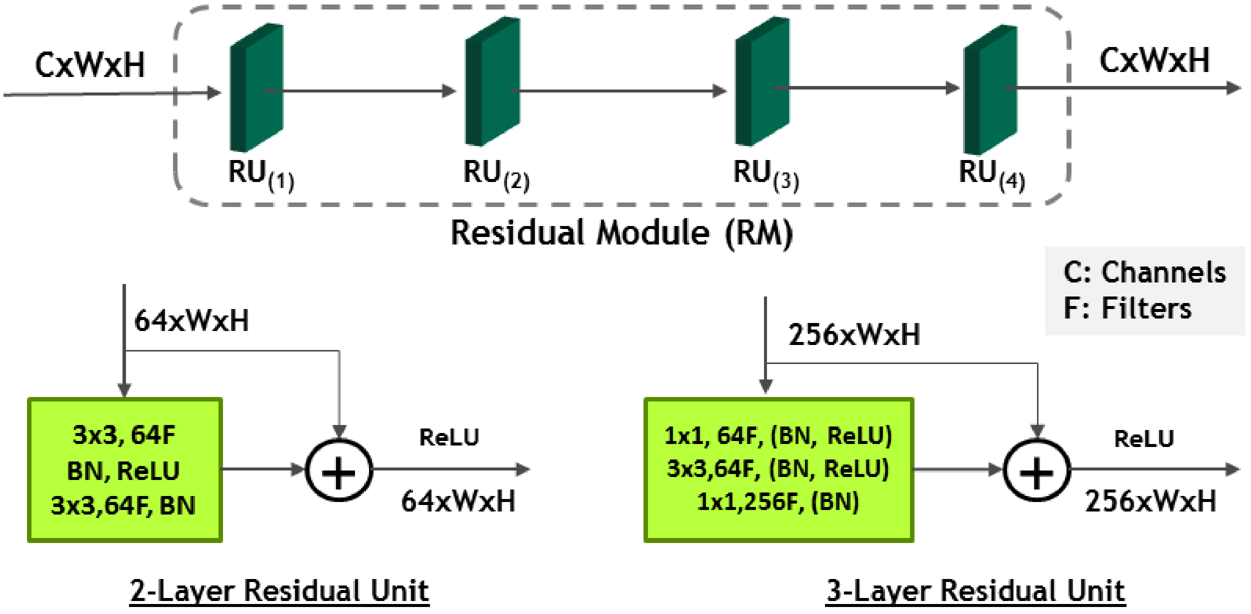


Figure 3.2: The Residual Module (RM) consists of either only 2- or 3-layers [26] based four residual units (RU).

effective rescaled input priors (I_1, I_3) with the original input (I_2) yields better crowd estimates, and consequently avoid huge crowd under- or over-estimation, as shown in the ablation study in Sec. 3.5.5.

3.3.2 Main Network (MN)

The main network is composed of three deep columns/branches, each with its own input prior feature maps, and also caters the respective crowd-density scale. The main network is divided into three phases from left to right, where each phase consists of one or more columns/branches. The total number of columns in a phase is equal to its phase number. All branches in a phase fuse feature maps with each other after each Residual Module (RM). At the end of each phase, the MN also feeds its lowest-resolution output into the auxiliary crowd regression heads (RH_1, RH_2, RH_3), as detailed in the next subsection 3.3.3. Each branch in the main network maintains its original input resolution throughout the branch, unlike other state-of-the-art multi-scale crowd estimation methods. The lower columns resolution and total channels in any phase depend on the highest-resolution branch ($i = 1$). Let C_1 and R_1 be the total channels and their resolution respectively in the highest-resolution column. Then, the remaining columns ($i = 2,3$) follow the below principle for their C_i and R_i in a given phase [99, 105].

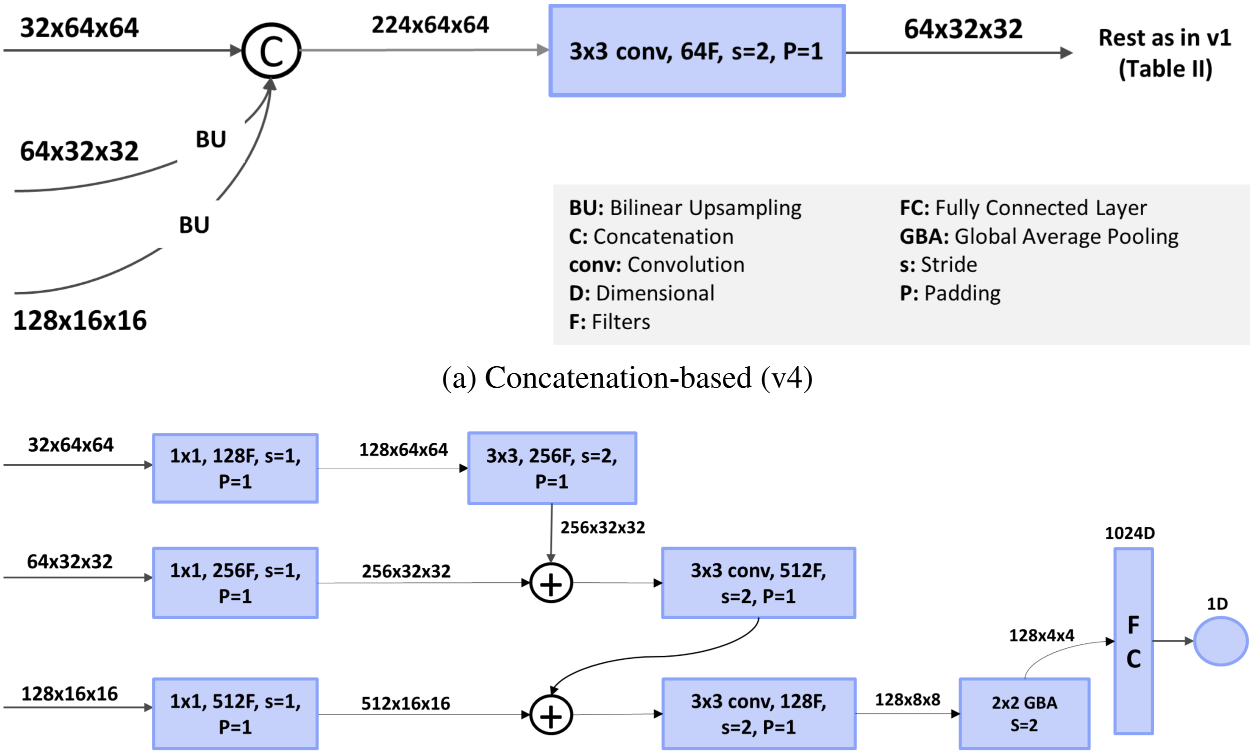
$$C_i = 2C_{i-1}, R_i = \frac{R_{i-1}}{2} \tag{3.1}$$

Residual Module: It consists of four residual units, where each unit is formed by either only 2-layer or 3-layer based residual block [26], as shown in Fig. 3.2. The 2-layer based residual block [26] contains two 3×3 convolution layers. Similarly, the 3-layer residual block [26] starts with a bottleneck layer, followed by one 3×3 convolution layer and a bottleneck layer. Each convolution operation in these units is followed by the batch Normalization (BN) [36] and the nonlinear ReLU [69] activation. Phase-1 uses the 2-layer based residual unit, whereas Phase-2 and 3 deploy the 3-layer based residual unit. The number of residual modules in each column per phase serves as a hyperparameter and discussed in ablation study in Sec. 3.5.4. Moreover, by the network design, total residual modules in each column of a specific phase remain the same.

Recurring Multi-resolution Fusions: The primary purpose of the multi-resolution fusion is to exchange the information between different resolutions/columns, so as to enhance the generalization ability of the proposed scheme towards huge crowd diversity in and across different images. We utilize one or more 3×3 convolution operations to fuse higher-resolution feature maps into the lower-level channels. To fuse the lower-resolution feature maps into the higher-level channels, bilinear upsampling followed by the bottleneck layer (to adjust the number of channels) have been deployed. Let Ch_i be the fusion source channels from column at ith index ($i = 1,2$ or 3), Ch_j be the fusion target column at index j ($j = 1,2$ or 3), and $f(.)$ be the transformation function. If $i < j$, then $f(Ch_i)$ downsamples the Ch_i channels by $2(j-i)$ times via $(j-1)$ stride-2 3×3 convolution(s). For example, fusing column-1 channels (Ch_1) into column-2 channels (Ch_2) first requires one stride-2 3×3 convolution ($f(Ch_1)$) for $2\times$ downsampling. Similarly, Ch_1 fusion into Ch_3 requires 2 stride-2 3×3 convolutions for $4\times$ downsampling before the fusion operation. If $i = j$, then $f(Ch_i) = Ch_i$, i.e., no transformation is done. If $i > j$, then $f(Ch_i)$ transformation upscales the Ch_i using the bilinear upsampling, followed by the bottleneck layer to adjust the number of channels accordingly before the summation-based fusion process. Each convolution operation is followed by the Batch Normalization (BN) [36] and the nonlinear ReLU activation [69].

Output Size	Filters (F) Operation
v1 (Highest-resolution)	
$32 \times 64 \times 64$	
$64 \times 32 \times 32$	(3×3) conv, stride 2, padding 1, 64F
$64 \times 16 \times 16$	(3×3) conv, stride 2, padding 1, 64F
$64 \times 8 \times 8$	(2×2) Avg Pooling, stride 2
1024D, FC	-
1D, FC (single neuron)	-
v2 (Middle-column)	
$64 \times 32 \times 32$	
$64 \times 32 \times 32$	(1×1) conv, stride 1, padding 0, 64F
	Rest continues as in v1 above
v3 (Lowest-resolution)	
$128 \times 16 \times 16$	
$64 \times 16 \times 16$	(1×1) conv, stride 1, padding 0, 64F
	Rest continues as in v1 above
RH_1 Configuration	
$64 \times 32 \times 32$	
$64 \times 16 \times 16$	(3×3) conv, stride 2, padding 1, 64F
	Rest continues as in v1 above
RH_2 and RH_3 Configuration	
$128 \times 16 \times 16$	
$64 \times 8 \times 8$	(3×3) conv, stride 2, padding 1, 64F
	Rest continues as in v1 above

Table 3.2: Standalone single-column output based RH_{final} head versions (v1, v2, v3) and Auxiliary Crowd Regression Heads (RH_1,RH_2,$RH3$) configurations. Each conv operation denotes the Convolution-BN-ReLU sequence. These configurations mainly consist of several conv layers followed by the global average pooling and one or more fully connected (FC) layers to finally yield the crowd estimate (single neuron).



(a) Concatenation-based (v4)

(b) Summation-based (v5)

Figure 3.3: Concatenation-based crowd regression head (v4) concatenates the lower-resolutions with the highest-level channels using the bilinear upsampling, whereas the summation-based head (v5) adds the higher-level channels into the lowest-resolution feature maps, before proceeding through the several deep layers to finally yield the crowd estimate (cc_{final}) [99, 105].

3.3.3 Crowd Regression Heads

The proposed approach contains three phase-wise crowd regression heads (RH_1, RH_2, RH_3) and the final regression head (RH_{final}).

Phase-wise Regression Heads: One of the primary purpose of phase based organization of the main network is to introduce auxiliary crowd regression heads (RH_1, RH_2, RH_3) at the end of each phase. The last lowest-resolution output of each phase serves as the input to its respective regression head. These heads mainly consist of several convolution based deep layers, followed by optional average pooling operation and one or more fully connected (FC) layers as detailed in Table 3.2. Finally, the single neuron ($1D, FC$) at the end of each head gives the corresponding crowd counts ($cc_{P(1)}, cc_{P(2)}, cc_{P(3)}$) for the input patch ($P$).

Final Regression Head (RH_{final}): Phase-3 outputs three blocks of feature maps, each from the respective column with varying resolution. These blocks have been exploited in different ways for possible and effective RH_{final} head configuration, as discussed below.

Standalone Single-Column Output based (v1,v2,v3). Here, we only use one of three phase-3 outputs for the RH_{final} configuration [99, 105]. Subsequent configurations are shown in Table 3.2, and named as v1 (highest-resolution), v2 (middle-column), and v3 (lowest-resolution), respectively. These representations consist of several deep layers, followed by the 1024 dimensional fully connected (FC) layer and the final single neuron to directly regress the crowd count.

Concatenation-based (v4). The lower-resolution feature maps concatenate at the highest-resolution branch, with configuration shown in Fig. 3.3(a) [99, 105].

Summation-based (v5). The higher-level feature maps are summed up into the subsequent lower resolution feature maps after respective downscaling, as shown in Fig. 3.3(b) [99, 105].

Employing one of the above configurations, the RH_{final} yields its crowd count (cc_{final}) for the input patch P. The final crowd count (CC_P) for the original input patch P is computed using all regression heads weighted crowd estimates as follows:

$$CC_P = w * cc_{P(1)} + x * cc_{P(2)} + y * cc_{P(3)} + z * cc_{final} \tag{3.2}$$

Where $w = x = y = 0.1$ and $z = 0.7$. The mean squared error (MSE) has been used as the loss function for each of the four regression heads (RH), given as follows:

$$L_{RH} = \frac{1}{N} \sum_{i=1}^{N} (F(x_i, \Theta) - y_i)^2 \tag{3.3}$$

where N represents the total training patches per batch, y_i denotes the ground truth crowd count for the input image patch x_i, and $F(.)$ represents the transformation function that learns the x_i to crowd count mapping with learnable weights Θ. Finally, the total loss for the input patch P is the weighted accumulation of all four regression head losses as below:

$$L_P = w * L_{RH_1} + x * L_{RH_2} + y * L_{RH_3} + z * L_{final} \tag{3.4}$$

3.4 Implementation Details

We employ the following two standard metrics, namely Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), for the evaluation and comparison of the proposed scheme with other state-of-the-art methods.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|CC_t - \hat{CC_t}|, RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(CC_t - \hat{CC_t})^2} \tag{3.5}$$

where T represents the total test images in a dataset, and CC_t and $\hat{CC_t}$ denote the actual and estimated crowd counts respectively for the test image t.

Training Details: We randomly extract $60,000$ patches of 256×256, 128×128, and 64×64 sizes with varying crowd number from the training images. Horizontal flip based data augmentation is then used to double the training samples quantity. We trained the proposed model for 100 epochs, used SGD optimizer with a weight decay of 0.0001 and a Nesterov momentum value of 0.9. Multi-step learning has been employed that initially starts at 0.001 and decreases by half after every 25 epochs. As per the standard literature convention, 10% data from the predefined training set has been separated for the model validation purpose.

3.5 Experimental results

In this section, we first perform standard quantitative analysis on three benchmark datasets: UCF-QNRF [35], ShanghaiTech [127], and AHU-crowd [32]. These benchmarks pose a great collective challenge for the proposed scheme to prove its effectiveness, as they vary significantly with each other in terms of average image resolution, average crowd number per image, total images, and lighting conditions. Next, we discuss the ablation experiments findings and the cross-dataset evaluation, followed by the qualitative evaluation. For comparison with other state-of-the-art methods,

	ShanghaiTech		UCF-QNRF	
Method	MAE	RMSE	MAE	RMSE
MCNN [127]	110.2	173.2	277	426
CMTL [94]	101.3	152.4	252	514
Switch-CNN [85]	90.4	135.0	228	445
SaCNN [126]	86.8	139.2	-	-
IG-CNN [3]	72.5	118.2	-	-
ACSCP [87]	75.7	102.7	-	-
CSRNet [54]	68.2	115.0	-	-
CL[35]	-	-	132	191
CFF [91]	65.2	109.4	93.8	146.5
RRSP [103]	63.1	96.2	-	-
CAN [59]	**62.3**	100.0	107	183
L2SM [115]	64.2	98.4	104.7	173.6
BL [65]	62.8	101.8	**88.7**	154.8
ZoomCount [82]	66.6	94.5	128	201
PRM-based[83]	67.8	86.2	94.5	141.9
v1/v2 (**ours**)	71.4/70.1	85.7/85.3	103.1/100.6	139.6/136.3
v3/v4 (**ours**)	69.8/67.9	84.7/81.9	101.7/98.4	137/135.1
v5 (**ours**)	67.1	**81.0**	96.9	**130.1**

Table 3.3: Experiments on ShanghaiTech [127] and UCF-QNRF [35] benchmarks. The proposed method (v5) outperforms the state-of-the-art methods (including the PRM based approach [83]) for the RMSE metric, while giving comparable results for the MAE metric.

we evaluate all five versions of the proposed method ($v1, v2, v3, v4, v5$) as discussed in Sec. 3.3.3.

3.5.1 Experiments on UCF-QNRF Dataset

UCF-QNRF [35] is one of the most diverse, realistic, and challenging dataset. It consists of $1,535$ free-view images with a predefined train/test division of $1,201/334$. It contains images with relatively very small (300×377) and very large (6666×9999) resolutions, with $1,251,642$ total people annotations that show its crowd complexity and diversity. We compare the proposed approach with the state-of-the-art methods (including the PRM based approach [83]) in Table 3.3. As shown, the proposed scheme (v5) outperforms the state-of-the-arts under the RMSE evaluation metric by $\sim$ 8.3% (from 141.9 to 130.1), amid performing reasonably well for the MAE.

Method	MAE	RMSE
Haar Wavelet [71]	409.0	-
DPM [20]	395.4	-
BOW–SVM [18]	218.8	-
Ridge Regression [12]	207.4	-
Hu et al. [32]	137	-
DSRM [120]	81	129
ZoomCount [82]	74.9	111
CC-2P (PRM-based)[83]	66.6	101.9
v1/v2/v3 (**ours**)	69.8/67.1/65.4	107.8/103.5/100.2
v4/**v5** (**ours**)	63.1/**60.2**	99.5/**91.7**

Table 3.4: AHU-Crowd dataset experiments.

3.5.2 Experiments on ShanghiTech Dataset

The ShanghaiTech Part-A benchmark [127] is another diverse and free-view crowd counting benchmark. It contains 482 images (predefined train/test division of $300/182$) with a total of $241,677$ people annotations and average image resolution of 589×868. Based on the quantitative comparison with the state-of-the-art methods (including the PRM based methods [83]) as shown in Table 3.3, the proposed approach (v5) decreases the RMSE error by $\sim 6\%$ (from 86.2 to 81.0). For the MAE metric, our schemes give reasonable and comparable results. The lowest RMSE value also demonstrates that our method is less susceptible to huge crowd over- and under-estimation.

3.5.3 Experiments on AHU-Crowd Dataset

The AHU-Crowd [32] dataset contains 107 crowd images with 58 to $2,201$ people annotations per image and $45,807$ annotations in total. As per the standard evaluation process, we perform 5-fold cross-validation, and final (MAE, RMSE) results are obtained by computing their average. Evaluation and comparison results are shown in Table 3.4, where our scheme (v5) outperforms other state-of-the-arts under both evaluation metrics with significant improvements i.e., the MAE error decreases by $\sim 9.6\%$ (from 66.6 to 60.2) and the RMSE improves by $\sim 10\%$ (from 101.9 to 91.7).

RM Modules Quantity Effect		
RM Modules per column in Phase-2 and 3	MAE	RMSE
1	79.3	111.4
2 (our default)	**67.1**	**81.0**
3	75.8	104.7

Input Priors Effect		
	MAE	RMSE
w/o (I_2, I_3)	77.1	108.8
w/o (I_3)	75.9	106.5
w/o (I_2)	73.8	101.4
only (I_1) with original input size (256×256)	80.1	124.5
with (I_1, I_2, I_3) (our default)	**67.1**	**81.0**

Auxiliary Regression Heads Effect		
	MAE	RMSE
w/o RH_1	76.2	107.0
w/o RH_2	71.7	115.2
w/o RH_3	73.9	103.1
w/o (RH_1, RH_2, RH_3)	78.5	120.7
with (RH_1, RH_2, RH_3) (our default)	**67.1**	**81.0**

Table 3.5: Three respective ablation studies on the effect of RM modules quantity per column in Phase-2 and 3, Input priors (I_1, I_2, I_3), and auxiliary crowd regression heads usage on the proposed network performance. The results demonstrate the fact that the priors and auxiliary heads are of vital importance, as the MAE and RMSE errors increase without them. These ablation experiments are done using the ShanghaiTech dataset, and performed on the proposed method version v5, as being the best of them quantitatively.

3.5.4 Effect of RM Modules Quantity

In this ablation study, we examine the effect of the number of RM modules in the Phase-2 and 3 of the proposed scheme. Instead of using 2 RM modules by default, we evaluate our method (v5) separately by utilizing only either 1 or 3 RM modules per column in each phase. As shown in Table 3.5 on the ShanghaiTech [127] dataset, our default choice of 2 RM modules per column in both phases (Phase-2 and 3) yields the most effective results. Using 1 or 3 RM modules per column in each phase cause the MAE, RMSE errors increase of $(15.4\%, 27.3\%)$ and $(11.5\%, 22.6\%)$ respectively. Thus, we have employed 2 RM modules per column in Phase-2 and 3.

Method	MAE	RMSE
Cascaded-mtl [94]	308	478
Switch-CNN [85]	301	457
CC-2P (PRM based) [83]	219	305
v1/v2/v3 (**ours**)	214/217/212	301/303/294
v4/**v5** (**ours**)	206/**201**	285/**278**

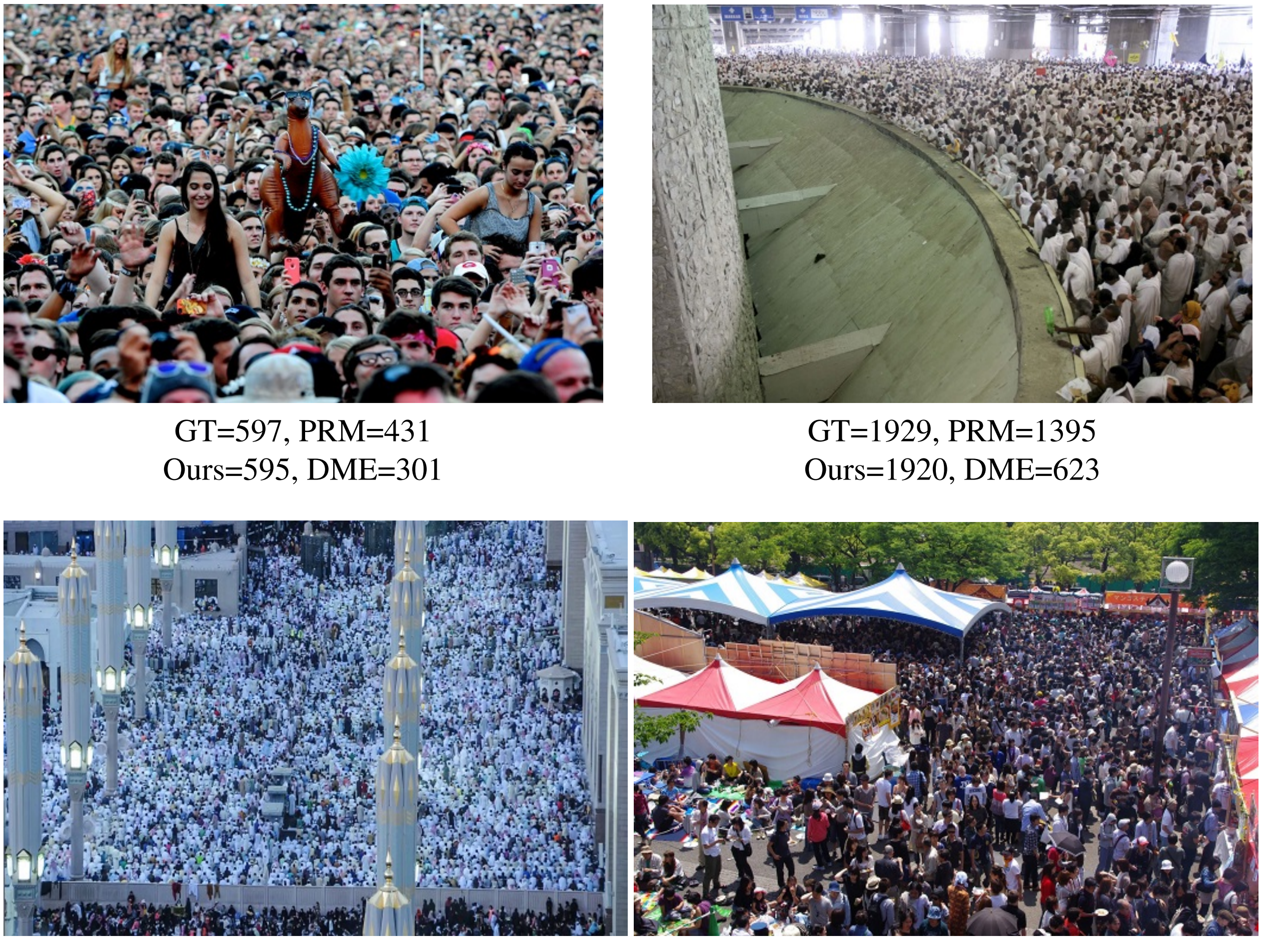
Table 3.6: Cross-dataset experiments.

3.5.5 Effect of Input Priors (I_1, I_2, I_3)

This section reveals the quantitative importance of the input priors. We remove these input priors in different experimental settings to analyze their effectiveness. In the first three separate experiments, we only use (I_1), (I_1, I_2) and (I_1, I_3) input prior(s) respectively. While, in the last setting, we only deployed the I_1 input, but with the original 256×256 input size without any rescaling. The consequent ablation results are shown in Table 3.5, from which we can see that removing these input priors significantly decreases the overall network performance (with minimum MAE, RMSE errors increase of $9.1\%, 20.1\%$ respectively). Thus, all three input priors are critical for the proposed method effectiveness.

3.5.6 Effect of Auxiliary Crowd Regression Heads

In this experiment, we analyze the quantitative effect of employing the auxiliary crowd regression heads (RH_1, RH_2, RH_3) in the proposed scheme. During this ablation study, we removed each auxiliary head one by one and evaluate the network (v5) on the ShanghaiTech Part-A [127] dataset. As shown in Table 3.5, the performance decreases significantly after removing these heads (RH_1, RH_2, RH_3). For instance, without using the RH_1 head, the MAE error increases the most with a jump of 11.9%. Similarly, the RMSE error is being affected the most by the RH_2 head removal with a 29.7% increase in error.



GT=597, PRM=431
Ours=595, DME=301

GT=1929, PRM=1395
Ours=1920, DME=623

GT=3653, PRM=2792
Ours=3639, DME=2792

GT=1070, PRM=1011
Ours=1072, DME=722

Figure 3.4: Ground truth (GT) based qualitative comparison.

3.5.7 Cross-Dataset Evaluation

During the cross-dataset experiment, all methods have been trained and tested on the ShanghaiTech Part-A [127] and the UCF-QNRF [35] datasets respectively. As shown in Table 3.6, the proposed method demonstrates better generalization capability as compared to the state-of-the-art methods (including the PRM-based scheme [83]) with MAE, RMSE errors decrease by 8.2% (from 219 to 201) and 8.9% (from 305 to 278) respectively. Similar to the previous experiments, the proposed approach version (v5) appears to be the most effective cross-dataset validation scheme with the lowest MAE, RMSE values.

3.5.8 Qualitative Evaluation

In this section, we demonstrate some qualitative results as shown in Fig. 3.4. We also compare our scheme with the PRM-based [83] and density-map estimation (DME) [35] based recent state-of-the-art methods. In comparison, it can be observed that the proposed scheme yields the best performance of all on these actual test images with hugely varying crowd-density, lighting condition, and image resolution.

3.6 Conclusion

To address the major crowd count challenges, we proposed a new multi-resolution fusion based end-to-end crowd counting network for the still images in this work. We also deployed a new and effective PRM substitute that uses three input priors, and proves to be much more accurate than the PRM. Both quantitative and qualitative results have revealed that the proposed network outperforms the state-of-the-art approaches under the RMSE evaluation metric. Cross-dataset evaluation also demonstrates better generalization capability of our approach towards new datasets.

Chapter 4

PRM-based Crowd Counting via the Multi-scale Fusion and Attention Network (Uni-Modal)

This chapter deals with the second multi-scale and uni-modal crowd estimation network. This work focuses on improving the recent plug-and-play patch rescaling module (PRM) based approaches for crowd counting. In order to make full use of the PRM potential and obtain more reliable and accurate results for challenging images with crowd-variation, large perspective, extreme occlusions, and cluttered background regions, we propose a new PRM based multi-resolution and multi-task crowd counting network by exploiting the PRM module with more effectiveness and potency. The proposed model consists of three deep layered branches with each branch generating feature-maps of different resolutions. These branches perform a feature-level fusion across each other to build the vital collective knowledge to be used for the final crowd estimate. Additionally, early-stage feature-maps undergo visual attention to strengthen later-stage channel's understanding of the foreground regions. The integration of these deep branches with the PRM module and the early-attended blocks proves to be more effective than the original PRM based schemes through extensive numerical and visual evaluations on three benchmark datasets. The proposed approach yields a significant improvement by a margin of 12.6% in terms of RMSE evaluation criterion. It also outperforms state-of-the-art methods in cross-dataset evaluations. Most content of this chapter comes from our paper [84].

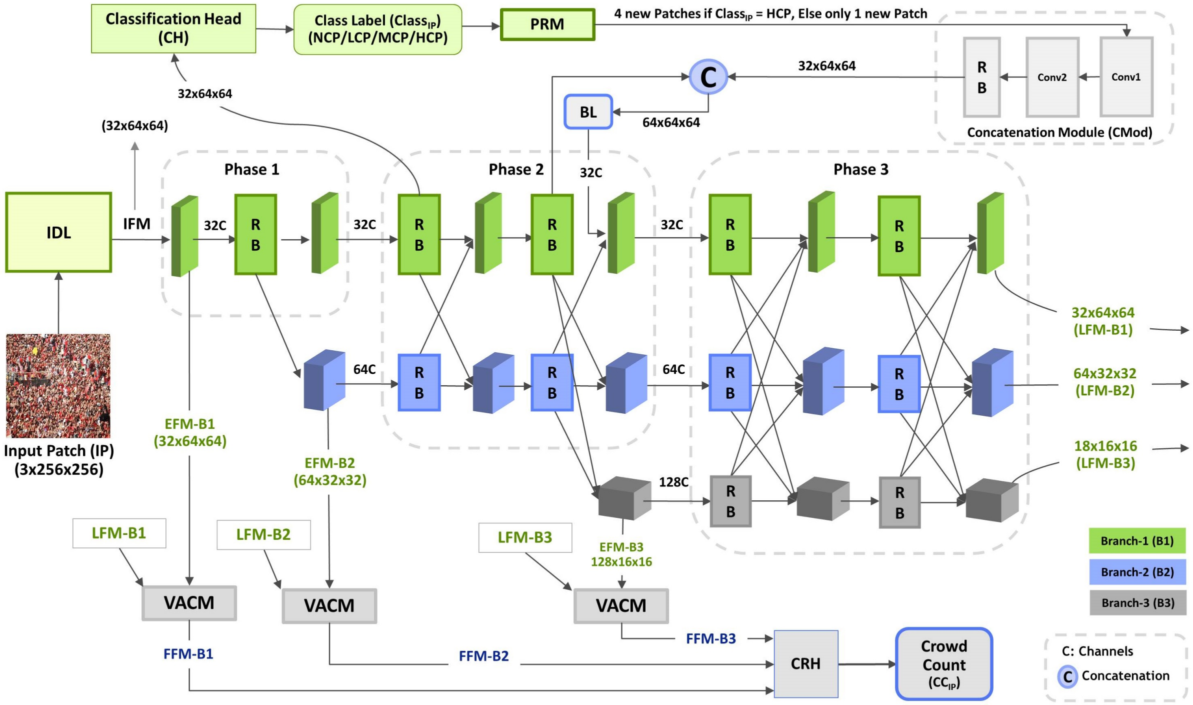


Figure 4.1: The proposed multi-task feature-level fusion based crowd counting network. The 256×256 input patch (IP) passes through the initial deep layers (IDL) to output the initial feature-maps (IFM) that are fed into the phase-based multi-branch network. Each of the three branches/-columns maintains its original resolution throughout that branch [99, 105]. The output channels from the first residual block (RB) of the (Branch-1, Phase-2) are branched-out to make the classification head (CH). The CH classifies the patch (IP) 4-way according to its crowd-density level. This labeling process is being used by the PRM module [83] to decide if appropriate rescaling and generation of new patches are required or not. The resultant patch(es) then passes through the concatenation module ($CMod$) to generate the channels (C) that are concatenated back into the network after adjusting the number of channels via the Bottleneck layer (BL). The network branches also do the feature-level fusion regularly to form the model collective knowledge. After the remaining processing, the Phase-3 outputs three later-stage feature-maps (LFM-B1, LFM-B2, LFM-B3). Each LFM passes through the VACM module along with their respective branch early-stage feature-maps (EFM-B1, EFM-B2, EFM-B3). The VACM generated final feature-maps (FFMs) are eventually inserted into the concatenation-based Crowd Regression Head (CRH) to obtain the final crowd count (CC_{IP}) for the input Patch (IP).

4.1 Introduction

Recently, Sajid *et al.* [83] proposed a counting-by-regression based method that uses a lightweight Patch Rescaling Module (PRM) to rescale the input image or patch accordingly based on its crowd-density level before the crowd estimation. They also proposed PRM-based schemes, which performed reasonably better as compared to other state-of-the-art methods. Although these schemes produce state-of-the-art results, they carry the following key shortcomings:

- They maintain a single branch/column architecture with only-one scale focus that limits their achievable performance and potential effectiveness.
- These models only utilize the PRM based input-level multi-resolution rescaling process, while lacking the beneficial feature-level multi-resolution process.

In this work, we present a new multi-resolution feature-level fusion based multi-task and visually-attended crowd counting method that aims to address the major crowd counting challenges as well as further explores and pushes the boundaries of the PRM module by tackling its shortcomings as mentioned above. The proposed PRM based scheme, as shown in Fig. 4.1, incorporates multiple columns or branches, each with feature-maps of different resolutions. Based on the high-resolution networks [99, 105], these branches perform the fusion or sharing across each other on a regular basis to form a collective knowledge that improves the overall network performance. We also perform the visual-attention process on the early-stage feature-maps from each of these branches to boost later-stage channel's understanding of foreground and background. On the other hand, the PRM based rescaling of the input image or patch has been used to select the appropriate input-level scale based on its crowd-density level. The experiments (Sec. 4.5) demonstrate the proposed PRM utilization to be remarkably more effective as compared to its previous implementations [83, 82].

4.2 Our Contributions

The key contributions of this chapter are listed as follows:

- We propose a new multi-resolution and multi-task PRM based visually-attended crowd counting network that effectively addresses major crowd counting challenges, including the issues of crowd-like background regions and huge crowd-variation.
- We deploy the plug-and-play PRM module so as to further push its boundaries and utilize it more effectively as compared to its previous deployments.
- We employ the visual attention mechanism in a unique and effective way on early-stage feature-maps that facilitates the later-stage channels to better understand the foreground regions.
- The proposed scheme outperforms the current-best methods in most cases, including the original PRM based schemes as demonstrated via both numerical and visual experiments on three benchmark datasets. The proposed model shows an improvement of up to 12.6% in terms of RMSE evaluation while performing almost equally best for the MAE evaluation criterion. During the cross-dataset evaluation, the proposed method outperforms the state-of-the-art.

4.3 Proposed Approach

This work aims to address the major crowd counting challenges (e.g., huge difference in image and scale resolution, severe occlusions, far-reaching perspective changes, etc.) as well as to exploit the PRM module more effectively while mitigating its shortcomings. The proposed multi-task network, as shown in Fig. 4.1, contains three deep-layered branches with different resolutions feature-maps (channels). The feature-level fusion [99, 105] occurs in between these branches on

Layer	Output	Filters (F)
Initial Deep Layers (IDL)		
IP	3 × 256 × 256	
	64 × 128 × 128	(3 × 3) conv, s = 2, p = 1, 64F
	64 × 64 × 64	(3 × 3) conv, s = 2, p = 1, 64F
IFM	32 × 64 × 64	(1 × 1) conv, s = 1, p = 0, 32F
Classification Head (CH)		
	32 × 64 × 64	
	64 × 32 × 32	(3 × 3) conv, s = 2, p = 1, 64F
	32 × 16 × 16	(3 × 3) conv, s = 2, p = 1, 32F
	32 × 8 × 8	(2 × 2) Global Avg. Pooling, s = 2
	1024D FC	-
	4D FC, Softmax	-
Concatenation Module (CMod)		
	3 × 256 × 256	
Conv1	64 × 128 × 128	(3 × 3) conv, s = 2, p = 1, 64F
Conv2	32 × 64 × 64	(3 × 3) conv, s = 2, p = 1, 32F
RB	32 × 64 × 64	-
Crowd Regression Head (CRH) (Continued from Fig. 3(b))		
	64 × 32 × 32	
	64 × 16 × 16	(3 × 3) conv, s = 2, p = 1, 64F
	64 × 8 × 8	(2 × 2) Avg. Pooling, s = 2
	1024D, FC	-
	1D, FC	-

Table 4.1: Configurations of IDL, CH, CMod, and CRH modules. Each *conv* operation denotes the Convolution-BN-ReLU sequence [33]. (s: stride, p: padding, FC: Fully Connected)

a regular basis, which helps them to form a collective knowledge about the input image or patch. As their final output, each of these branches produces early- (EFM) and later-stage (LFM) feature-maps. The EFMs undergo visual attention process before concatenation with the LFM channels via the visual attention and concatenation (VACM) module. The visually-attended EFMs enable the LFMs to clearly distinguish between foreground and background image regions. The concatenated final feature-maps (FFMs) then pass through the crowd regression head (CRH) to output the final crowd estimate. The PRM module has been deployed to rescale the input to its appropriate scale. To start with, we first divide the input image into fixed-size 256 × 256 non-overlapping patches. Following that, we estimate the crowd count for each resultant patch separately using the proposed scheme. Finally, the image total people count is equal to the summation of all its patches crowd estimates.

The input patch (IP) first proceeds through the initial deep layers (IDL), as detailed in Table 4.1, to produce the initial feature maps (IFMs). The IDL also reduces the input resolution to $1/4$ (from 256×256 to 64×64). Next, these feature-maps are routed to Phase-1 of the multi-branch network. The phase-wise multi-branch network comprises of five key components: Multi-resolution branches and phases, Multi-resolution fusion, PRM module deployment, VACM module, and Crowd Regression Head (CRH). They are detailed as follows.

4.3.1 Multi-resolution Branches and Phases

The network comprises of three phases (Phase-1, 2, and 3). The initially generated feature-maps (IFMs) or channels pass through the phase-wise organized multi-branch deep layers, starting from Phase-1. Each phase contains the total number of multi-resolution branches equal to its phase number. Consequently, Phase-1, 2, and 3 contain one, two, and three deep branches, respectively. Each branch also maintains its channels resolution throughout that branch [99, 105]. These branches also perform feature-level fusion across each other on a regular basis to form a collective knowledge-based learning process, as detailed in the next subsection 4.3.2. The channel resolution and the total number of channels in a specific branch depend on the highest-resolution branch configuration (Branch-1). The channel resolution decreases by half in each subsequent lower-resolution branch. However, the total number of channels increases $2\times$ times as we move from higher to lower-resolution branches. Thus, Branch-1, 2, 3 contain $(32 \times 64 \times 64)$, $(64 \times 32 \times 32)$, $(128 \times 16 \times 16)$ channels respectively, where it is denoted by $(Channels \times Width \times Height)$.

Each phase also contains the Residual Blocks (RB). Each RB comprises of four residual units [26] that are 3-layer based residual modules as shown in Fig. 4.2. Batch Normalization (BN) [36] and the non-linear ReLU activation [69] follow each convolution operation. The number of such RB modules remains the same in each branch of a specific phase.

Eventually, the phase-based three-branch structure outputs six separate blocks of feature-maps;

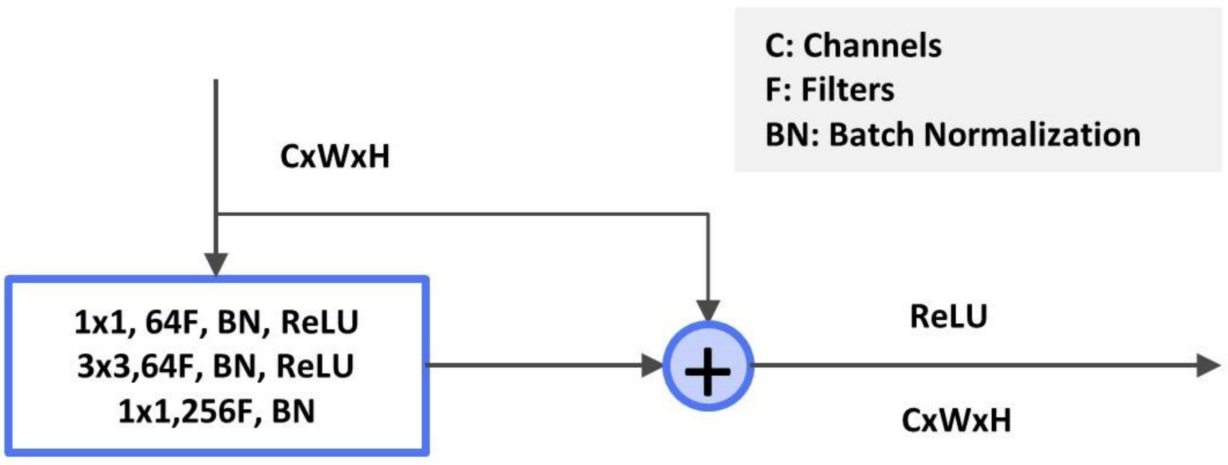


Figure 4.2: 3-Layered Residual Unit [26] being used in the Residual Block (RB). Each RB is composed of four such units.

three early- (EFM) and three later-stage (LFM) feature-maps. The phase-3 output blocks from the respective branch Bx (x=1,2, or 3) serve as the LFM-Bx channels. To obtain the early-stage EFM-Bx channels, we take the very first channels being produced in that specific branch Bx. These channels proceed forward for further processing.

4.3.2 Repeating Multi-resolution Fusion

The branches in a phase regularly share their channels across each other via the summation-based fusion. This sharing process helps in learning and building the collective information and knowledge from all branches that naturally enhances the generalization potential of the proposed network towards huge crowd-density and scale variation. The higher-branch channels are fused into the lower-resolution channels using the $(3 \times 3, stride-2, padding-1)$ convolution(s) to down-size the resolution accordingly [99, 105]. To fuse the Branch-1 channels into the Branch-2, Branch-1 channels are down-sampled by using one such 3×3 convolution operation. Similarly, the Branch-1 fusion into Branch-3 requires this convolution operation twice. To fuse the lower-resolution channels into the higher-level branch, bilinear upsampling has been applied to lower-resolution features to up-size them accordingly before the fusion process [99, 105].

4.3.3 PRM Module Deployment

The purpose of the Patch Rescaling Module (PRM) [83] is to rescale the input patch (*IP*) based on its crowd-density level. As defined in [83], we first require the 4-way crowd-density classification (No-Crowd (NCP), low-Crowd (LCP), Medium-Crowd (MCP), High-Crowd Patch (HCP)) for input (*IP*) before using the PRM module. Thus, we branch-out the output channels from the first RB module of (Branch-1, Phase-2). These channels then proceed through the crowd-density Classification Head (*CH*) that performs the required 4-way classification (NCP, LCP, MCP, HCP) [83]. Depending on the designated class label ($Class_{IP}$), the PRM rescales the input patch (*IP*) accordingly as given in [83]. Consequently, it generates one or more new 256×256 size patches. For ($Class_{IP} = NCP, LCP, MCP$), the PRM generates $(1, 1, 1)$ new patches respectively. In case of ($Class_{IP} = HCP$), the PRM divides the input patch (*IP*) into four new 128×128 patches, then upscales each by 2× to output the final 256×256 size patches [83]. The PRM generated patches then separately go through the concatenation module (CMod) to generate the initial feature-maps. Eventually, these features go through the concatenation process with the (Branch-1, Phase-2) second RB-block output, followed by the Bottleneck layer (*BL*) to adjust the number of channels before proceeding further.

The *CH* configuration is shown in Table 4.1. It utilizes the softmax based 4-way classification activation with the cross-entropy loss, given as follows:

$$Loss_{CH} = -\sum_{i=1}^{4} y_i log(\hat{y}_i) \tag{4.1}$$

where y_i denotes the actual class (1 or 0) and $\hat{y}_i$ indicates the predicted class label. Similarly, the concatenation module (CMod) configuration is shown in Table 4.1, where it consists of several deep layers to eventually yield the final ($32 \times 64 \times 64$) channels to be used next for the concatenation. It is also worth mentioning that the input patch (*IP*), classified as the NCP label during the test time, will be automatically discarded without any further processing. This is very effective especially in the case of discarding cluttered background regions in the images (e.g. tree leaves)

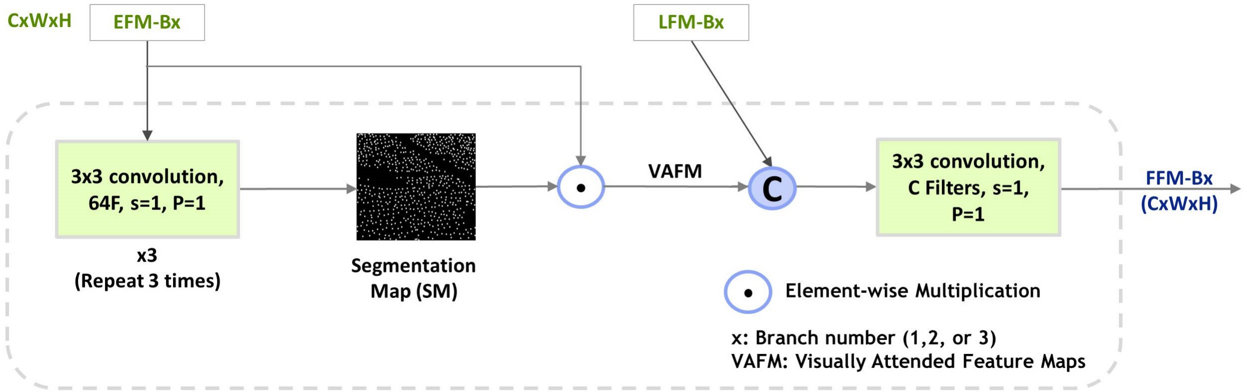

CxWxH
EFM-Bx
LFM-Bx
3x3 convolution, 64F, s=1, P=1
x3
(Repeat 3 times)
Segmentation Map (SM)
VAFM
C
3x3 convolution, C Filters, s=1, P=1
FFM-Bx (CxWxH)
Element-wise Multiplication
x: Branch number (1,2, or 3)
VAFM: Visually Attended Feature Maps

Figure 4.3: Visual Attention and Concatenation module (VACM). The early-stage Feature-Maps (EFM-Bx) from the Branch Bx (=1,2, or 3) get visual attention and output the segmentation map (SM) that undergoes an element-wise multiplication with the original EFM-Bx channels to generate the visually-attended feature-maps (VAFM). These VAFM maps concatenate with the later-stage feature-maps (LFM-Bx) of the same branch before passing through the channel-adjusting deep layer to output the final feature-maps (FFM-Bx) for the specific branch Bx.

that look very similar to the dense-crowd region.

4.3.4 Visual Attention and Concatenation Module (VACM)

The purpose of the VACM module is to visually attend the EFM feature-maps and share them with the LFM channels for better foreground vs background understanding. The VACM module visually attends the early-stage feature-maps (EFM-B1, EFM-B2, EFM-B3) and concatenates them separately with their respective branch later-stage channels (LFM-B1, LFM-B2, LFM-B3). The resultant final feature-maps (FFM-B1, FFM-B2, FFM-B3) then proceed through the crowd regression head (CRH) for the final crowd estimate. As shown in Fig. 4.3, the input EFM-Bx block passes through three deep convolution layers to produce the attention-based segmentation map (SM $\in [0,1]_{W\times H}$). The SM undergoes element-wise multiplication with the original EFM-Bx block to produce visually attended feature-maps (VAFM). These VAFMs are then concatenated with the input LFM-Bx block and channel adjusted to output final feature-maps (FFM-Bx) for the branch Bx.

This early-attention mechanism helps in strengthening the early-stage feature-maps of each

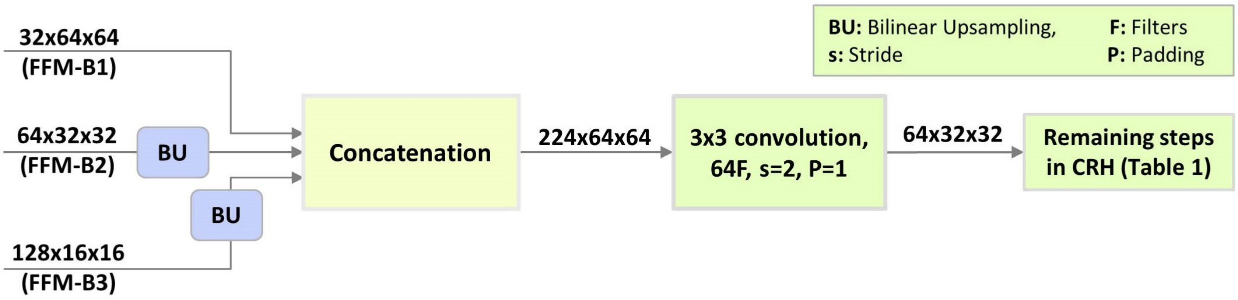


Figure 4.4: Concatenation-based Crowd Regression Head (CRH) concatenates the lower-level branches with the highest-level (Branch-1), followed by several deep layers to form the regression head.

branch towards better understanding and distinguishing the spatial foreground in comparison to the background. More importantly, this information becomes an integral part of the later-stage channels collective knowledge by concatenation-based sharing with them. Consequently, later-stage channels use this information to boost areas of interest and neglect background pixels. The SM weights are trained using the cross-entropy error ($Loss_{SM}$) between the SM and the ground-truth map. To compute the ground-truth map, we use each person's localization information already available in the benchmark datasets. This attention process significantly improves the network performance as shown in the ablation study in Sec. 4.5.4.

4.3.5 Crowd Regression Head (*CRH*)

The VACM module outputs three separate sets of final feature-maps (FFB-B1, FFB-B2, FFB-B3), each for the respective branch. These channels are then routed to the Crowd Regression Head (CRH). The CRH concatenates the FFB-B2 and FFB-B3 outputs with the FFB-B1 output channels using the bilinear upsampling (BU) as shown in Fig. 4.4. Following that, it passes through various deep convolution and Fully Connected (FC) layers and eventually through the final 1-dimensional single neuron (1D, FC) to regress the final crowd count (CC_{IP}) for the input patch (IP) as shown in Table 4.1. When class label $Class_{IP} = HCP$, then the crowd count for the input patch (IP) is the sum of all four PRM generated and rescaled (256×256 size) patches, given as follows:

$$CC_{IP} = \sum_{i=1}^{4} cc_{p(i)}, \ (if\ Class_{IP} = HCP) \tag{4.2}$$

where $cc_{p(i)}$ with $(i = 1,2,3\ and\ 4)$ denotes the four rescaled patches created by the PRM for the input patch (*IP*) being classified with the *HCP* label. The regressor utilizes Mean Square Error (MSE) as its training loss function, defined as follows:

$$Loss_{Regressor} = \frac{1}{T} \sum_{s=1}^{T} (F(x_s, \Theta) - y_s)^2 \tag{4.3}$$

where T represents the total training samples per batch, y_s indicates the actual crowd number for the input patch x_s, and $F(.)$ represents the mapping or transformation function with the learnable weights parameters Θ that learns to regress the crowd number for the input patch. Finally, the total network loss ($Loss_{total}$) is the sum of 4-way classification, segmentation map (SM), and the crowd regression losses:

$$Loss_{total} = Loss_{Regressor} \ + \ Loss_{CH} + \ Loss_{SM} \tag{4.4}$$

4.4 Evaluation and Training Details

We utilize the commonly used crowd counting metrics for numerical evaluation: Mean Absolute Error (*MAE*) and Root Mean Squared Error (*RMSE*), as defined below:

$$MAE = \frac{1}{TI} \sum_{h=1}^{TI} |CC_h - \hat{CC_h}| \tag{4.5}$$

$$RMSE = \sqrt{\frac{1}{TI} \sum_{h=1}^{TI} (CC_h - \hat{CC_h})^2} \tag{4.6}$$

where TI indicates the total number of test images, and CC_h *and* $\hat{CC_h}$ denote the true and the predicted crowd counts respectively for the test image h.

Network Training Details. To train the network, we randomly take out $75,000$ patches of 128×128, 256×256, and 512×512 sizes from the predefined training images. The resultant patches with mixed crowd numbers are resized to 256×256 size as per required. We double the training patches number by performing the horizontal-flip based data augmentation on each patch. The proposed network has been trained for 120 epochs with a batch size of 16. Stochastic gradient descent (SGD) has been used as the optimizer with a weight decay of 0.0001 and the Nesterov Momentum equal to 0.9. We also employed the multi-step learning rate (α) that starts with the value of 0.001 and decreases to 1/2 after every 30 epochs. Also, 10% training data has been separated for the network validation purpose during the model training.

4.5 Experiments

This section presents both numerical and visual results based on different experiments conducted on three extensively used crowd counting benchmarks: UCF-QNRF [35], ShanghaiTech [127], and the AHU-Crowd [32] dataset. These datasets are totally different from each other as they vary remarkably in terms of image resolution, average people count per image, maximum/minimum people per image, background regions, total images, and different lighting conditions. First, we compare our scheme numerically with the state-of-the-art models on these benchmarks. Next, we discuss the ablation experiments and the cross-dataset evaluation. Finally, visual analysis has been presented to discuss qualitative performance.

4.5.1 UCF-QNRF Dataset Numerical Evaluation

The UCF-QNRF [35] dataset contains a total of $1,535$ images with a pre-established training/testing division of 1201/334 respectively. The images contain a wide range of crowd-density and vary greatly in image resolution and background setting. The total people annotations in the dataset

	ShanghaiTech		UCF-QNRF	
Method	MAE	RMSE	MAE	RMSE
MCNN [127]	110.2	173.2	277	426
CMTL [94]	101.3	152.4	252	514
Switch-CNN [85]	90.4	135.0	228	445
SaCNN [126]	86.8	139.2	-	-
IG-CNN [3]	72.5	118.2	-	-
ACSCP [87]	75.7	102.7	-	-
CSRNet [54]	68.2	115.0	-	-
CL[35]	-	-	132	191
CFF [91]	65.2	109.4	93.8	146.5
RRSP [103]	63.1	96.2	-	-
CAN [59]	62.3	100.0	107	183
L2SM [115]	64.2	98.4	104.7	173.6
BL [65]	62.8	101.8	88.7	154.8
RRP [13]	63.2	105.7	93	156
HA-CCN [95]	62.9	94.9	118.1	180.4
ADSCNet [6]	**55.4**	97.7	**71.3**	132.5
RPNet [119]	61.2	96.9	-	-
ZoomCount [82]	66.6	94.5	128	201
PRM-based[83]	67.8	86.2	94.5	141.9
Ours	56.1	**79.8**	**71.3**	**120.7**

Table 4.2: Numerical experiments on the UCF-QNRF [35] and the ShanghaiTech [127] benchmarks. Our proposed scheme outperforms the state-of-the-art models (including the original PRM-based) under the RMSE standard criterion, while indicating closer or equal to the best results for the MAE evaluation metric.

equal to $1,251,642$, while the images resolution varies between (300×377) and (6666×9999). We compare the proposed approach with the state-of-the-art models as reported in Table 4.2. The evaluation demonstrates that our model performs the best in comparison to the state-of-the-art for the RMSE evaluation criterion with 8.9% performance boost (from 132.5 to 120.7) amid performing equally best for the MAE metric. It may also be noted that the proposed model performs better for both metrics as compared to the original PRM based scheme (CC-2P) [83].

Method	MAE	RMSE
Haar Wavelet [71]	409.0	-
DPM [20]	395.4	-
BOW–SVM [18]	218.8	-
Ridge Regression [12]	207.4	-
Hu et al. [32]	137	-
DSRM [120]	81	129
ZoomCount [82]	74.9	111
CC-2P (PRM-based)[83]	66.6	101.9
Ours	**57.5**	**89.0**

Table 4.3: AHU-Crowd [32] benchmark dataset based experiments indicate that the proposed scheme appears to be the best for both the evaluation metrics in contrast to the state-of-the-arts including the original PRM-based model.

4.5.2 ShanghaiTech Dataset Numerical Evaluation

The ShanghaiTech [127] dataset, comprises of a total 482 images with diverse crowd range, images resolution, and varying lighting conditions. These images are already divided into 300 training and 182 testing images. The average image resolution in the benchmark is 589×868 with a total of $241,677$ human annotations in total. We analyze the proposed scheme on this benchmark and also compare it with the state-of-the-art (including the original PRM-based [83]) models. The results, as shown in Table 4.2, from which we can see that our model yields the best results for RMSE criterion with an improvement of 7.4% (from 86.2 to 79.8), and also performs reasonably well in terms of the MAE evaluation metric.

4.5.3 Numerical Experiments on AHU-Crowd Dataset

AHU-Crowd dataset [32] poses a great challenge with totally different statistics as compared to the ShanghaiTech and the UCF-QNRF datasets. It only contains 107 images with 58 to $2,201$ ground-truth people count per image. The dataset also contains a total of $45,807$ people annotations. Based on the standard literature practice, we carried out the 5-fold cross-validation for the (MAE, RMSE) based numerical evaluation. In each fold, 96 images were selected for training, and the remaining 11 images for the testing purpose. We report the numerical evaluation and comparison results

Total Branches Choice		
Total Branches	MAE	RMSE
1	91.6	131.4
2	73.6	109.7
3 (our default choice)	**56.1**	**79.8**
4	69.6	97.5

Residual Units (RUs) per RB Block Quantity Effect		
RU units per RB block	MAE	RMSE
2	77.9	101.1
3	71.5	98.3
4 (our default choice)	**56.1**	**79.8**
5	71.7	97.2
6	73.6	102.8

2-Layered vs 3-Layered Residual Unit Choice		
	MAE	RMSE
2-Layered	70.6	96.9
3-Layered (our default choice)	**56.1**	**79.8**

Branching-out Positioning in the Network		
From	MAE	RMSE
RB (of Phase1)	71.0	97.9
1st RB (of Branch-1, Phase2)	**56.1**	**79.8**
2nd RB (of Branch-1, Phase2)	69.1	94.8
1st RB (of Branch-1, Phase3)	70.9	96.1
Visual Attention (VACM) Effect		
	MAE	RMSE
w/o VACM	63.9	82.4
w VACM (our default choice)	**56.1**	**79.8**

Table 4.4: Five different sets of ablation experiments validate our selection of the few vital hyper-parameters for the proposed network.

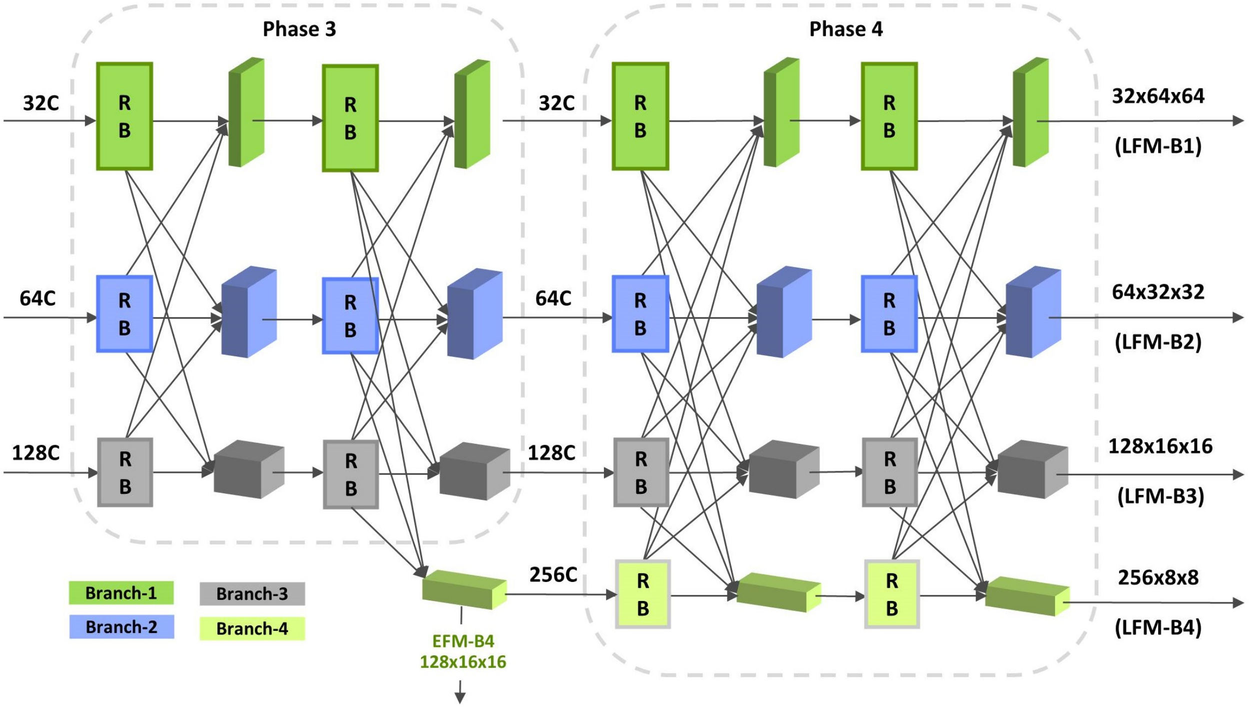


Figure 4.5: Four-branch based version of the proposed network for the branch quantity ablation study. In this setting, a new branch (Branch-4) has been added into the proposed (three-branch) model that naturally results in a new phase (Phase-4) as well. Also, one additional EFM (EFM-B4) block emerges from Branch-4 that passes through VACM along with its respective LFM block (LFM-B4). The outputs from Phase-4 now serve as the LFM feature-maps for the remaining crowd estimation process.

in Table 4.3. These findings indicate that the proposed method outperforms other state-of-the-art methods (including the original PRM based scheme) with significant improvement of 13.7% (from 66.6 to 57.5), 12.6% (from 101.9 to 89.0) for the MAE and RMSE metrics respectively.

4.5.4 Ablation Experiments Study

In this section, we present five ablation studies on the ShanghaiTech dataset [127] to investigate the effect of different components of the proposed scheme.

Effect of the total number of Columns/Branches. First ablation study discusses the consequences related to the quantity of multi-resolution branches (or columns) being used in the proposed network. We explore this critical hyper-parameter by experimenting separately with different quantities of such branches. The study results are shown in Table 4.4. Using only one branch ($Branch-1$) produces the worst results, since it only contains a single column and thus lacks any fusion or information-sharing. Two-Branch ($Branch-1, Branch-2$) network does not contain the $Branch-3$ column, but performs better than one-branch based model. The proposed network, with three multi-resolution branches, gives the best performance in comparison to the above configurations as well as the model with four multi-resolution branches as indicated in the same table. Due to this ablation study outcome, the proposed model has been designed with three multi-resolution branches. In the four-branch network experiment, we deployed an additional branch ($Branch-4$) with $2\times$ down-scaled resolution (8×8) and double the channels (256) than the $Branch-3$ as shown in Fig. 4.5. The four-branch based model naturally contains an additional phase ($Phase4$) to cover the fusion process for $Branch-4$ with other branches. Naturally, it also contains one more early-stage block (EFM-B4) emerging from Branch-4. The EFM-B4 is routed to the VACM module along with its respective later-stage block (LFM-B4). Additionally, the $Phase4$ outputs now serve as the LFM blocks.

Effect of the number of residual units in RB blocks. Here, we investigate the effect of using four 3-layered residual units (RU) per RB block as compared to deploying other potential quantities ($2,3,5$, or 6 RU units per RB block). As shown in the ablation experiments results in Table 4.4, 4 RU units per RB block yield the best results with the lowest MAE and RMSE errors. Thus, it acts as our default and preferred choice in the proposed network.

Effect of using 3-layered vs 2-layered residual unit. We have two major choices for the Residual Unit: 2-layered or 3-layered deep residual unit as given in [26]. Results for both choices are shown in Table 4.4. It is evident that the 3-layered RU performs much better than the 2-layered residual unit.

Effect of branching-out location in the network. As shown in the proposed network in Fig. 4.1, we branch-out the output features of the first RB block in the (Branch-1, Phase-2) to feed into the classification head. Here, we investigate the effect of the location of this branching-out by re-positioning it to other Branch-1 RB blocks output. As shown in Table 4.4, our default choice

Method	MAE	RMSE
Switch-CNN [85]	301	457
Cascaded-mtl [94]	308	478
CC-2P (PRM based) [83]	219	305
Ours	**197**	**271**

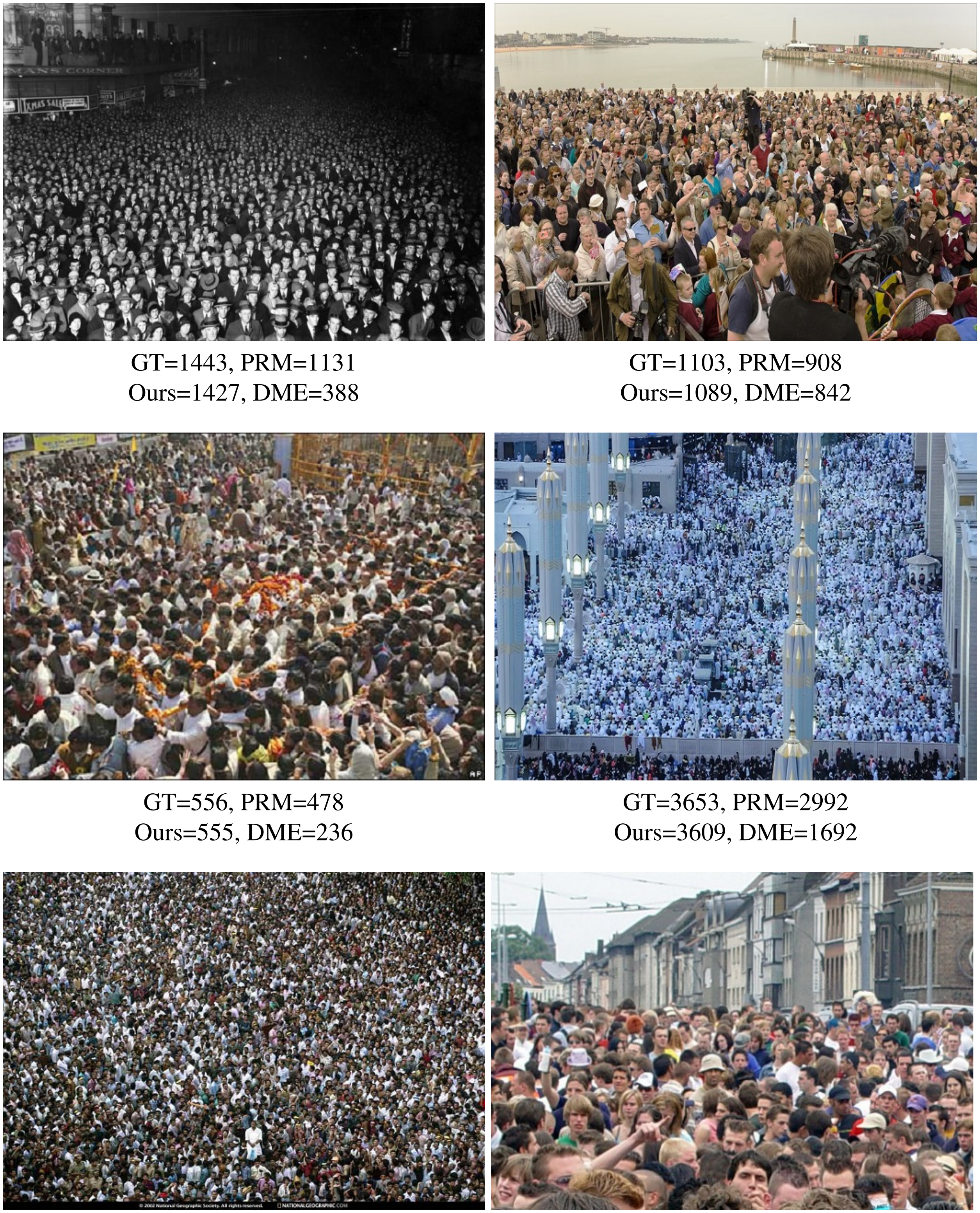
Table 4.5: Based on the comparison with the state-of-the-art methods during the cross-dataset evaluation, our approach outperforms them under both evaluation criteria.

of branching-out from 1st RB of (Branch-1, Phase-2) gives the lowest MAE and RMSE error. Additionally, in each ablation experiment setting, the concatenation back into the network happens with the RB block of the Branch-1 that is next and subsequent to the RB block responsible for the branching-out.

VACM Module Effect. Visual attention on the early-stage channels helps in enriching the later-stage feature-maps. Consequently, the VACM module should improve the overall network performance. As shown in Table 4.4, the visual attention process boosts the network effectiveness by (12.2%,3.2%) in terms of (MAE,RMSE) respectively.

4.5.5 Cross-Dataset Evaluation

To further assess the proposed model, we conducted the cross-dataset evaluation. ShanghaiTech [127] benchmark has been used for all models training, while the testing has been conducted using the UCF-QNRF dataset. Table 4.5 reports the proposed approach cross-dataset performance and also compares it to other state-of-the-art schemes. It is clear from the results that the proposed model outperforms other methods including the original PRM-based scheme. These findings also indicate the better generalization potential of our scheme towards unseen images with different dynamics and crowd diversity.



GT=1443, PRM=1131
Ours=1427, DME=388

GT=1103, PRM=908
Ours=1089, DME=842

GT=556, PRM=478
Ours=555, DME=236

GT=3653, PRM=2992
Ours=3609, DME=1692

GT=2472, PRM=2017
Ours=2488, DME=1370

GT=207, PRM=130
Ours=204, DME=109

Figure 4.6: Visual evaluation.

4.5.6 Visual Analysis

In this section, we show a few visual results. Six original test images are shown in Fig. 4.6, where we analyze our scheme against the state-of-the-art PRM [83] and the density-map estimation (DME) [35] based methods. These images contain hugely varying crowd-density and scale with fluctuating lighting conditions and background. As shown in Fig. 4.6, the proposed scheme yields the best results that are closer to the ground-truth in comparison to the two competing models.

4.6 Conclusion

We have proposed a new multi-resolution fusion and multi-task based crowd counting network with visual attention in this chapter by further exploring and more effectively utilizing the PRM module. The proposed method relies on the PRM module and builds the collective knowledge using the feature-level fusion across the multi-resolution branches as well as visually attending the early-stage channels to boost the foreground vs background understanding of later-stage channels. This integration technique outperforms the state-of-the-art (including the original PRM based) methods as demonstrated through extensive standard numerical and visual experiments and comparisons. The proposed scheme also demonstrates better generalization ability during the cross-dataset evaluations.

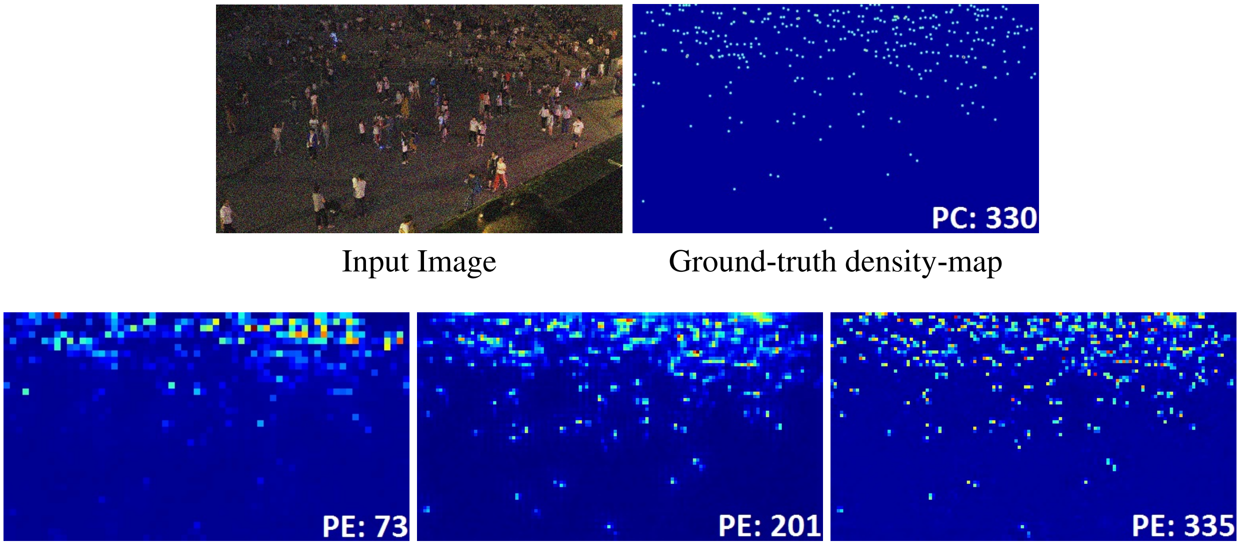
Chapter 5

Audio-Visual Transformer Based Crowd Counting (Multi-Modal)

This chapter deals with the multi-modal crowd counting network that combines audio and image inputs for effective people count. Most content of this chapter comes from our paper [79]. The most recent study [31] tries to exploit auditory information to aid the visual models, however, the performance is limited due to the lack of an effective approach for feature extraction and integration. This work proposes a new audio-visual multi-task network to address the critical challenges in crowd counting by effectively utilizing both visual and audio inputs for better modalities association and productive feature extraction. The proposed network introduces the notion of auxiliary and explicit image patch-importance ranking (PIR) and patch-wise crowd estimate (PCE) information to produce a third (run-time) modality. These modalities (audio, visual, run-time) undergo a transformer-inspired cross-modality co-attention mechanism to finally output the crowd estimate. To acquire rich visual features, we propose a multi-branch structure with transformer-style fusion in-between. Extensive experimental evaluations show that the proposed scheme outperforms the state-of-the-art networks under all evaluation settings with up to 33.8% improvement. We also analyze and compare the vision-only variant of our network and empirically demonstrate its superiority over previous approaches.

5.1 Introduction

Crowd estimation requires one to count the total people in the given image. It finds many applications in real-world scenarios, e.g., better management of crowd gatherings, safety and secu-



Input Image
Ground-truth density-map
Visual-only [31]
Audio-Visual [31]
Audio-Visual (Ours)

Figure 5.1: For the low-quality input image with severe conditions such as noise, low-illumination, or low-resolution, the proposed audio-visual model yields the best and more fine-grained people estimate (PE) as evaluated using the ground-truth density-map and people count (PC).

rity, and circumventing any undesirable incident. Many deep learning-based image-only schemes [81, 85, 83, 35, 83, 40, 127, 60] have been proposed to date, ranging from single and multi-branch networks [127, 81, 83], multi-regressors [85] based to trellis networks [40]. Although they show reasonable performance in regular images, they fail to generalize well in many practical scenarios such as low illumination and lighting conditions, noise, severe occlusion, and low-resolution images, where visual information is scarce. Consequently, they give huge crowd under-estimation as shown in Fig. 5.1. Lack of visual clues may also invoke highly sensitized behavior in these models towards different image regions, resulting in large over-estimation. Moreover, in the case of regular images, sub-optimal capabilities of these state-of-the-arts implicate that there is a lot of room for further improvement.

One compelling way to address these challenges is to investigate the effect of utilizing more than one modality (e.g., image and audio). Recently, Hu *et al.* [31] introduced a novel audio-visual crowd counting task and proposed an estimation model that jointly learns both visual and audio features and fuses them together. The results demonstrate that combining the related audio modality with the visual input significantly improves the crowd estimate in such conditions. However, it only accounts for the parametric influence of audio features on the visual ones without making full use of the audio-visual information, thus, under- or over-estimating the crowd as shown in Fig. 5.1.

On the other hand, the learning and fusion of visual and audio modalities have been applied with reasonable success to other computer vision problems, e.g. classification tasks [112, 9, 30, 45], event localization [57, 116], and speech recognition [123, 25, 68, 96]. However, these schemes are generally not suitable for the crowd estimation task because of very few pixels per person, and thus require a specifically tailored method to obtain pixel-perfect results. Moreover, these schemes (including [31]) mostly focus on improving the intra- or inter-modality fusion process, and often ignore the significant visual feature extraction part by normally using the conventional VGG [93] or ResNet-based [26] standard structures for that.

To address these major challenges and issues, we propose a new transformer-based [101] audio-visual multi-task crowd counting network as shown in Fig. 5.2. It consists of an Audio-Visual Transformer (AVT) that generates two auxiliary network outputs, image patch-importance ranking (PIR) and patch-wise crowd estimate (PCE), as part of the inter-modality fusion process. This explicit PIR and PCE information also helps AVT module in generating a third run-time audio-visual attended modality that consequently helps in constructive association and transformer-style co-attention of audio-visual features. Furthermore, no extra ground-truth annotation process is required to embed the PIR and PCE into the proposed network. Second, instead of deploying the conventional and standard structure for visual feature extraction, we use the multi-scale branches that also undergo the unique transformer-inspired inter-scale fusion process to yield rich and productive visual representations. Extensive experiments show that the proposed model outperforms the state-of-the-art methods in all settings with up to 33.8% improvement, especially in challenging situations such as shown in Fig. 5.1.

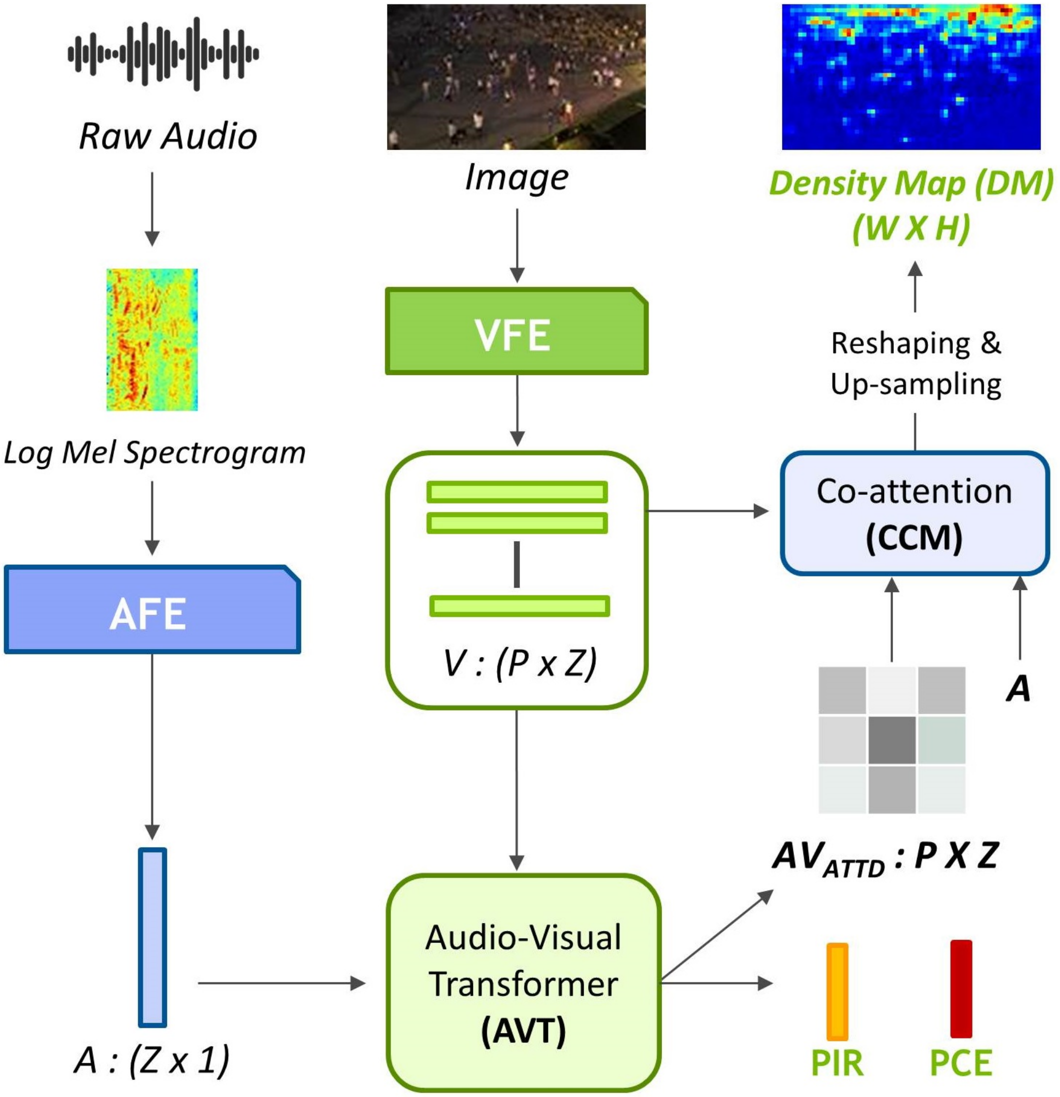


Figure 5.2: The proposed audio-visual crowd counting network. The extracted audio-visual features (V,A) go through the AVT module to obtain two auxiliary network outputs (PIR,PCE) and third (run-time) modality (AV_{ATTD}). The AV_{ATTD} undergoes the cross-modality co-attention fusion with V and A via the CCM module, followed by getting the final crowd density-map (DM).

5.2 Our Contributions

The key contributions of this work are listed as follows:

- We propose a novel audio-visual multi-task crowd counting network for effective estimation in both regular and severe conditions. To the best of our knowledge, this is the first attempt to use the transformer-style mechanism for this task.
- We introduce the notion of auxiliary PIR and PCE information, and empirically shown that it is beneficial for better modalities association and extracting rich visual features without requiring any extra ground-truth annotation process.
- We also design an image-only variant of our model. Extensive experimental evaluations on benchmark datasets indicate that the proposed networks significantly outperform the state-of-the-art.

5.3 Proposed Approach

The proposed multi-task model, as shown in Fig. 5.2, exploits both input image and audio modalities for effective crowd estimation. First, we separately extract rich features for both modalities, then pass them through the Audio-Visual Transformer (AVT) to generate two auxiliary network outputs: Patch-Importance Ranking (PIR) and Patch-wise Crowd Estimate (PCE). The explicit PIR and PCE vectors play a crucial role in improving the final crowd estimate, and also help the AVT in generating the audio-visually attended channels. These attended channels then undergo the cross-modality co-attention process along with the original audio-visual features (V,A) via the CCM module. Finally, the CCM output goes through the reshaping and up-sampling steps to give the crowd-density map, where we sum-up all its pixel values to yield the final crowd count. The network components are detailed below.

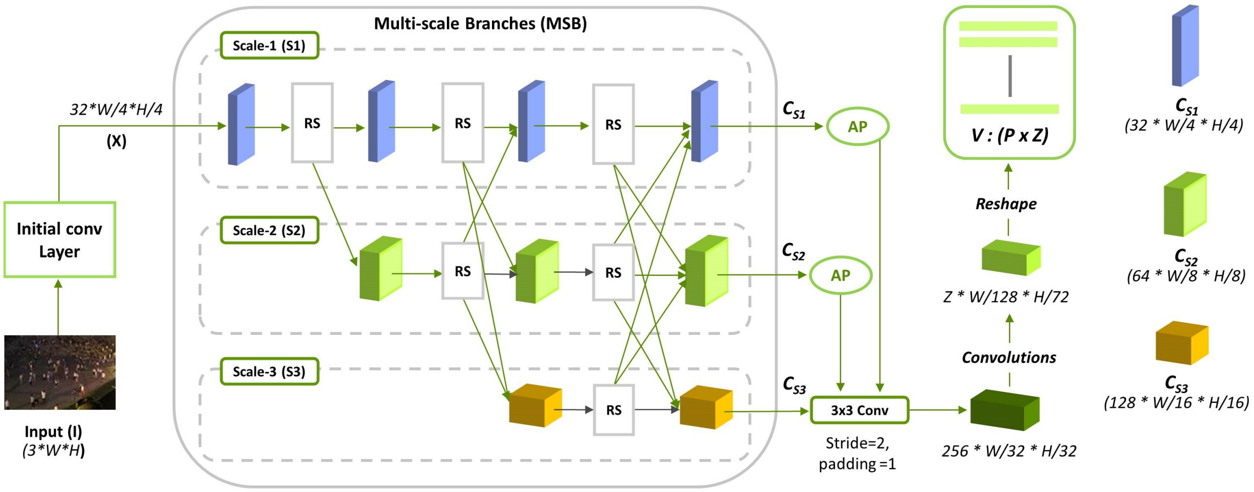


Figure 5.3: The framework of Visual Feature Extraction (VFE) block.

5.3.1 Audio Feature Extraction (AFE)

To extract the audio features embedding, we deploy the ResNet-like CNN structure [29] (pre-trained on the AudioSet dataset [48]) and apply it on the conventionally computed [31] Log Mel-Spectogram (LMS) representation of the raw one-second duration input audio signal. For the given Audio LMS ($A_{LMS} \in \mathbb{R}^{64*96}$), audio CNN ($AFE$) yields the vector output as follows:

$$A = AFE(A_{LMS}) \tag{5.1}$$

where $A \in \mathbb{R}^{Z*1}$ represents the extracted audio embedding.

5.3.2 Visual Feature Extraction (VFE)

The VFE component, as shown in Fig. 5.3, comprises of three multi-scale branches (MSB) with repeated inter-branch fusion. The input image ($\mathrm{I} \in \mathbb{R}^{3*W*H}$) passes through two initial (3×3) convolutional layers to obtain the down-scaled channels ($X \in \mathbb{R}^{32*\frac{W}{4}*\frac{H}{4}}$). These features then proceed through the multi-scale branches ($S1, S2, S3$) that are composed of several residual structures (RS). The RS block contains four residual units, where each unit is composed of a three-layer based ResNet building block [26]. Similar to the high-resolution networks [99, 105], each branch retains its channel quantity and resolution throughout that branch. Channel quantity doubles each time as we move from S_1 to S_3, while the resolution decreases by half. The MSB module outputs three separate sets of channels ($C_{S1} \in \mathbb{R}^{32*\frac{W}{4}*\frac{H}{4}}, C_{S2} \in \mathbb{R}^{64*\frac{W}{8}*\frac{H}{8}}, C_{S3} \in \mathbb{R}^{128*\frac{W}{16}*\frac{H}{16}}$).

5.3.2.1 Inter-Branch Fusion

The purpose of inter-branch fusion is to develop coordinated knowledge in-between the multi-scale branches. We denote this process as $\{S\} \rightarrow T$, indicating the fusion from one or two source branches (S) channels into the target branch (T) features. We deploy the transformer-inspired attention mechanism to achieve such fusion. All cases are detailed below as well as illustrated in Fig. 5.4. It is worth mentioning that later stage three-branch fusions also integrate the audio embedding (A) during the fusion process, which empirically proves beneficial and also unique multi-modal strategy to the proposed method. The fusion process significantly helps the VFE in preparing constructive and co-attended visual features (V) for the next steps.

S1→ S2 only (and vice versa). In this step, we first down-sample the source branch channels (C_{S1}) via 3×3 convolution to match the resolution and quantity of the S2 channels. The resultant channels (C'_{S1}) are converted into attention-weights (AW), which separately undergo the attention mechanism with the respective target branch channels (C_{S2}) to give visually-attended features (C'_{S2}) as shown in Fig. 5.4. Mathematically, it is defined as:

$$C'_{S2} = AW * C_{S2} = softmax(C'_{S1} * C'^{T}_{S1}) * C_{S2} \tag{5.2}$$

where $*$ and T denote matrix multiplication (MatMul) and transpose respectively. In case of **(S2→ S1 only) fusion**, the approach remains same except that the lower-branch channels (C_{S2}) are bi-linearly up-sampled to match C_{S1} features dimensions before fusing together as shown in Fig. 5.4. **{S1, S2}→ S3 Fusion case.** Both higher-branch source channels (C_{S1},C_{S2}) get down-scaled to match the lowest-branch channels (C_{S3}) dimensions. The generated channels (C'_{S1},C'_{S2}) are added element-wise to produce features C_{S12}. After the linear-layer operation on the audio embedding row-vector (A^T), it separately performs element-wise addition with each row of C_{S12}. The resultant

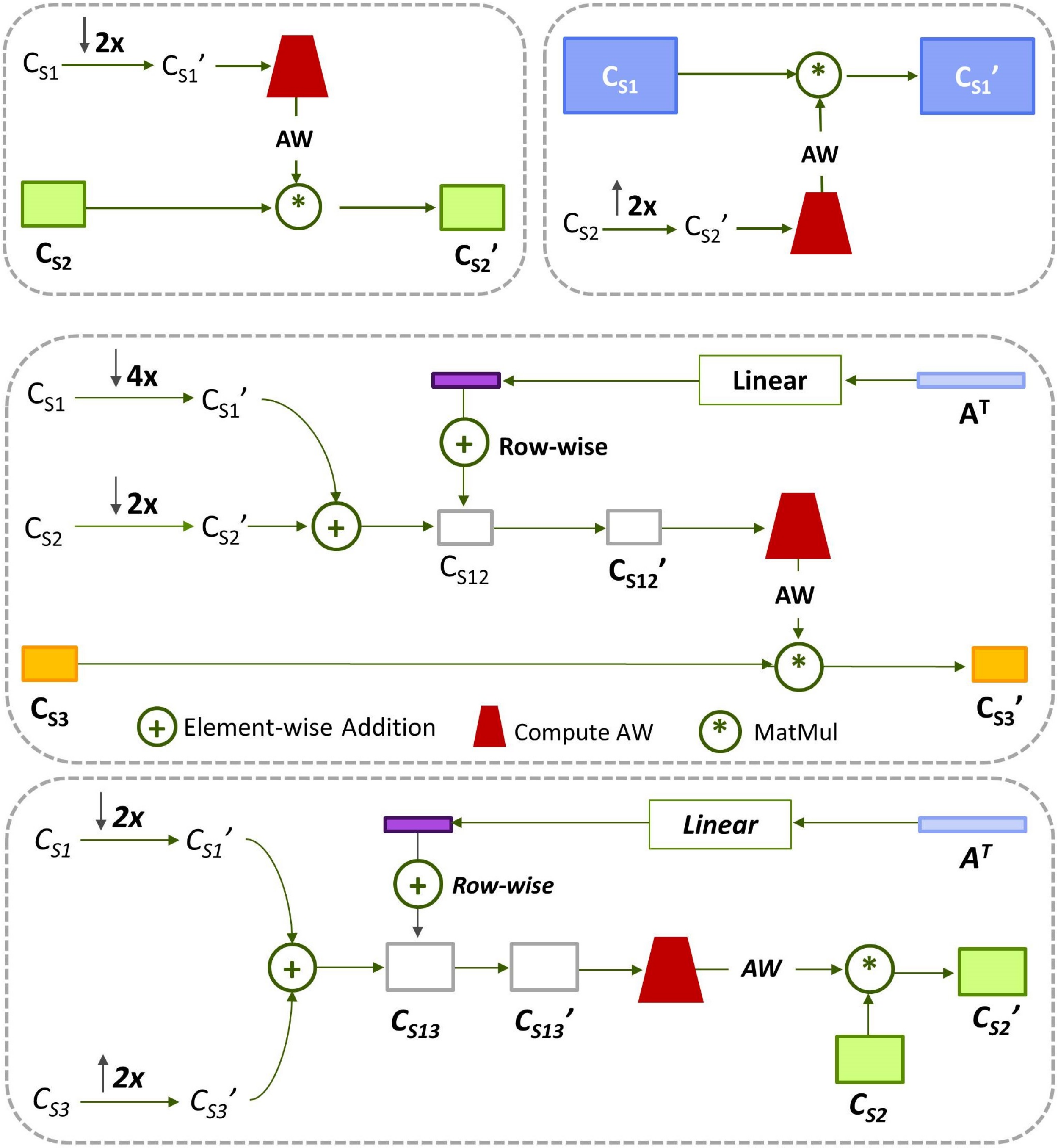


Figure 5.4: Illustration of different transformer-inspired inter-branch fusion cases. (S1 → S2 only (top-left), S2 → S1 only (top-right), {S1, S2} → S3 (middle), {S1, S3}, → S2 (bottom))

C'_{S12} is being used next to obtain the attention weights (AW). The AW finally gets applied on the target branch channels (C_{S3}) to produce the audio-visual attended channels (C'_{S3}) as shown in Fig. 5.4. It is defined as:

$$C'_{S3} = AW * C_{S3} = softmax(C'_{S12} * C'^{T}_{S12}) * C_{S3} \tag{5.3}$$

where $C'_{S12} = C_{S12} \oplus Linear(A^T)$. Similarly, **the ({S2, S3}→ S1) case** takes the same direction as stated above except that now the source channels (C_{S2}, C_{S3}) first get up-scaled to match the dimensions of the target channels (C_{S1}).

{S1, S3}→ S2 fusion case. The first (C_{S1}) and third (C_{S3}) branch channels are down- and up-sampled respectively by 2× to match the $S2$ dimensions, followed by their element-wise summation to generate C_{S13}. We apply the linear-layer on the audio embedding (A^T), which is separately added to each row of C_{S13} via element-wise summation. The produced channels (C'_{13}) are used to obtain the attention-weights (AW) that get applied on target channels (C_{S2}) to yield audio-visual attended features (C'_{S2}) as shown in Fig. 5.4. We can define it as:

$$C'_{S2} = AW * C_{S2} = softmax(C'_{S13} * C'^{T}_{S13}) * C_{S2} \tag{5.4}$$

where $C'_{S13} = C_{S13} \oplus Linear(A^T)$.

5.3.2.2 Visual Features Generation

The MSB higher-scales outputs (C_{S1}, C_{S2}) are merged together with the lowest-branch output channels (C_{S3}) through (3 × 3) convolution after down-scaling higher features via required average pooling (AP). The generated channels ($\in \mathbb{R}^{256 * \frac{W}{32} * \frac{H}{32}}$) employ several convolution layers defined as follows: {Conv2d(256,144,3,(1,1),1)-BN-ReLU, Conv2d (144,144,3,(4,1),1)-BN-ReLU}. Where Conv2d (I,O,F,P,S) indicates I: input channels, O: output channels, F: F×F filter, P: padding in (H,W), S: stride, and BN and ReLU denote Batch-Normalization [36] and ReLU [69] activation function. The resultant channels ($\in \mathbb{R}^{Z * \frac{W}{128} * \frac{H}{72}}$) are reshaped to give the VFE module output as follows:

$$V = VFE(X), \tag{5.5}$$

where $V \in \mathbb{R}^{P*Z}$, and P $(= \frac{W}{128} * \frac{H}{72})$ represents the total patches/regions in the input image. Intuitively, the V matrix can be perceived as containing the Z-dimensional embedding for each image-patch, with P total patches.

5.3.3 Audio-Visual Transformer (AVT)

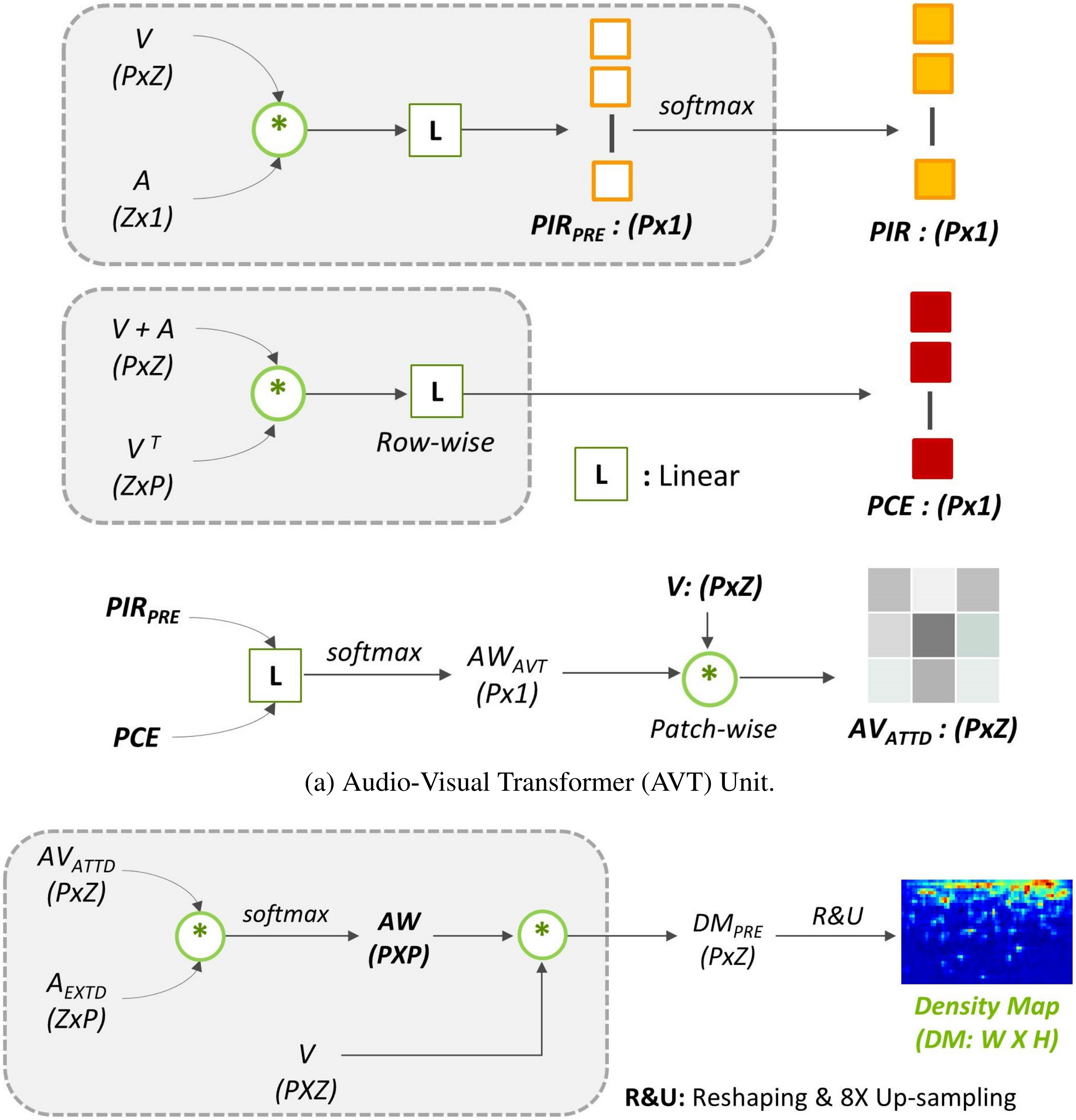
The purpose of the AVT module is twofold: 1) Calculate and output auxiliary Patch-Importance Ranking (PIR) and Patch-wise Crowd Estimate (PCE) Information, 2) Combine this information to generate third run-time modality to be used by the subsequent Co-attention (CCM) module. The AVT process, as shown in Fig. 5.5(a), contains two separate streams to compute PIR and PCE. The AVT calculations are primarily inspired by the transformer-style dot-product attention amid using both visual (V) and audio (A) features. The PIR computation is defined as:

$$PIR = softmax(PIR_{PRE}) \tag{5.6}$$

where $PIR_{PRE} = Linear(V * A)$ and $PIR \in \mathbb{R}^{P*1}$. Intuitively, the PIR probability vector ith value gives the percentage of total image people contained in the ith image-patch. To set the ground-truth PIR vector jth value ($PIR_{GT}(j)$) for training, we use following formula:

$$PIR_{GT}(j) = \frac{CC_{GT}(j)}{CC_{GT}(image)} \tag{5.7}$$

where $CC_{GT}(j)$ and $CC_{GT}(image)$ denote the actual crowd-count in the jth patch and whole input image respectively. The KL-Divergence based loss function has been used to measure similarity between the PIR probability vector and the ground-truth probability distribution (PIR_{GT}):



(a) Audio-Visual Transformer (AVT) Unit.

(b) Cross-modality Co-attention Module (CCM).

Figure 5.5: Illustration of PIR, PCE, AV_{ATTD}, and DM computations.

$$Loss_{PIR} = \frac{1}{\sqrt{P}} \sum_{i=1}^{P} PIR_{GT}(i) \log(\frac{PIR_{GT}(i)}{PIR(i)}) \tag{5.8}$$

where $\frac{1}{\sqrt{P}}$ acts as a scaling factor. Similarly, the PCE vector is computed as:

$$PCE = Linear_{Row-wise}((V + A) * V^{T}) \tag{5.9}$$

where $Linear_{Row-wise}$ indicates the row-wise linear-layer operation on the ($P \times P$) matrix to obtain ($PCE \in \mathbb{R}^{P*1}$) as shown in Fig. 5.5(a). Intuitively, the ith value in the PCE vector gives the network estimate for the ith image patch. The ground-truth PCE vector computation strategy is the same as for PIR. The squared-normalized-difference loss function has been deployed for the PCE output, given as follows:

$$Loss_{PCE} = \sum_{i=1}^{P} (\frac{PCE_{GT}(i) - PCE(i)}{\sum_{j=1}^{P} PCE_{GT}(j)})^{2} \tag{5.10}$$

where PCE_{GT} indicates the ground-truth PCE vector and $\sum_{j=1}^{P} PCE_{GT}(j)$ denotes whole image actual people-count. The PIR and PCE information looks the same, but they invoke different yet relevant and effective behavior in the network because of different operational inputs being used for their calculation. In addition, the nature of both outputs differs as the PIR is probability-based, while the PCE directly regresses the crowd-count patch-wise. Next, the PIR_{PRE} and PCE pass through the linear-layer and softmax to produce the attention-weights (AW_{AVT}). The AW_{AVT} is then applied on the original visual features (V) to give the PIR-PCE attended AVT output ($AV_{ATTD} \in \mathbb{R}^{P*Z}$), which acts as the third modality to be used in the next steps. This unique AVT strategy helps the network in focusing more on image regions with higher crowd-number and ignore the background patches. More importantly, the auxiliary mid-network PIR-PCE outputs aid both earlier and later-stage layers learning during the training process, and thus, resulting in significant improvement as demonstrated in experiments Sec. 5.5.

5.3.4 Cross-Modality Co-attention Module (CCM)

The co-attention module exploits the visual features (V) to perform the image-level crowd-estimation by jointly considering the audio features (A) and PIR-PCE attended channels (AV_{ATTD}). The transformer-inspired attention process is shown in Fig. 5.5(b) and defined as:

$$DM_{PRE} = softmax(AV_{ATTD} * A_{EXTD}) * V \tag{5.11}$$

where $DM_{PRE} \in \mathbb{R}^{P*Z}$, and A_{EXTD} is the $(Z \times P)$ matrix containing P times repeated vector A.

5.3.5 Final Crowd Estimate (CE_{FINAL})

The DM_{PRE} gets re-shaped and up-sampled 8× to output the final crowd Density-Map ($DM \in \mathbb{R}^{W*H}$) as shown in Fig. 5.5(b). We sum all DM pixel-values to obtain the final crowd estimate (CE_{FINAL}) for the input image-audio. We deploy L_2-norm as the DM loss-function, given as:

$$Loss_{DM} = \sum_{m=1}^{W} \sum_{n=1}^{H} (DM_{mn} - DM'_{mn})^2 \tag{5.12}$$

where $DM \in \mathbb{R}^{W*H}$, $DM' \in \mathbb{R}^{W*H}$ indicate estimated and ground-truth density-maps, respectively. The network total multi-task loss ($Loss_{TOTAL}$) will be as follows:

$$Loss_{TOTAL} = Loss_{PIR} + Loss_{PCE} + Loss_{DM} \tag{5.13}$$

Unlike other existing audio-visual mechanisms [31], our scheme employs both global and local learning (inter-pixel and inter-patch) in an explicit manner with the joint consideration of audio features, which empirically improves network performance significantly. It also helps in suppressing background regions at pixel, patch, and image-level.

5.4 Evaluation and Training Details

The only available audio-visual crowd counting dataset to-date (DISCO) [31] contains images with the same 1920 × 1080 resolution. As per convention, we resize them to 1024 × 576 for better resources usage. Consequently, (C_{S1}, C_{S2}, C_{S3}) channels have $(32 \times 256 \times 144)$, $(64 \times 128 \times 72)$, $(128 \times 64 \times 36)$ dimensions respectively. Therefore, we have 64 image patches in total (i.e. $P = \frac{W}{128} * \frac{H}{72} = 8 * 8 = 64$), and the value of Z is set to 144. In case of low-resolution setting experiments, we have 128 × 72 size input images as per the norm. During this setting, we train without any down-sampling in the initial convolution layers, giving dimensions of (C_{S1}, C_{S2}, C_{S3}) as $(32 \times 128 \times 72)$, $(64 \times 64 \times 36)$, $(128 \times 32 \times 18)$ respectively and $P = \frac{W}{32} * \frac{H}{18} = 4 * 4 = 16$.

To generate the ground-truth density map, we apply the 15 × 15 Gaussian kernel $(G \sim \mathcal{N}(0, 4.0))$ on binary annotations, where the ground-truth annotations are available in terms of people head center location in the image. We employ Adam optimizer [47] and the learning rate with an initial value of $1e{-5}$ that decays by 0.99 every epoch with total 500 epochs. The training batch size is set to 4 and model evaluation takes place after every epoch. To mitigate over-fitting, linear-layers are followed by the dropout layer with the drop-probability of 0.3, and weight-decay ($\lambda = 1e{-4}$) has been used.

Evaluation Details. We evaluate and compare our method with the state-of-the-art using standard evaluation metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), defined as follows:

$$MAE = \frac{1}{N}\sum_{n=1}^{N} |E_n - C_n|, \; RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (E_n - C_n)^2} \tag{5.14}$$

where C_n and E_n indicate the ground-truth and estimated crowd for the test audio-image input n respectively, and N denotes the total test audio-image samples in the dataset.

5.5 Experiments

We first discuss the numerical evaluation on the audio-visual and vision-only benchmark datasets, followed by the ablation study and visual analysis.

Method	Regular		Low Res.		Gaussian Noise				Low Illum. & Gaussian				Avg. Score	
			128×72		$\sigma = 25/255$		$\sigma = 50/255$		R=0.2,B=25		R=0.2,B=50			
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCN [127]	53.4	84.1	60.1	89.3	53.4	84.0	53.9	84.0	70.7	96.11	70.5	96.1	60.3	88.9
CAN [59]	15.4	28.9	22.1	39.6	13.3	27.2	14.2	28.0	26.0	49.1	33.1	58.2	20.7	38.5
CSR [54]	13.8	28.7	17.1	30.6	13.7	28.0	14.5	29.1	35.7	62.76	45.8	75.4	23.5	42.4
ACSR [31]	14.2	28.0	16.8	31.4	13.0	27.4	13.7	28.6	25.0	51.5	27.3	45.1	18.3	35.4
CC-V	12.9	25.7	16.9	32.8	13.3	28.7	13.9	29.0	26.0	55.7	27.5	58.6	18.4	38.4
CC-AV	**9.24**	**19.8**	**11.1**	**26.2**	**10.1**	**19.7**	**10.3**	**19.7**	**20.1**	**44.5**	**21.1**	**40.8**	**13.7**	**28.5**
Boost (%)	**33.4**	**29.4**	**33.8**	**14.3**	**22.3**	**27.4**	**24.2**	**29.4**	**19.6**	**9.2**	**22.5**	**9.5**	**25.4**	**19.5**

Table 5.1: Quantitative Evaluation on the DISCO Benchmark [31] based on regular and several low-quality images settings. (Here R, B denote the hyper-parameters being used for the illumination decay-rate and Gaussian-noise standard deviation computations respectively as defined in [31])

5.5.1 Experiments on Audio-Visual Dataset

DISCO [31] is an up-to-date and only-available diverse audio-visual crowd dataset. It contains a total of $1,935$ high-resolution images ($1,920 \times 1,080$) and corresponding one-second audio signals. We have $170,270$ people annotations in total with the minimum, maximum and average people per image equal to 1, 709, and 88, respectively. The (train, validation, test) split is pre-defined as (1435, 200, 300) respectively. We evaluate our network using both audio-visual and vision-only versions. Audio-visual (CC-AV) version is the same as discussed above, whereas the vision-only variant (CC-V) only uses the image input and is detailed in sub-section 5.5.2. As per the standard practice, we compare the proposed scheme with the state-of-the-art for three pre-defined image settings.

Regular Images. In this case, we use test images without any modification. The results, as shown in Table 5.1, indicate that both proposed network versions outperform the state-of-the-arts under all evaluation metrics with CC-AV giving 33.4% and 29.4% error decrease for the MAE and RMSE metrics respectively. CC-AV performs significantly better than the CC-V, which directly implicates the benefit of including the audio modality.

Low-Quality Images. To check the robustness under severe conditions, we evaluate the model on three pre-defined standard settings: low-resolution, low-illumination, and strong noise. In the low-

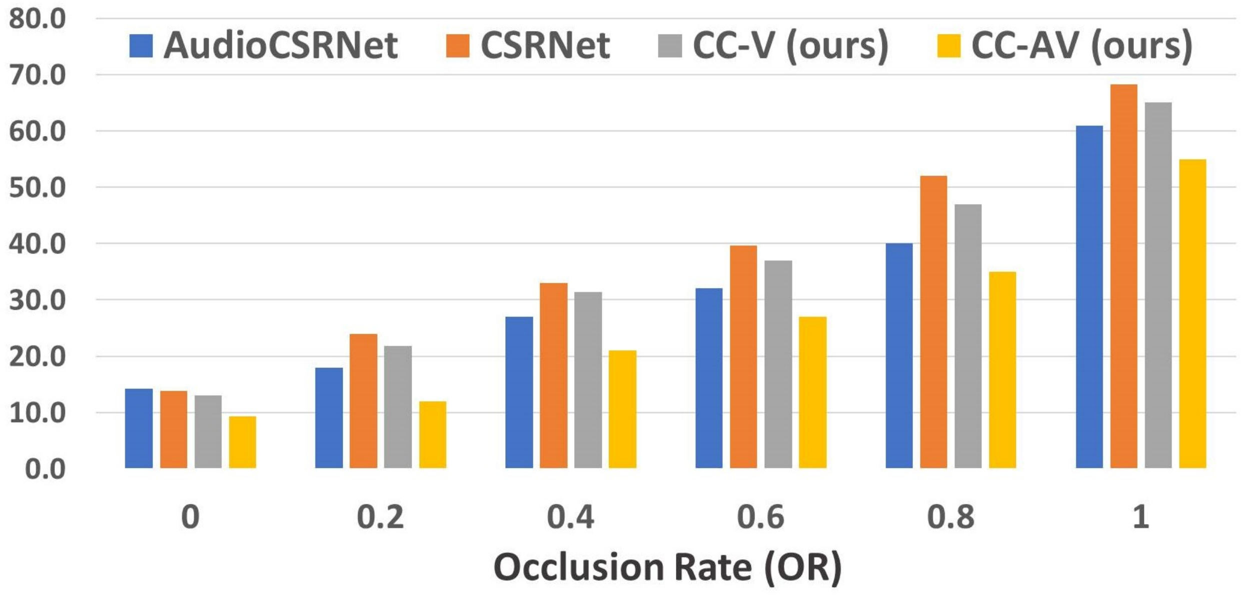


Figure 5.6: Occluded images setting based evaluation using MAE metric.

resolution setting, images are just 128 × 72 in size. During the low-illumination study, random brightness reduction is followed by the Gaussian noise addition as defined in [31]. Lastly, the Gaussian noise has been added in the strong noise case as given in [31]. Observing the results for all three settings, as shown in Table 5.1, the proposed model (CC-AV) appears as the best choice with improvement up to 33.8% and 29.4% for MAE and RMSE respectively. The CC-V variant performance decreases in such extreme conditions because the visual information alone proves insufficient without any further aid.

Occluded Images. In this setting, we occlude the image with a black rectangle using the Occlusion Rate (OR). The OR value lies in [0,1], meaning that image occlusion ranges from no occlusion ($OR = 0$) to completely occluded ($OR = 1$). The results, as shown in Fig. 5.6, show that the CC-AV model gives the best performance for the whole OR range as compared to the state-of-the-art methods (AudioCSRNet[31] and CSRNet [54]) on the MAE metric. All methods experience performance degradation as we increase the OR value due to the lack of more visual information. Our CC-V model yields a bigger error jump than CC-AV with the increase of OR values because it only relies on the visual information. Interestingly, in the case of no visual information (OR=1),

	ShanghaiTech [127]		UCF-QNRF [35]	
Method	MAE	RMSE	MAE	RMSE
MCNN [127]	110.2	173.2	277	426
Switch-CNN [85]	90.4	135.0	228	445
CSRNet [54]	68.2	115.0	-	-
CL[35]	-	-	132	191
CAN [59]	62.3	100.0	107	183
RRP [13]	63.2	105.7	93	156
HA-CCN [95]	62.9	94.9	118.1	180.4
ADSCNet [6]	**55.4**	97.7	**71.3**	132.5
RPNet [119]	61.2	96.9	-	-
PRM-based[83]	67.8	86.2	94.5	141.9
CC-V (Ours)	58.7	**81.3**	75.4	**125.6**

Table 5.2: MAE and RMSE based evaluation on image-only datasets.

CC-AV still performs better, indicating its robustness and better utilization of the audio-modality as compared to the best audio-visual models.

5.5.2 Experiments on Image-only Datasets

First, we discuss the design of the image-only variant (CC-V) of the proposed network. The CC-V structure remains the same as the CC-AV except that there is no available audio information (A) and thus the following changes have been made. 1) No A based operation in the MSB three-branch fusion, PIR, PCE, and Co-attention processes. 2) Matrix operations required to compute PIR_{PRE} have been replaced by the same set of operations being used for PCE. 3) Replace A_{EXTD} with V^T in the co-attention module.

We compare our CC-V model on two image-only diverse benchmark datasets: UCF-QNRF [35] and ShanghaiTech Part-A [127]. The UCF-QNRF dataset comprises of $1,535$ ($1,201$ train, 334 test) images with total $1,251,642$ people annotations. On the other hand, ShanghaiTech dataset contains a diverse collection of 482 crowd images (300 train, 182 test). To avoid over-fitting in the case of ShanghaiTech dataset training, we use the model pre-trained on the UCF-QNRF benchmark, and train for only 250 epochs instead of 500. The images have been resized to $1,024 \times$

Ablation Setting	MAE	MAE Increase (%)	RMSE	RMSE Increase (%)
W/o explicit PIR, PCE	16.0	42.3	28.2	29.8
W/o PIR branch in AVT	16.7	44.7	27.6	28.2
W/o PCE branch in AVT	15.5	40.4	28.1	29.5
W/o AVT	18.9	51.1	39.7	50.1
W/o CCM	17.3	46.6	31.4	36.9
W/o A^T in the MSB fusion	13.9	33.5	25.0	20.8
W only S1 branch in MSB	14.3	35.4	26.8	26.1
Default (CC-AV)	**9.24**	-	**19.81**	-

Table 5.3: Seven independent ablation studies on the effect of PIR, PCE, MSB, AVT and CCM components on the proposed network performance.

576 with zero-padding if required. The results on both datasets are shown in Table 5.2, where the proposed model CC-V yields the best performance for the RMSE metric (5.2% improvement for UCF-QNRF and 5.7% for ShanghaiTech) amid producing reasonable results for the MAE as compared to the state-of-the-art schemes. These results demonstrate that the proposed scheme is also practical, robust, and highly effective in vision-only scenarios.

5.5.3 Ablation Study

In addition to the previous sub-section 5.5.1 analysis on audio-visual DISCO dataset various settings, here we further analyze and investigate the effect of different components on overall network performance during the following independent ablation studies.

W/o explicit PIR, PCE. No PIR vector output as well as no $Loss_{PRE}$ and $Loss_{PCE}$, i.e. $Loss_{TOTAL} = Loss_{DM}$.

W/o PIR or PCE branch in AVT unit. In the first setting, we exclude the whole PIR computation stream and $Loss_{PIR}$, and only use the PCE stream and vector. In the second setting, we do vice versa by only keeping the PIR stream, and $Loss_{TOTAL} = Loss_{PIR} + Loss_{DM}$.

W/o AVT. No AVT module being deployed. Consequently, the CCM block uses V instead of AV_{ATTD}.

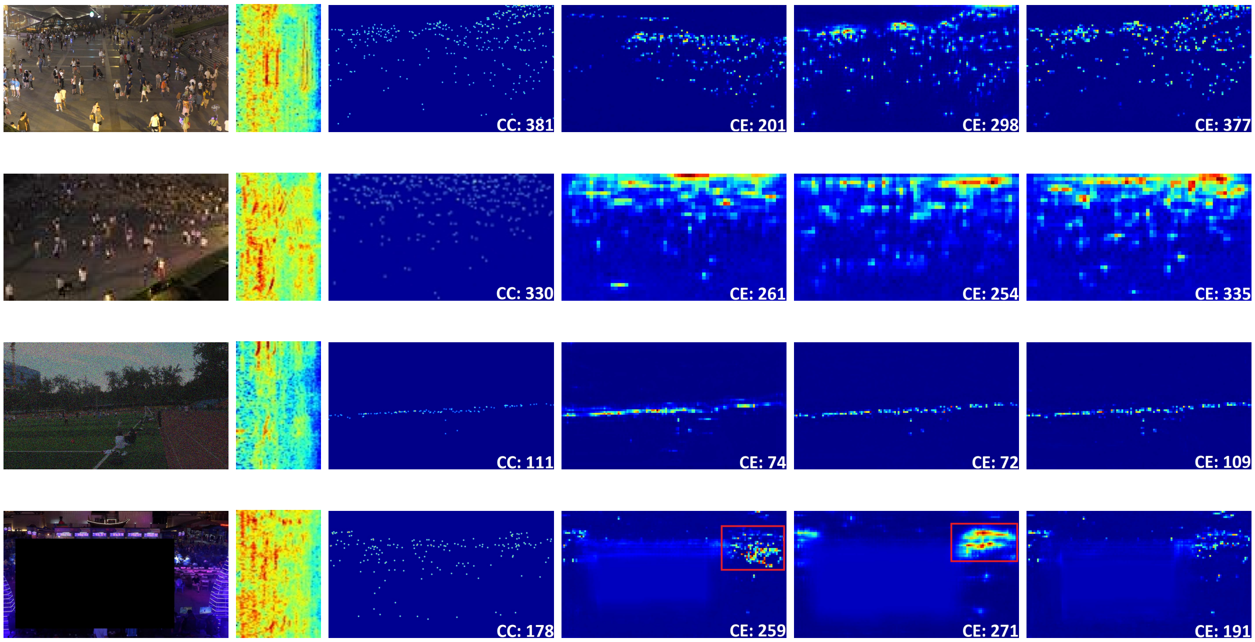


Figure 5.7: Ground truth (GT) density-map and crowd-count (CC) based qualitative comparison (From Left to Right Column: Input Image, Audio Log Mel-Spectogram, GT density-map, AudioCSRNet network [31] estimated density-map, CC-V model (ours) density-map, CC-AV (ours) density-map). Each row respectively indicates following four cases: Regular Image (1024×576), Low-Resolution (128×72), Noisy case ($\sigma = 50/255$), and 50% Occluded case.

W/o CCM. No CCM module usage. The AV_{ATTD} is considered as DM_{PRE}.

W/o A^T in the MSB fusion. No Audio information (A) has been used in any MSB three-branch fusion process.

Using only single (S1) branch. We only use one (S1) branch in the MSB multi-branch structure.

The results are listed in Table 5.3, where we can observe that the (MAE,RMSE) errors increase by a noticeable margin in each case with as low as (33.5%,20.8%) and as high as (51.1%,50.1%) respectively. These evaluations indicate the effective importance of several network components including PIR, PCE, MSB, AVT, and CCM modules.

5.5.4 Qualitative Analysis

We present a few visual results as shown in Fig 5.7. These results contain both regular (top row) and low-quality image cases (last three rows). For each input image, we display the input image, Log Mel-Spectrogram (LMS), ground-truth crowd density-map and count (CC) as well as predicted density-map and crowd-estimate (CE) being generated by our CC-AV, CC-V networks, and state-of-the-art AudioCSRNet [31]. We can easily observe that the proposed audio-visual model (CC-AV) yields the most effective and fine-grained results as compared to the visual-only variant (CC-V) and AudioCSRNet [31] in both regular and low-quality cases. However, the CC-V model experiences more error increase in low-quality cases due to lack of audio modality. These results also demonstrate that the proposed CC-AV network has significantly improved performance because of the better inclusion of the audio modality. Interestingly, the CC-AV performance is naturally better for regular images as visual information fades away in low-quality cases. One mentionable case is that of 50% random image occlusion (last row of Fig. 5.7). CC-V highly over-estimates in the non-occluded regions (highlighted in the red rectangular area) to compensate for the occluded area, and lacks the audio-modality aid to better estimate for the hidden region. Similarly, AudioCSRNet [31] also over-estimates in the same manner due to under-utilization of the audio information. On the other hand, our CC-AV model performs more robustly for both occluded and non-occluded regions.

5.6 Conclusion

In this work, we have presented a new audio-visual multi-task network for effective people counting by introducing explicit PIR and PCE information for better modalities association, and also producing a third run-time modality. This modality greatly helps the cross-modality fusion process to yield a better crowd estimate. We have also deployed a unique multi-branch structure to extract rich visual features and also proposed the image-only variant of our model. Experimental evaluation on standard benchmarks reveals the superior performance of our networks.

Chapter 6

Application of Multi-scale Crowd Counting network to the Scene-text Recognition

Here, we apply multi-scale fusion and attention-based crowd counting technique as given in Chapter 4 to another important and active computer vision task i.e. Scene Text Recognition. We propose a new text recognition network for scene-text images. Many current state-of-the-art methods employ the attention mechanism either in the text encoder or decoder for the text alignment. The encoder-based attention yields promising results, but these schemes inherit noticeable limitations. They perform the feature extraction (FE) and visual attention (VA) sequentially, which bounds the attention mechanism to rely only on the FE final single-scale output. Moreover, the utilization of the attention process is limited by only applying it directly to the single scale feature-maps. To address these issues, we propose a new multi-scale and encoder-based attention network for text recognition that performs the multi-scale FE and VA in parallel. The multi-scale channels also undergo regular fusion with each other to develop the coordinated knowledge together. Quantitative evaluation and robustness analysis on the standard benchmarks demonstrate the better effectiveness of the proposed network in most cases. Most content of this chapter comes from our paper [80].

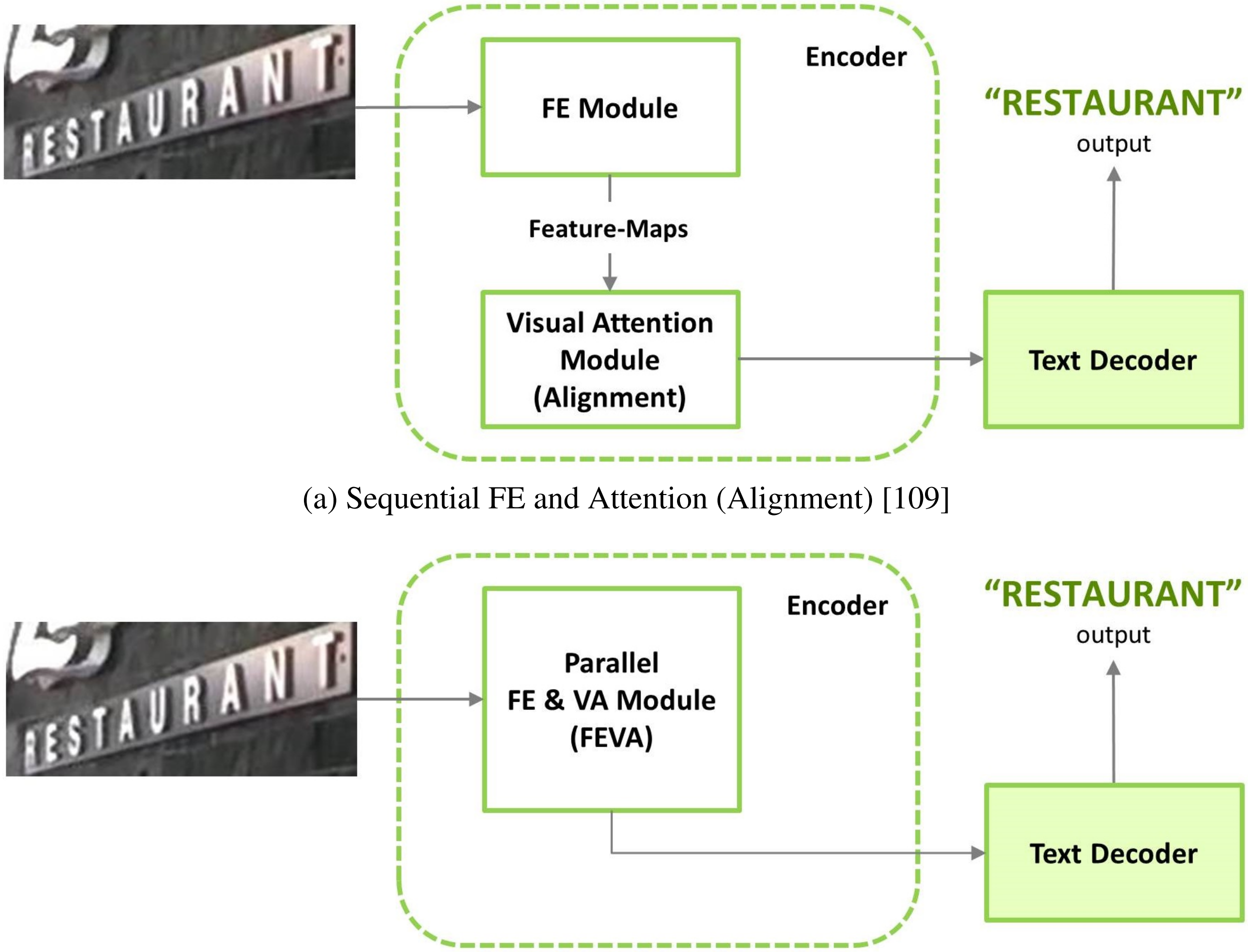
6.1 Introduction

Scene text recognition aims at extracting the screen text from the given input image. It serves as an active research area in the field of computer vision. The recognition task comes up with many key challenges and issues like huge background variation in and across different images, different font styles, big fluctuation in text appearance and scale. Automated text recognition remains more desirable as manual system proves to be very tedious and time-consuming. Recently, deep learning-based automated methods have shown superior performance in this domain. Some schemes perform character-level text recognition, while most methods do word/sentence level recognition. The later one is more preferred due to relatively easier and less tedious annotation process.

Among the best state-of-the-art deep networks, most of them [109, 121, 15, 88, 90, 5, 51] are based on the attention mechanism [4, 101]. The purpose of the attention mechanism is to align the text characters followed by their recognition. Generally, these methods incorporate the attention-based alignment and recognition into the decoder part of the network. But these networks inherit an important limitation as the decoder gets highly over-burdened and sensitized with the dual task of text alignment and recognition. Consequently, it generates huge error propagation and aggregation within the decoder and thus compromises the effectiveness of the whole network. One possible solution is to decouple the attention/alignment mechanism from the decoder and integrate it with the feature extraction process inside the encoder block of the network. Recently, Wang et. al. [109] proposed such decoupled attention network (DAN) with promising results. However, the encoder first sequentially performs the feature extraction (FE) followed by the visual attention (VA) process as shown in Fig. 6.1(a). This limits the DAN network efficacy as the attention mechanism only depends on and utilizes the final output feature-maps from the FE module. Consequently, the attention is not applied directly to each of the multi-scale feature-maps separately, but only to the final set of accumulated single-scale channels. Therefore, our focus revolves around two main objectives in this work:

- Design a scale-wise visual attention-based scene text recognition network to address the key issues and challenges in this domain.

- Utilize the encoder-based and scale-wise attention process in parallel to the feature extraction (FE) instead of standard sequential processing from FE to the Visual attention module.



(a) Sequential FE and Attention (Alignment) [109]

(b) Parallel scale-wise FE and Alignment

Figure 6.1: (a) Recent state-of-the-art Encoder-based attention mechanism [109] sequentially performs the feature extraction (FE) followed by the (attention) alignment module. (b) The proposed method performs the parallel FE and visual attention (VA) process on feature-maps with different scales within the encoder.

In this work, we propose a new multi-scale and scale-wise visually attended text recognition network to achieve the above objectives. As shown in Fig. 6.1(b), the feature extraction and visual alignment/attention (FEVA) have been done in parallel on different scale features within a single module, followed by the recognition-focused decoder to extract the scene text. In this way, we separately attend feature-maps from different scales directly instead of just attending the final single-scale channels. Moreover, we also deploy different and simpler visual attention process in contrast to the conventional deep up- and down-scaling FCN [61] based visual attention being used in DAN [109]. Several experiments on different standard benchmark datasets demonstrate the effectiveness of our scheme on both regular and irregular scene-texts as presented in the experiments section 6.5.

6.2 Our Contributions

The main contributions of this work include:

- We propose a new parallel FEVA-based encoder and multi-scale text recognition network to address the key recognition task challenges and limitations in similar state-of-the-art architectures.

- We deploy the visual attention mechanism in an effective and unique way on multiple scales to enable the network in making a clearer distinction between the foreground and background pixels.

- Experimental evaluation on the standard benchmark datasets demonstrates the fact that the proposed network outperforms the state-of-the-arts in most cases on both regular and irregular scene-texts.

6.3 Related Work

Text recognition problem remains an active research area in the computer vision field due to different challenges like varying text scale and size, partial occlusion, and non-axis aligned text. In pre deep-learning era, document text recognition remained the main focus. [8] adopted the binarization process to extract the segmented text characters. But these methods are not applicable to scene-text due to different nature of issues like varying scale and style, and complex background. Most of the classical recognizers utilized the low-level information including the connected components [70], gradients descriptors (HoG) based on some feature-extraction mechanism [106]. Recently, deep-learning based method hugely surpass and outperform the traditional methods. They can be categorized as segmentation-based and segmentation-less text recognizers.

Segmentation-based methods undergo the character-wise detection followed by the word formation. [7] designed five hidden fully-connected layers and ReLU Units [69] with softmax-based classification. [108] developed an CNN with convolution and average pooling layers and the non-maximum suppression for character-wise text recognition. [38] used weight-shared CNN for three sub-tasks of dictionary, character sequence, and bag-of-N-gram encoding to perform the text recognition.

The segmentation-less schemes directly recognize the whole word or sentence from the given input image. [39] performed an CNN-based 90,000-way classification, where each category/class corresponds to one whole word. Shi *et al.* [88] integrated the CNN and RNN-based network to extract the string features, and the Connectionist Temporal Classification (CTC) based decoder to finally yield the recognized text. [89] employed the attention mechanism for text alignment before the recognition. Most following methods utilize the attention mechanism [4, 101] in one way or the other. Cheng *et al.* [15] designed the deep focused attention network (FAN) after observing and aiming to address the "attention-drift" problem in the recognition process, but it requires character-level annotations. [89, 63, 125] aimed at addressing non-axis aligned and distorted text via attention-based mechanism. [97, 2] utilized the recurrent neural networks (RNNs) with Long Short Term Memory (LSTM) networks to perform the sequential word recognition. [27] integrated the CNN and RNN to design the deep-text recurrent network (DTRN) for recognizing text. Shi *et al.* [90] explicitly handled the text rectification by using the control points based rectification module and also applied the attention-based bi-directional LSTM decoder for text prediction. Li *et al.* [51] proposed a simple LSTM-based encoder-decoder framework via the 2D attention process. Wang *et al.* [109] designed the decoupled attention network (DAN) that performed the text alignment via convolution based visual attention. Yu *et al.* [121] proposed the semantic reasoning

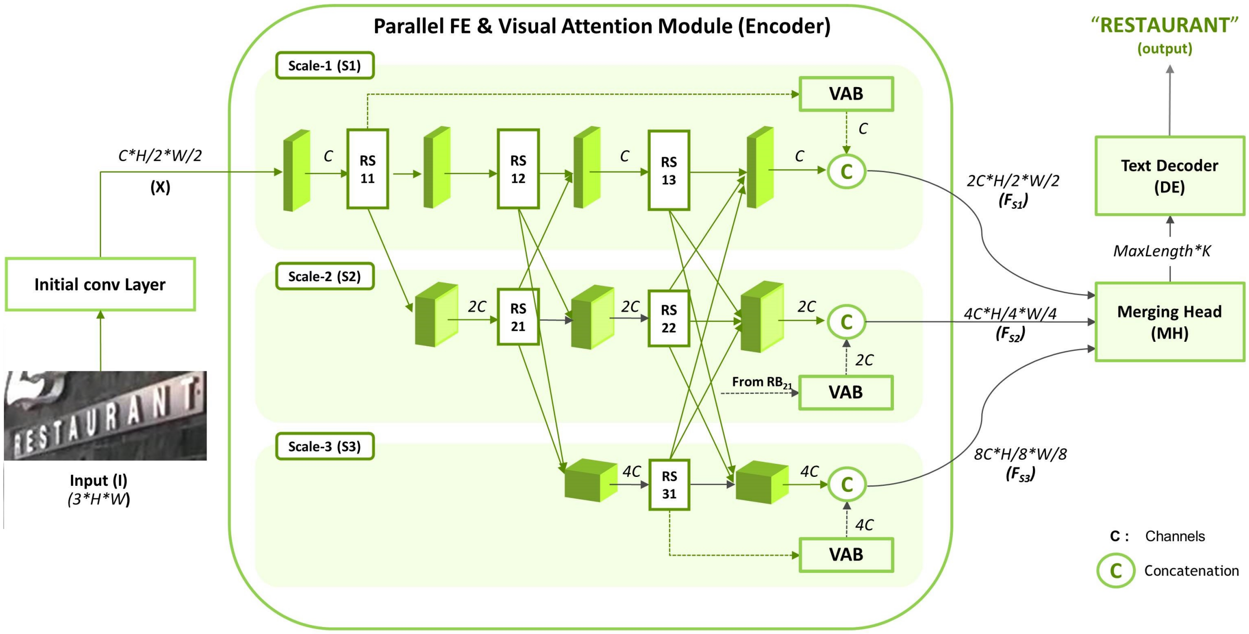


Figure 6.2: The proposed text recognition network. Initially extracted features (X) from the input image (I) first pass through the multi-scale feature-extraction (FE) and visual attention (VA) based encoder (EN). The encoder performs both intra-scale processing and inter-scale fusion between three scale-modules (S1,S2,S3). During the intra-scale processing, the channels are processed with the residual connections-based residual structures (RS) as well as undergo the VA mechanism via the visual attention block (VAB), followed by their concatenation to produce the respective scale-module output. Consequently, the encoder outputs three sets of feature-maps that go through the merging head (MH) for channel and resolution adjustment. Finally, the text decoder (DE) outputs the recognized text character-wise.

network (SRN) for irregular scene-text that fuses the visual attention and semantic context modules while avoiding the RNN-based sequential processing.

Although these schemes produce good results, yet they fail to utilize the promising and beneficial attention mechanism explicitly on different multi-scale features. In this work, we work towards utilizing the multi-scale feature-extraction and visual-attention in parallel for better efficacy.

6.4 Proposed Approach

The chapter proposes a new scene-text recognition network to address the major recognition challenges as detailed in Sec. 6.1, as well as performs the visual attention on multi-scale feature-maps directly in parallel. The proposed network, as shown in Fig. 6.2, downscales the input image I ($\in \mathbb{R}^{3*H*W}$) resolution by half ($C*\frac{H}{2}*\frac{W}{2}$) using the initial convolutional layer. The resultant feature-maps go through the text encoder (EN). The EN block comprises of three parallel multi-scale modules (S1,S2,S3) with each module handling one specific scale. Each multi-scale module also visually attends the feature-maps in parallel to their conventional deep layers-based processing, followed by the concatenation together to generate their respective output. The visual attention helps the model to have a clearer understanding between the foreground and background pixels. Inspired by the high-resolution networks [105, 99], these multi-scale modules also fuse their channels on regular intervals to develop the accumulated knowledge together. The encoder outputs three multi-scale channels (F_{S1},F_{S2},F_{S3}) that are merged together via the merging head (MH). The text decoder (DE) finally outputs the recognized text. The proposed network architecture consists of three major components: Text Encoder (EN), Merging Head (MH) and Text Decoder (DE) as detailed next.

6.4.1 Encoder (EN)

The purpose of the encoder is to simultaneously perform the feature extraction (FE) and visual attention/alignment (VA) on the multi-scale feature-maps. The input channels ($X \in \mathbb{R}^{C*\frac{H}{2}*\frac{W}{2}}$) pass through three multi-scale modules(S1,S2,S3) to finally yield three respective output feature-maps with different dimensions. The encoder processes the input feature-maps as follows:

$$(F_{S1},F_{S2},F_{S3}) = Encoder(X), \tag{6.1}$$

Where $F_{S1} \in \mathbb{R}^{2C*\frac{H}{2}*\frac{W}{2}}$, $F_{S2} \in \mathbb{R}^{4C*\frac{H}{4}*\frac{W}{4}}$, $F_{S3} \in \mathbb{R}^{8C*\frac{H}{8}*\frac{W}{8}}$ and C indicates the total number of input channels. As we move from S1 to S3, the number of channels become twice as many as their subsequent upper scale. Similarly, the feature-maps resolution (scale) decreases to half with each scale-module as we move from S1 to S3. It may be noted that each scale module keeps the channel resolution same throughout that module [99, 105].

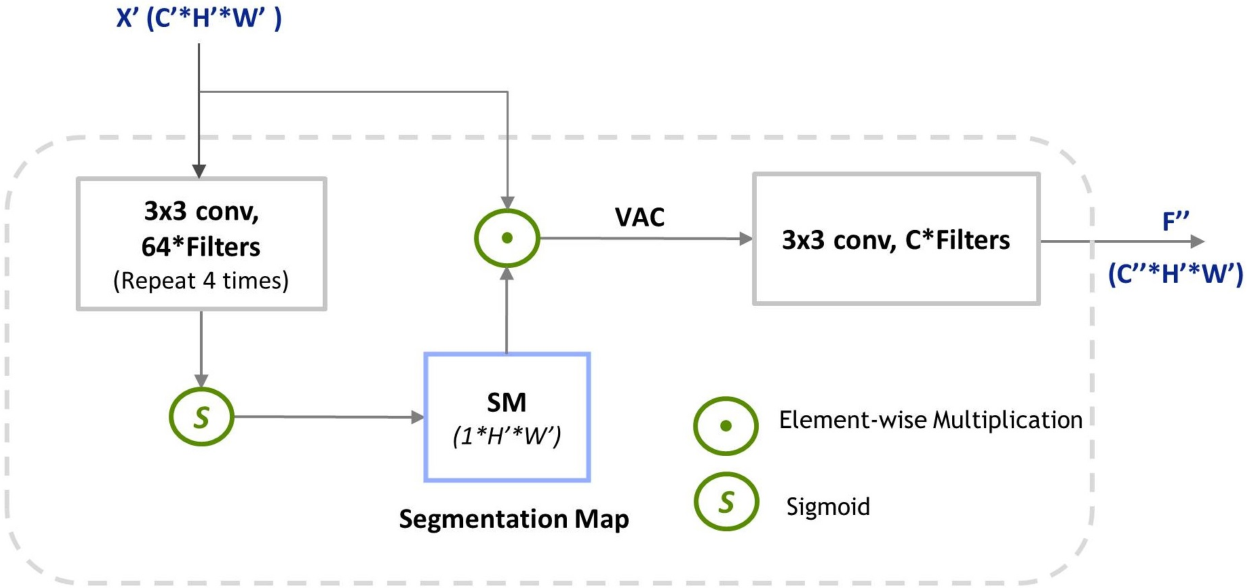


Figure 6.3: Visual Attention Block (VAB). The input channels go through the convolution operation for four repeated times. Subsequent single-filter 1×1 convolution and sigmoid function give the segmentation map (SM). The SM undergoes the element-wise multiplication with the original input channels to yield the visually attended channels (VAC) that are channel-adjusted to become the VAB output.

6.4.1.1 Intra-Scale Processing

Within every scale module (S1,S2,S3), the input channels pass through one or more residual structures (RS) and the visual attention process.

Residual Structure (RS). Each RS block comprises of five residual units (RU). The RU unit is a 3-layered residual building block as given in [26] that contains three convolution layers ($1 \times 1, 3 \times 3, 1 \times 1$) and a residual connection. Each convolution in the chapter is followed by the Batch-Normalization (BN) [36] and the ReLU activation [69] unless stated otherwise. The RS blocks are denoted as *RS(xy)*, where *x* denotes the scale-module number (1,2 or 3) and *y* indicates their location or index within that module (starting from left to right). Thus, *RS12* denotes the second RS block in the S1 scale-module.

Visual Attention Block (VAB). The scale-modules (S1,S2,S3) also visually attend (align) their feature-maps independently. This helps the network in making a better understanding regarding the foreground and background image pixels at different feature-scales. The first RS block output

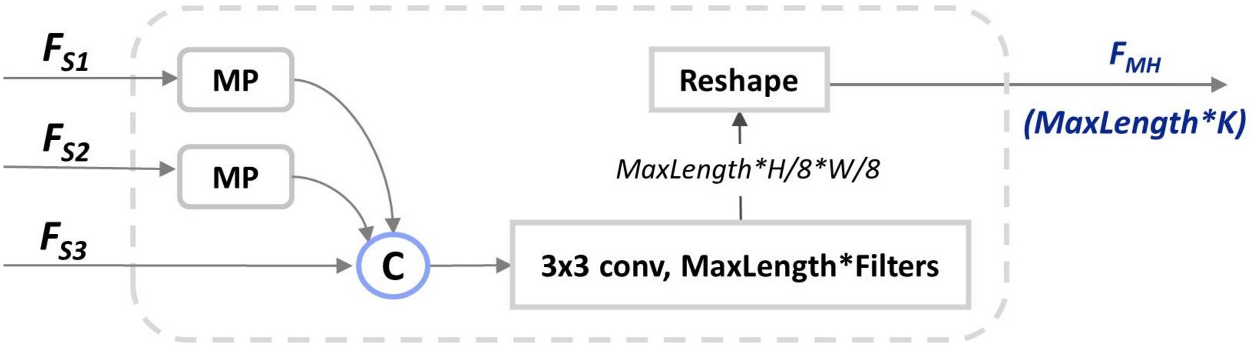


Figure 6.4: Merging Head (MH). The higher scale-modules (S1,S2) output channels (F_{S1}, F_{S2}) are max-pooled before concatenation with the S3 scale-module output. Convolution and reshaping operations finally output the K-dimensional vectors with *MaxLength* such vectors in total.

channels ($\in \mathbb{R}^{C'*H'*W'}$) in any scale-module undergo the attention mechanism via the *VAB* block. The attended feature-maps are then concatenated back at the end of the respective scale-module. The VAB process is shown in Fig. 6.3, where the input feature-maps first go through the four consecutive convolution layers. Next, a single feature-map is obtained via a simple 1×1 convolution operation. The sigmoid function is then applied on the resultant channel to obtain the segmentation map ($SM \in \mathbb{R}^{1*H'*W'}$). The SM undergoes element-wise multiplication with the original input feature-maps to yield the visually-attended channels (VAC). The VAC feature-maps serve as the VAC module final output after being channel-adjusted via the 3×3 convolution operation. The VAB input feature-maps $X' \in \mathbb{R}^{C'*H'*W'}$ get the visual attention as follows:

$$F'' = VAB(X'), \tag{6.2}$$

Where $F'' \in \mathbb{R}^{C''*H'*W'}$ and we set $C'' = C'$. This attention process is different from the conventional and complex convolutional and deconvolutional layers based mechanism [109], and proves to be more effective as demonstrated in the experiments Sec. 6.5.

6.4.1.2 Repetitive Inter-scale Fusion

Inspired by the high-resolution networks [105, 99], the scale-modules (S1,S2,S3) also fuse channels with each other on regular intervals. It enables the network to form the accumulated and coordinated knowledge from the multi-scale channels and learn the valuable information better. To fuse the higher-scale source channels into the lower-scale target feature-maps, they undergo the $(n+1)$ times 3×3 convolution operation (with stride: 2, padding: 1). Here n $(= 0, 1)$ denotes the number of scale-modules in-between the source and target scale-modules. Thus, fusion from S1 channels into S3 requires two such convolution operations on S1 scale feature-maps to down-scale them to the S3 scale. Similarly, the lower-to-higher scale fusion requires the bilinear upsampling of the lower-scale source feature-maps. No re-scaling transformation is done when the source and target scale-modules are same. Once all source channels have been adjusted for channels quantity and target scale, they undergo the summation-based fusion with the target channels to obtain the fused feature-maps.

6.4.2 Merging Head (MH)

The encoder outputs three separate sets of feature-maps (F_{S1}, F_{S2}, F_{S3}) from the respective scale-modules (S1,S2,S3). The merging head (MH) combines them to output the feature-maps to be used for the text decoding. The MH block, as shown in Fig. 6.4, down-samples the S1 and S2 output channels using the max-pooling (MP) operation, so as to rescale them to the S3 output channels (F_{S3}) resolution. Next, they are concatenated together followed by the channel-adjustment via the convolutional layer. The resultant channels $(\in \mathbb{R}^{MaxLength * \frac{H}{8} * \frac{W}{8}})$ are reshaped into the K-dimensional vectors to give the MH final output $(F_{MH} \in \mathbb{R}^{MaxLength*K})$. Thus, the input channels are merged as follows:

$$F_{MH} = MH(F_{S1}, F_{S2}, F_{S3}), \tag{6.3}$$

Here, the maxLength refers to the maximum length of text characters to be recognized. The output vectors are then routed to the text decoder for further processing.

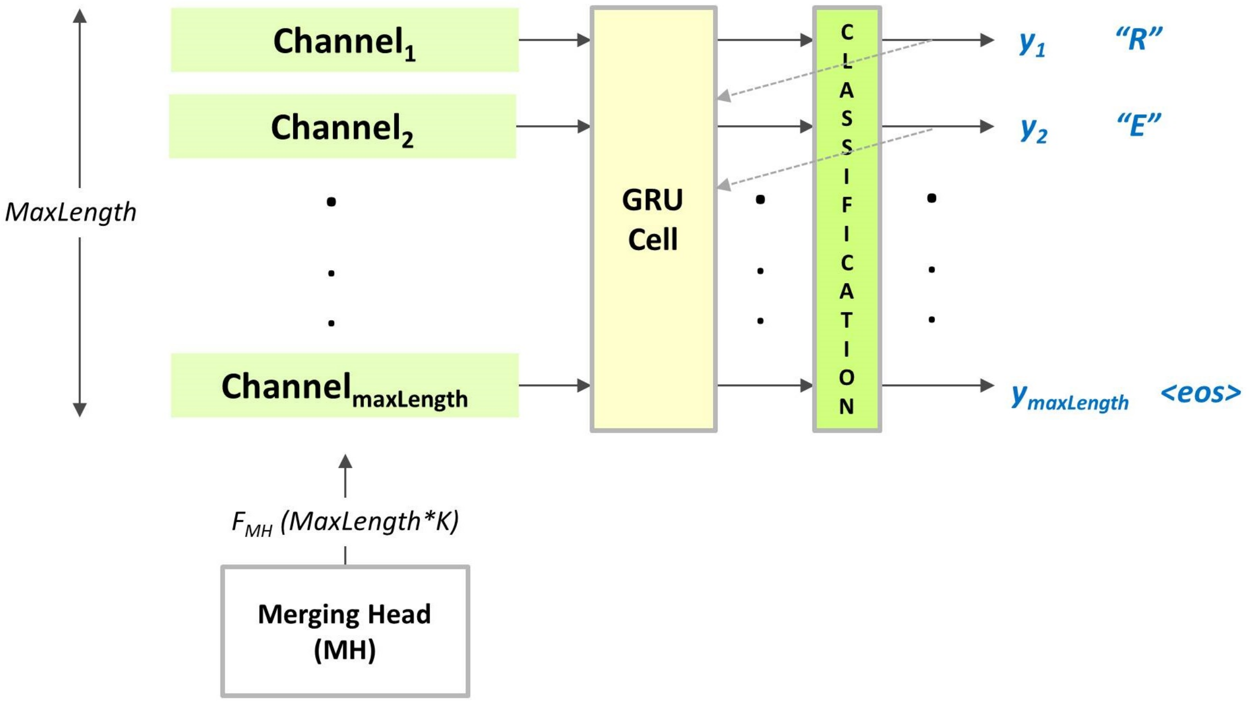


Figure 6.5: Text Decoder (DE). The decoder predicts the recognized text character-by-character via the GRU and the classification layers. Here *eos* means the end-of-sequence character.

6.4.3 Text Decoder (DE)

The responsibility of our text decoder is to perform recognition only. That makes it more focused towards one task rather than the dual task of text alignment and recognition. We adopted the text decoder from the DAN network [109]. As shown in Fig. 6.5, the MH output channels ($F_{MH} \in \mathbb{R}^{maxLength*K}$) go through the GRU [17] cell one-by-one at time ($t' = 1,2,3,...maxLength$) as K-dimensional vectors. The classification layer outputs the recognized text character at time t' with the output $p(y_{t'})$ as follows:

$$p(y_{t'}) = softmax(w * hidden_{t'} + b), \tag{6.4}$$

Where $hidden_{t'}$ denotes the GRU cell hidden state, given as follows:

$$hidden_{t'} = GRU((embd_{t'-1}, Channel_{t'}), hidden_{t'-1}), \tag{6.5}$$

Where $embd_{t'-1}$ is the embedding belonging to the previous classification $y_{t'-1}$. The network loss function is defined as follows:

$$L = -\sum_{t'=1}^{T'} \log P(y'_{t'} | Input, \theta) \tag{6.6}$$

Where $y'_{t'}$ is the actual or ground-truth text character at time t' and θ denotes the learnable parameters of the network.

6.5 Quantitative and Qualitative Evaluation

This section deals with the experimental evaluation of the proposed network. First, we discuss the quantitative evaluation on seven standard benchmark datasets followed by the ablation study. At the end, we analyze the visual results.

6.5.1 Experiments on Standard Benchmarks

Datasets. To evaluate the effectiveness of the proposed network, we test on seven different scene-text datasets. They are either regular (IIIT-5k [67], IC03 [62], IC13 [44], SVT [106]) or irregular (IC15 [43], SVT-P [73], CUTE80 [77]) scene-text datasets.

IIIT-5k [67] is an internet-based scene-text dataset that contains 3,000 cropped text images for testing.

Street View Text (SVT) [106] comprises of 647 text-based test images collected via Google Street View. For diversity and variation, drastic corruption has been incorporated in the form of noise, blurriness and low resolution.

ICDAR 2003 (IC03) [62] has 251 scene-text images with 867 test bounding boxes. As per the standard protocol [106], 860 cropped images have been retained after removing words with non-alphanumeric characters or less than 3 characters.

ICDAR 2013 (IC13) [44] is a regular scene-text dataset that contains total 1,015 cropped images, and most of them come from the IC03 dataset. Using the standard practice as given in

[106], images with non-alphanumeric characters or less than three characters have been filtered-out.

ICDAR 2015 (IC15) [43] contain irregular scene-text images taken via the Google Glasses with slight focusing and positioning. Only 1,811 test images have been utilized after removing some with extreme distortions as part of the standard pre-processing practice [15].

SVT-P [73] is an irregular scene-text dataset with 639 cropped images taken from Google street view. Mostly, they are single-angle based and highly perspective-distorted images.

CUTE80 [77] mainly deals with curved scene-text and consists of 80 images. We cropped 288 test samples from these high-resolution images using their bounding-box annotations.

Implementation Details. The input image gets resized with fixed height of 32 pixels and width up to 128 based on the aspect ratio. The proposed network is trained using two synthetic datasets until convergence: Synth90k [38] and SynthText [24]. Batch size of 64 has been used with 32 images each from Synth90k and SynthText. The value of total channels (C) in the encoder has been set to 32, so the scale-modules (S1,S2,S3) contain (32,64,128) channels respectively after every intra-scale processing step. MaxLength is set to 25, and the total number of character classes is 94 including the upper- and lower-case alphabets, 0-9 digits, and 32 ASCII punctuation symbols. The number of hidden units in the decoder equals to 256. The ADADELTA-based optimization [124] has been employed with the initial learning rate of 1.0 and decreased to 0.1 after third epoch.

Experimental Evaluation. Here, we compare our method quantitatively with the current state-of-the-art models. The comparison is done without using the lexicon information as it is generally the case in practice. As per the standard convention, the evaluation is done using the case-insensitivity for word accuracy computation. The results are shown in Table 6.1, where our method outperforms other methods on 4 out of 7 datasets while performing reasonably competitive on the remaining three benchmarks. In comparison to the specifically designed rectification-based methods [90, 125, 63], our model gives better or competitive results without any rectification. For the regular scene-text dataset (IIIT-5K and IC13), we obtain an increase of (0.8% and 1.1%) respectively. While for the irregular scene-text datasets (IC15 and SVT-P), the proposed network improves the accuracy

	Rect.	Regular Datasets				Irregular Datasets		
Method		IIIT-5K	SVT	IC03	IC13	IC15	SVT-P	CUTE80
Jaderberg *et. al*[39]		-	80.7	93.3	90.8	-	-	-
Jaderberg *et. al*[37]		-	71.7	89.6	81.8	-	-	
Shi *et. al*[88]		81.2	82.7	91.9	89.6	-	-	-
Lyu *et. al*[64]		94.0	90.1	94.3	92.7	76.3	82.3	86.8
Xie *et. al*[113]		82.3	82.6	92.1	89.7	68.9	70.1	82.6
Liao *et. al*[56]		91.9	86.4	-	91.5	-	-	-
Cheng *et. al*[15]		87.4	85.9	94.2	93.3	70.6	-	-
Cheng *et. al*[16]		87.0	82.8	91.5	-	68.2	73.0	76.8
Bai *et. al*[5]		88.3	87.5	94.6	94.4	73.9	-	-
Yang *et. al*[118]		-	-	-	-	-	75.8	69.3
Shi *et. al*[90]	✓	93.4	89.5	94.5	91.8	76.1	78.5	79.5
Zhan *et. al*[125]	✓			-	91.3	76.9	79.6	83.3
Yang *et. al*[117]		93.3	90.2	91.2	93.9	78.7	80.8	87.5
Li *et. al*[51]		91.5	84.5	-	91.0	69.2	76.4	83.3
Liao *et. al*[55]		93.9	90.6	-	95.3	77.3	82.2	**87.8**
Wang *et. al*[109]		94.3	89.2	**95.0**	93.9	74.5	80.0	84.4
Yu *et. al*[121]		94.8	**91.5**	-	95.5	82.7	85.1	**87.8**
Ours		**95.9**	90.8	94.6	**96.3**	**83.9**	**86.0**	86.9

Table 6.1: Quantitative Evaluation on the Standard Benchmarks. The results demonstrate the proposed scheme as the most effective in most cases as compared to SOTA methods. The bold and underline numbers indicate the best and the second-best methods respectively.

by (1.4% and 1.0%) respectively . The accuracy boost is mainly due to the inclusion of multi-scale visual attention and inter-scale fusion within the encoder. It is empirically shown during the ablation study as given in next paragraphs.

Ablation Study. We perform five different ablation experiments to analyze different components of the proposed network.

1) *Effect of VAB Block*: The VAB block provides the most important visual attention mechanism that improves the network performance. As shown in Table 6.2, the network under-performs on both regular and irregular scene-text datasets without using the VAB block. Thus, it's imperative to include VAB block.

2) *Number of Residual Units*: The number of residual units (RU) in the RS block plays an important role in better feature extraction. We experimented with different RU units quantity per

VAB Block Effect						
	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
w/o VAB	86.6	81.9	88.8	77.4	80.7	75.2
w VAB (ours)	**95.9**	**90.8**	**96.3**	**83.9**	**86.0**	**86.9**

Number of Residual Units (RUs) per RS Block						
	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
1	61.5	58.7	61.2	55.3	59.5	62.1
2	72.6	65.5	68.0	61.0	65.9	67.0
3	83.0	76.1	77.9	71.7	74.8	77.5
4	90.1	84.9	83.5	79.6	82.3	84.2
5 (ours)	**95.9**	**90.8**	**96.3**	83.9	**86.0**	**86.9**
6	94.3	88.8	95.6	**84.0**	85.5	86.2

S2 and S3 scale-modules Effect						
Scale-Modules	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
S1 only	87.4	81.5	90.1	79.9	81.3	82.1
S1,S2 only	92.9	87.3	93.6	82.2	83.7	84.5
S1,S2,S3 (ours)	**95.9**	**90.8**	**96.3**	**83.9**	86.0	**86.9**
S1,S2,S3,S4	95.1	90.6	85.4	82.9	**86.5**	86.0

MaxLength Effect						
MaxLength	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
25 (ours)	**95.9**	**90.8**	**96.3**	**83.9**	**86.0**	**86.9**
50	95.5	90.7	96.2	83.7	85.9	86.9
75	95.6	90.7	96.1	83.8	86.0	86.8
100	95.8	90.6	96.2	83.6	85.9	86.9

Table 6.2: Ablation studies on the proposed network. Several experiments on different components of the proposed network indicate their vitality.

RS block as shown in Table 6.2. As per the results, we found five RU units per RS block to be the most effective choice with the highest accuracy.

3) *Effect of S2 and S3 scale-modules inclusion*: As given in Table 6.2, using the S2 and S3 scale-modules in addition to S1 increases the network effectiveness. However, adding another scale-module S4 does not enhance the accuracy significantly. Thus, the (S1,S2,S3) combination has been employed.

4) *MaxLength Value Selection*: The maxLength value has to be selected so that it covers the maximum length an output word can possibly have in a dataset. Beyond that, increasing it should not have any noticeable effect on the network efficacy. As given in Table 6.2, increasing the

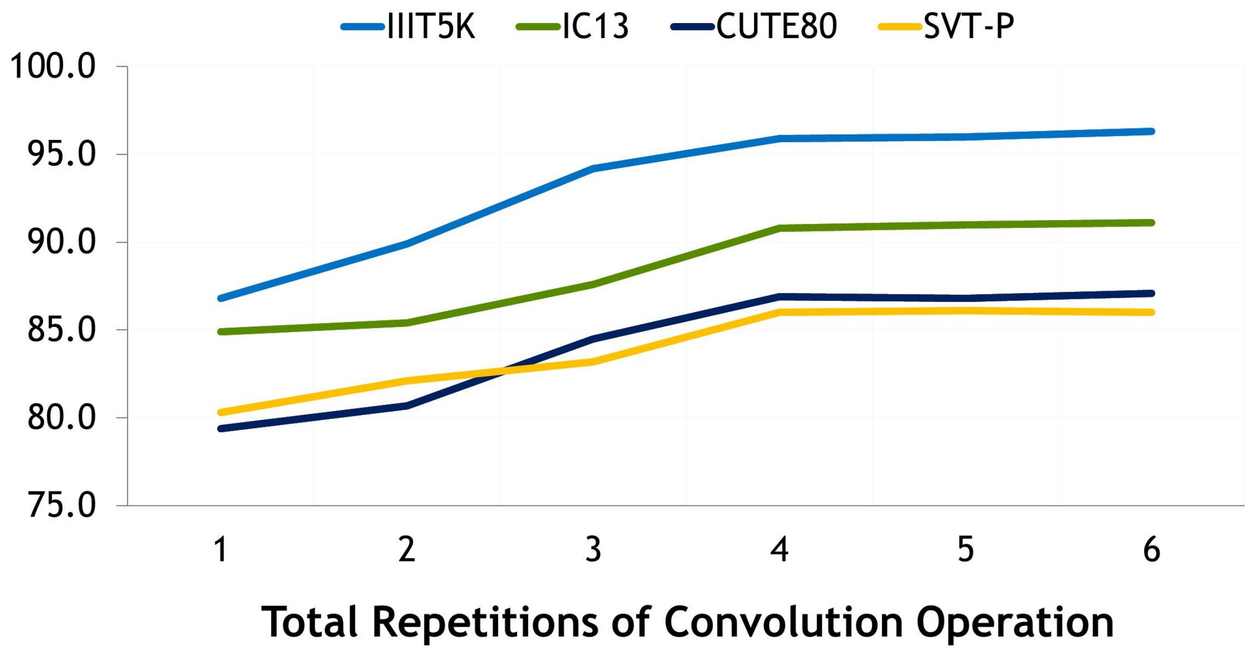


Figure 6.6: VAB convolution quantity analysis graph. The graph indicates that repeating the convolution operation four times in the VAB block before the segmentation-map generation yields the optimal accuracy as tested on four different datasets.

maxLength value from default value of 25 does not alter the performance by much.

5) *Total Convolution Operations in VAB Block*: We investigate the effect of total number of convolution operations before the segmentation map creation. To analyze the effect, we perform convolution operations quantity experiments on four datasets (IIIT5k, IC13, CUTE80, SVT-P). The results are shown in Fig. 6.6, where repeating four convolution operations before the segmentation map generation in the VAB block proves to be the best choice.

Robustness Analysis. Here, we check for robustness of the proposed scheme against different modifications on the input images. We compare our scheme with two recent SOTA methods (DAN [109] and CA-FCN [56]) on two datasets (IIIT-5K [67] and IC13 [44]). Following the practice as given in [109], the variations introduced into these datasets are as follows:

IIIT-padded: 100% padding of the input images in IIIT-5k in both horizontal and vertical direction via border pixels replication. **IIIT-r-padded:** Stretching the image vertices using a random scale value up to 20% for both height and width respectively. Next, border pixels are repeated for

	IIIT	IIIT-padded			IIIT-r-padded			IC13	IC13-expanded			IC13-r-expanded		
Method	acc	acc	diff.	change (%)	acc	diff.	(%)	acc	acc	diff.	(%)	acc	diff.	change (%)
[56]	92.0	89.3	-2.7	2.9	87.6	**-4.4**	4.8	91.4	87.2	-3.7	4.1	83.8	-6.9	7.6
1D [109]	93.3	91.5	-1.8	1.9	88.2	-5.1	5.4	94.2	91.2	**-3.0**	**3.2**	86.9	-7.3	7.7
2D [109]	94.3	92.1	-2.2	2.3	89.1	-5.2	5.5	93.9	90.4	-3.5	3.7	86.9	-7.0	7.5
Ours	**95.9**	**94.0**	**-1.9**	**2.0**	**91.4**	-4.5	**4.7**	**96.3**	**92.5**	-3.8	3.9	**89.5**	**-6.8**	**7.1**

Table 6.3: Robustness Analysis. This study demonstrates the proposed model better robustness towards different changes in the input images. (acc: accuracy, diff.: Accuracy Difference of performance from the original dataset, change (%): Percentage change (decrease) in the accuracy).

filling it. Finally, we crop the axis-aligned rectangles. **IC13-expansion:** The input images in IC13 are expanded into image frames with relatively extra 10% height and width followed by cropping. **IIIT-r-expansion:** Expansion of the IC13 images using a random scale up to 20% height and width, followed by cropping the axis-aligned rectangular images.

As shown in Table 6.3, it can be observed that the proposed method appears as the most stable and resilient to these input distortions and variations in majority cases, hence, demonstrating the robustness of our scheme.

6.5.2 Qualitative Analysis

Here, we present some good and bad qualitative results. We evaluate the proposed scheme with and without the visual attention block (VAB). The results are shown in Fig. 6.7, where the first two rows indicate the good results followed by the failure cases in the last row. Following the practice in [121], under each image, the first line shows the text recognition made by the proposed scheme without using the VAB block followed by our network text prediction with the VAB module in the second line. Characters colored as red indicate wrong predictions. As shown in the good results, the proposed scheme without the VAB block lacks the visual attention and struggles to differentiate between highly similar characters (e.g. 'e' and 'c' or 'o' and 'a') when they lack clear visual exposure, skewed perspective, or partial occlusion. The VAB block coupled with the multi-scale fusion helps in overcoming these issues and produces accurate results as shown.

The bad results, as shown in the last row of Fig. 6.7, mainly occur when the visual attention

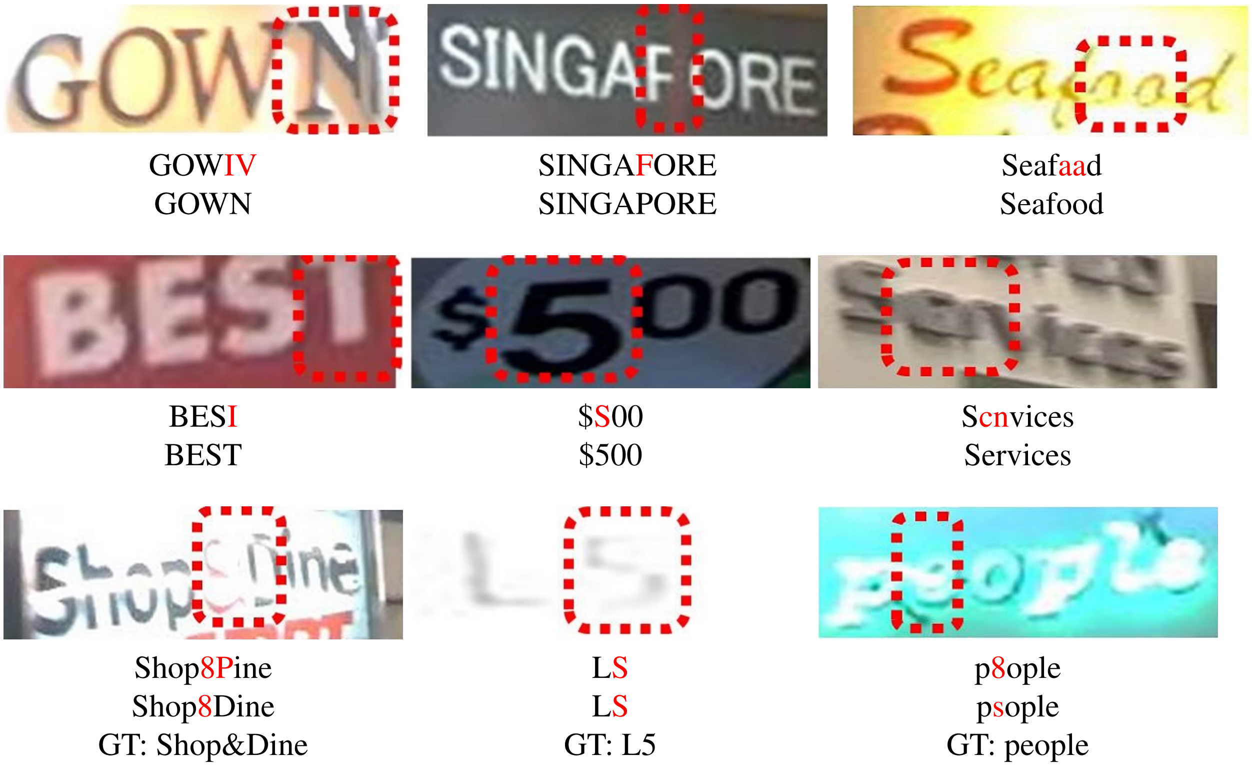


Figure 6.7: Ground truth (GT) scene-text based qualitative comparison. First two rows demonstrate the good prediction results followed by the bad recognition cases in the last row. Under each image, the first line indicates our network text prediction without using the VAB block, whereas second line shows our model with the VAB. The red-colored characters indicate the wrong predictions.

does not align the characters perfectly and results in failure as compared to the ground-truth (GT) recognition text.

6.6 Conclusion

In this chapter, we proposed a new multi-scale and scale-wise visually attended text recognition network to address key scene-text challenges. The multi-scale feature extraction and visual attention have been performed in parallel to utilize different feature scales explicitly in a more effective way. The network also undergoes multi-scale fusion with each other to develop the coordinated information. Experimental evaluation on standard benchmarks indicates better accuracy in most cases as compared to the SOTA methods.

Chapter 7

Conclusion and Future Work

This chapter concludes the dissertation and present some ideas for the future work. Some part of this chapter comes from our research papers [79, 80, 81, 82, 83, 84].

7.1 Conclusion

Here, we first present the concluding remarks regarding our uni-modal research work, followed by a discussion about the multi-modal domain.

7.1.1 Uni-Modal Crowd Counting Networks

We extensively study the crowd estimation problem for uni- and multi-modal inputs and design several novel modules and networks to address key issues and challenges in this domain amid mitigating major SOTA limitations. We propose a conceptually simple yet effective and plug-and-play based patch rescaling module (PRM) to address the major huge crowd diversity issue in crowd counting problems. We present three new uni-modal and uni-branch crowd counting multi-task frameworks that utilize the lightweight PRM module instead of computationally expensive recent multicolumn or multi-regressor based architectures. We devise a new multi-resolution and multi-branch feature-level fusion based end-to-end crowd counting approach for still images that effectively deals with significant variations of crowd-density, lighting conditions, and large perspective. We also propose an alternative patch re-scaling module by more effectively using the input priors. Unlike the PRM, the proposed module fully utilizes all three crowd density levels without requiring any compromising or additional crowd-density classification process. We also present a novel multi-resolution and multi-task PRM based visually attended crowd counting network for static images that effectively address major crowd counting challenges, including the issues of crowd-like background regions and huge crowd-variation. We deploy the plug-and-play PRM module so as to further push its boundaries and utilize it more effectively as compared to its previous deployments. We employ the visual attention mechanism in a unique and effective way on early-stage feature-maps that facilitate the later-stage channels to better understand the foreground regions. Experimental evaluation demonstrates that the proposed networks outperform the state-of-the-art methods in majority cases on four different benchmark datasets with up to **12.6%** improvement in terms of the RMSE evaluation metric. The better cross-dataset performance also validates the better generalization ability of our schemes.

7.1.2 Multi-Modal Crowd Estimation

In the multi-modal domain, we design a novel audio-visual crowd counting model that focuses on the effective utilization of both visual and audio inputs for better modalities association and productive feature extraction. In the same work, we also introduce the notion of auxiliary PIR and PCE information, and empirically show that it is beneficial for effective modalities association and extracting rich visual features without requiring any extra ground-truth annotation process. We have also deployed a unique multi-branch structure to extract rich visual features and also proposed the image-only variant of our model. Experimental evaluation on standard benchmarks reveals the superior performance of our networks under all standard evaluation settings with up to **33.8%** improvement.

7.2 Future Work

In the future, we aim to work on the following problems:

- Investigate the multi-scale crowd counting models in other computer vision tasks (segmentation, pose estimation etc.).

- Inspect the PRM-less CC-FIP network (Chapter 3) with the attention mechanism.
- Design the Multi-Modal (AudioVisual) Crowd counting using the Image+Audio input with special focus on improving the intra- and inter-modality fusion process.

Although we have explored the audio-visual space quite comprehensively, however, we still believe that extracting the valuable and strong coherent features from both (audio and visual) modalities for improved effectiveness can be further explored. Only a few countable works have been done to date in this domain. Therefore, we aim to design a more effective multi-modal (audiovisual) crowd counting model that gives better performance amid addressing the above crowd counting challenges.

References

[1] Alberti, C., Ling, J., Collins, M., & Reitter, D. (2019). Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*.

[2] B. Su and S. Lu (2017). Accurate recognition of words in scenes without character segmentation using recurrent neural network. *Pattern Recognition*, 63, 397–405.

[3] Babu Sam, D., Sajjan, N. N., Venkatesh Babu, R., & Srinivasan, M. (2018). Divide and Grow: Capturing Huge Diversity in Crowd Images With Incrementally Growing CNN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3618–3626).

[4] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[5] Bai, F., Cheng, Z., Niu, Y., Pu, S., & Zhou, S. (2018). Edit probability for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1508–1516).

[6] Bai, S., He, Z., Qiao, Y., Hu, H., Wu, W., & Yan, J. (2020). Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4594–4603).

[7] Bissacco, A., Cummins, M., Netzer, Y., & Neven, H. (2013). Photoocr: Reading text in uncontrolled conditions. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 785–792).

[8] Casey, R. G. & Lecolinet, E. (1996). A survey of methods and strategies in character segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 18(7), 690–706.

[9] Cen, F., Zhao, X., Li, W., & Wang, G. (2021). Deep feature augmentation for occluded image classification. *Pattern Recognition*, 111, 107737.

[10] Chan, A. B. & Vasconcelos, N. (2009). Bayesian poisson regression for crowd counting. In *IEEE International Conference on Computer Vision* (pp. 545–551).

[11] Chen, J., Chen, Z., Chi, Z., & Fu, H. (2014). Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction* (pp. 508–513).

[12] Chen, K., Loy, C. C., Gong, S., & Xiang, T. (2012). Feature mining for localised crowd counting. In *BMVC*, volume 1 (pp.$\tilde{3}$).

[13] Chen, X., Bin, Y., Gao, C., Sang, N., & Tang, H. (2020). Relevant region prediction for crowd counting. *Neurocomputing*.

[14] Chen, X. & Wang, G. (2021). Few-shot learning by integrating spatial and frequency representation. *arXiv preprint arXiv:2105.05348*.

[15] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., & Zhou, S. (2017). Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision* (pp. 5076–5084).

[16] Cheng, Z., Xu, Y., Bai, F., Niu, Y., Pu, S., & Zhou, S. (2018). Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5571–5579).

[17] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[18] Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization

with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1 (pp. 1–2).

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[20] Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, CVPR* (pp. 1–8).

[21] Gan, C., Huang, D., Zhao, H., Tenenbaum, J. B., & Torralba, A. (2020). Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10478–10487).

[22] Ge, W. & Collins, R. T. (2009). Marked point processes for crowd counting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2913–2920).: IEEE.

[23] Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

[24] Gupta, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2315–2324).

[25] Gurban, M., Thiran, J.-P., Drugman, T., & Dutoit, T. (2008). Dynamic modality weighting for multi-stream hmms in audio-visual speech recognition. In *Proceedings of the 10th international conference on Multimodal interfaces* (pp. 237–240).

[26] He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

[27] He, P., Huang, W., Qiao, Y., Loy, C., & Tang, X. (2016b). Reading scene text in deep convolutional sequences. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

[28] Hershey, J., Attias, H., Jojic, N., & Kristjansson, T. (2004). Audio-visual graphical models for speech processing. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5 (pp. V–649).: IEEE.

[29] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131–135).: IEEE.

[30] Hu, D., Li, X., et al. (2016a). Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3574–3582).

[31] Hu, D., Mou, L., Wang, Q., Gao, J., Hua, Y., Dou, D., & Zhu, X. X. (2020). Ambient sound helps: Audiovisual crowd counting in extreme conditions. *arXiv preprint arXiv:2005.07097*.

[32] Hu, Y., Chang, H., Nian, F., Wang, Y., & Li, T. (2016b). Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation*, 38, 530–539.

[33] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

[34] Huang, J., Tao, J., Liu, B., Lian, Z., & Niu, M. (2020). Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3507–3511).: IEEE.

[35] Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., & Shah, M. (2018). Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 532–546).

[36] Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[37] Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*.

[38] Jaderberg, Max and Simonyan, Karen and Vedaldi, Andrea and Zisserman, Andrew (2014). Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

[39] Jaderberg, Max and Simonyan, Karen and Vedaldi, Andrea and Zisserman, Andrew (2016). Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116(1), 1–20.

[40] Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D., & Shao, L. (2019). Crowd counting and density estimation by trellis encoder-decoder networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6133–6142).

[41] Jin, Q. & Liang, J. (2016). Video description generation using audio and visual cues. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (pp. 239–242).

[42] Kan, S., Zhang, L., He, Z., Cen, Y., Chen, S., & Zhou, J. (2020). Metric learning-based kernel transformer with triplets and label constraints for feature fusion. *Pattern Recognition*, 99, 107086.

[43] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., et al. (2015). Icdar 2015 competition on

robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1156–1160).: IEEE.

[44] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L. G., Mestre, S. R., Mas, J., Mota, D. F., Almazan, J. A., & De Las Heras, L. P. (2013). Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 1484–1493).: IEEE.

[45] Kiela, D., Grave, E., Joulin, A., & Mikolov, T. (2018). Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[46] Kim, Y., Lee, H., & Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3687–3691).: IEEE.

[47] Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[48] Kong, Q., Xu, Y., Wang, W., & Plumbley, M. D. (2018). Audio set classification with attention model: A probabilistic perspective. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 316–320).: IEEE.

[49] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

[50] Li, C. & Qian, Y. (2020). Deep audio-visual speech separation with attention mechanism. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7314–7318).: IEEE.

[51] Li, H., Wang, P., Shen, C., & Zhang, G. (2019a). Show, attend and read: A simple and strong

baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33 (pp. 8610–8617).

[52] Li, M., Zhang, Z., Huang, K., & Tan, T. (2008). 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* (pp. 1–4).: IEEE.

[53] Li, W., Li, H., Wu, Q., Meng, F., Xu, L., & Ngan, K. N. (2019b). Headnet: An end-to-end adaptive relational network for head detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

[54] Li, Y., Zhang, X., & Chen, D. (2018). CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1091–1100).

[55] Liao, M., Lyu, P., He, M., Yao, C., Wu, W., & Bai, X. (2019a). Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE transactions on pattern analysis and machine intelligence*.

[56] Liao, M., Zhang, J., Wan, Z., Xie, F., Liang, J., Lyu, P., Yao, C., & Bai, X. (2019b). Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33 (pp. 8714–8721).

[57] Lin, Y.-B. & Wang, Y.-C. F. (2020). Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*.

[58] Liu, J., Gao, C., Meng, D., & Hauptmann, A. G. (2018a). DecideNet: counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5197–5206).

[59] Liu, W., Salzmann, M., & Fua, P. (2019). Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5099–5108).

[60] Liu, X., van de Weijer, J., & Bagdanov, A. D. (2018b). Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7661–7669).

[61] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

[62] Lucas, S. M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., et al. (2005). Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2-3), 105–122.

[63] Luo, C., Jin, L., & Sun, Z. (2019). Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90, 109–118.

[64] Lyu, P., Yang, Z., Leng, X., Wu, X., Li, R., & Shen, X. (2019). 2d attentional irregular scene text recognizer. *arXiv preprint arXiv:1906.05708*.

[65] Ma, Z., Wei, X., Hong, X., & Gong, Y. (2019). Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 6142–6151).

[66] McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.

[67] Mishra, A., Alahari, K., & Jawahar, C. (2012). Scene text recognition using higher order language priors.

[68] Mroueh, Y., Marcheret, E., & Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2130–2134).: IEEE.

[69] Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML*.

[70] Neumann, L. & Matas, J. (2012). Real-time scene text localization and recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3538–3545).: IEEE.

[71] Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., & Poggio, T. (1997). Pedestrian detection using wavelet templates. In *cvpr*, volume 97 (pp. 193–199).

[72] Petridis, S., Wang, Y., Li, Z., & Pantic, M. (2017). End-to-end audiovisual fusion with lstms. *arXiv preprint arXiv:1709.04343*.

[73] Quy Phan, T., Shivakumara, P., Tian, S., & Lim Tan, C. (2013). Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 569–576).

[74] Rahman, W., Hasan, M., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., Hoque, E., et al. (2019). Integrating multimodal information in large pretrained transformers. *arXiv preprint arXiv:1908.05787*.

[75] Ranjan, V., Le, H., & Hoai, M. (2018). Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 270–285).

[76] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

[77] Risnumawan, A., Shivakumara, P., Chan, C. S., & Tan, C. L. (2014). A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18), 8027–8048.

[78] Ryan, D., Denman, S., Fookes, C., & Sridharan, S. (2009). Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications, DICTA* (pp. 81–88).

[79] Sajid, U., Chen, X., Sajid, H., Kim, T., & Wang, G. (2021a). Audio-visual transformer based crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2249–2259).

[80] Sajid, U., Chow, M., Zhang, J., Kim, T., & Wang, G. (2021b). Parallel scale-wise attention network for effective scene text recognition. *arXiv preprint arXiv:2104.12076*.

[81] Sajid, U., Ma, W., & Wang, G. (2021c). Multi-resolution fusion and multi-scale input priors based crowd counting. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 5790–5797).: IEEE.

[82] Sajid, U., Sajid, H., Wang, H., & Wang, G. (2020). Zoomcount: A zooming mechanism for crowd counting in static images. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10), 3499–3512.

[83] Sajid, U. & Wang, G. (2020). Plug-and-play rescaling based crowd counting in static images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2287–2296).

[84] Sajid, U. & Wang, G. (2021). Towards more effective prm-based crowd counting via a multi-resolution fusion and attention network. *Neurocomputing*.

[85] Sam, D. B., Surya, S., & Babu, R. V. (2017). Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5744–5752).

[86] Shami, M., Maqbool, S., Sajid, H., Ayaz, Y., & Cheung, S.-C. S. (2018). People counting in dense crowd images using sparse head detections. *IEEE Transactions on Circuits and Systems for Video Technology*.

[87] Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., & Yang, X. (2018). Crowd counting via adversarial

cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5245–5254).

[88] Shi, B., Bai, X., & Yao, C. (2016a). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11), 2298–2304.

[89] Shi, B., Wang, X., Lyu, P., Yao, C., & Bai, X. (2016b). Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4168–4176).

[90] Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., & Bai, X. (2018). Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9), 2035–2048.

[91] Shi, Z., Mettes, P., & Snoek, C. G. (2019). Counting with focus for free. *arXiv preprint arXiv:1903.12206*.

[92] Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., & Bartlett, M. (2013). Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 517–524).

[93] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[94] Sindagi, V. A. & Patel, V. M. (2017). CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6).

[95] Sindagi, V. A. & Patel, V. M. (2019). HA-CCN: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29, 323–335.

[96] Sterpu, G., Saam, C., & Harte, N. (2018). Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 111–115).

[97] Su, B. & Lu, S. (2014). Accurate scene text recognition based on recurrent neural network. In *Asian Conference on Computer Vision* (pp. 35–48).: Springer.

[98] Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019a). Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7464–7473).

[99] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019b). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5693–5703).

[100] Tian, Y., Shi, J., Li, B., Duan, Z., & Xu, C. (2018). Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 247–263).

[101] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

[102] Walach, E. & Wolf, L. (2016). Learning to count with cnn boosting. In *European Conference on Computer Vision* (pp. 660–676).: Springer.

[103] Wan, J., Luo, W., Wu, B., Chan, A. B., & Liu, W. (2019). Residual regression with semantic prior for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4036–4045).

[104] Wang, C., Zhang, H., Yang, L., Liu, S., & Cao, X. (2015). Deep people counting in ex-

tremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1299–1302).

[105] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020a). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.

[106] Wang, K., Babenko, B., & Belongie, S. (2011). End-to-end scene text recognition. In *2011 International Conference on Computer Vision* (pp. 1457–1464).: IEEE.

[107] Wang, M. & Wang, X. (2011). Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3401–3408).

[108] Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012). End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)* (pp. 3304–3308).: IEEE.

[109] Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., & Cai, M. (2020b). Decoupled attention network for text recognition. In *AAAI* (pp. 12216–12224).

[110] Wei, J., Pei, E., Jiang, D., Sahli, H., Xie, L., & Fu, Z. (2014). Multimodal continuous affect recognition based on lstm and multiple kernel learning. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific* (pp. 1–4).: IEEE.

[111] Wu, B. & Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *IEEE International Conference on Computer Vision* (pp. 90–97).

[112] Wu, Z., Jiang, Y.-G., Wang, J., Pu, J., & Xue, X. (2014). Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 167–176).

[113] Xie, Z., Huang, Y., Zhu, Y., Jin, L., Liu, Y., & Xie, L. (2019). Aggregation cross-entropy for sequence recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6538–6547).

[114] Xiong, H., Lu, H., Liu, C., Liu, L., Cao, Z., & Shen, C. (2019). From open set to closed set: Counting objects by spatial divide-and-conquer. *arXiv preprint arXiv:1908.06473*.

[115] Xu, C., Qiu, K., Fu, J., Bai, S., Xu, Y., & Bai, X. (2019). Learn to scale: Generating multipolar normalized density map for crowd counting. *arXiv preprint arXiv:1907.12428*.

[116] Xuan, H., Zhang, Z., Chen, S., Yang, J., & Yan, Y. (2020). Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 (pp. 279–286).

[117] Yang, M., Guan, Y., Liao, M., He, X., Bian, K., Bai, S., Yao, C., & Bai, X. (2019). Symmetry-constrained rectification network for scene text recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 9147–9156).

[118] Yang, X., He, D., Zhou, Z., Kifer, D., & Giles, C. L. (2017). Learning to read irregular text with attention mechanisms. In *IJCAI*, volume 1 (pp.$\tilde{3}$).

[119] Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., & Sebe, N. (2020). Reverse perspective network for perspective-aware object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4374–4383).

[120] Yao, H., Han, K., Wan, W., & Hou, L. (2017). Deep spatial regression model for image crowd counting. *arXiv preprint arXiv:1710.09757*.

[121] Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., & Ding, E. (2020a). Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12113–12122).

[122] Yu, J., Zhang, S.-X., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., & Yu, D. (2020b). Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6984–6988).: IEEE.

[123] Yuhas, B. P., Goldstein, M. H., & Sejnowski, T. J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11), 65–71.

[124] Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

[125] Zhan, F. & Lu, S. (2019). Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2059–2068).

[126] Zhang, L., Shi, M., & Chen, Q. (2018). Crowd counting via scale-adaptive convolutional neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1113–1121).: IEEE.

[127] Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 589–597).