Audio-Visual Transformer Based Crowd Counting

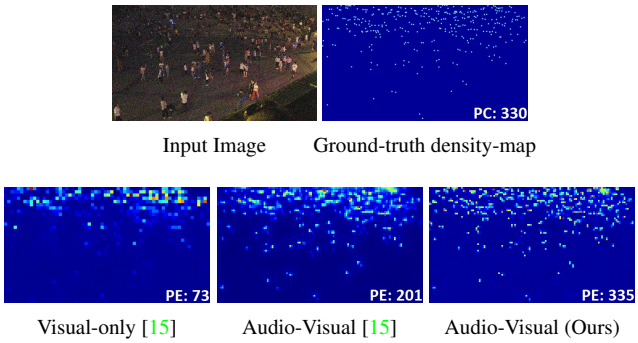
Usman Sajid[1], Xiangyu Chen[1], Hasan Sajid[2], Taejoon Kim[1], Guanghui Wang[3]
[1]*Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA, 66045*
[2] *School of Mechanical and Manufacturing Engineering, NUST, Islamabad, Pakistan*
[3] *Department of Computer Science, Ryerson University, Toronto, ON, Canada M5B 2K3*
Email: {usajid,xychen,taejoonkim}@ku.edu[1], hasan.sajid@smme.nust.edu.pk[2], wangcs@ryerson.ca[3]

Abstract

Crowd estimation is a very challenging problem. The most recent study 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. The paper 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.*

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 security, and circumventing any undesirable incident. Many deep learning-based image-only schemes [39, 42, 41, 17, 41, 19, 60, 32] have been proposed to date, ranging from single and multi-branch networks [60, 39, 41], multi-regressors [42] based to trellis networks [19]. 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. 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.



Input Image | Ground-truth density-map

Visual-only [15] | Audio-Visual [15] | Audio-Visual (Ours)

Figure 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).

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.* [15] 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. 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 [55, 3, 14, 22], event localization [30, 56], and speech recognition [59, 10, 34, 47]. 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 [15]) 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 [45] or ResNet-based [11] standard structures for that.

To address these major challenges and issues, we propose a new transformer-based [51] audio-visual multi-task crowd counting network as shown in Fig. 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. 1. The main contributions of our work include:

- 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. The source code of the models will be released soon.

2. Related Work

Audio-Visual Learning. Audio-visual representation learning aims to aid the visual modality with audio or vice-versa. Early speech perception research [33] demonstrates that the visual information can change what people hear, i.e., McGurk Effect. Since then, vision and audio modalities

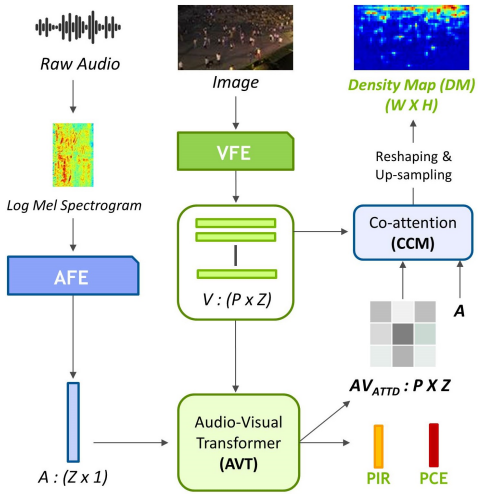


Figure 2. The proposed audio-visual crowd counting network. The extracted audio-visual features (V, A) go through the AVT module to obtain two auxiliary network outputs (PIR, PCE) and third (run-time) modality (AV_{ATTD}). The AV_{ATTD} undergoes the cross-modality co-attention fusion with V and A via the CCM module, followed by getting the final crowd density-map (DM).

are widely explored in speech recognition[59, 10, 34, 47], video classification [55], emotion recognition [23] and video description [20]. Multiple kernels are broadly implemented as the fusion module by feeding the kernels with data from different modalities [4, 44, 54]. Another fusion method is based on graphical models considering its advantages in temporal related tasks [10, 12]. Besides, neural networks raise more attention in fusion especially since the appearance of RNN and LSTM [36, 47]. More recently, transformer-based [51] fusion raises growing attention [1, 48, 37, 16, 21], especially after its application in vision [7]. In addition to that, there are also some model-agnostic fusion methods, including the simple concatenation [27, 6, 58] and element-wise operation [8, 50].

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 [43, 28] detect each person via some standard object detectors (e.g. Faster-RCNN [9], YOLO [38]). These methods give unsatisfactory results in the high-density crowd scenarios. The DReg models [39, 41, 52, 40, 52] directly regress the crowd number using CNN-based structures. Wang *et al.* [52] deployed the AlexNet [26] variant for direct crowd regression. Recently, Sajid *et al.* designed two different types of direct-regression counting methods [41, 40] that use the patch-rescaling module (PRM) and branch structure to deal with varying crowd levels. But these models fail to utilize

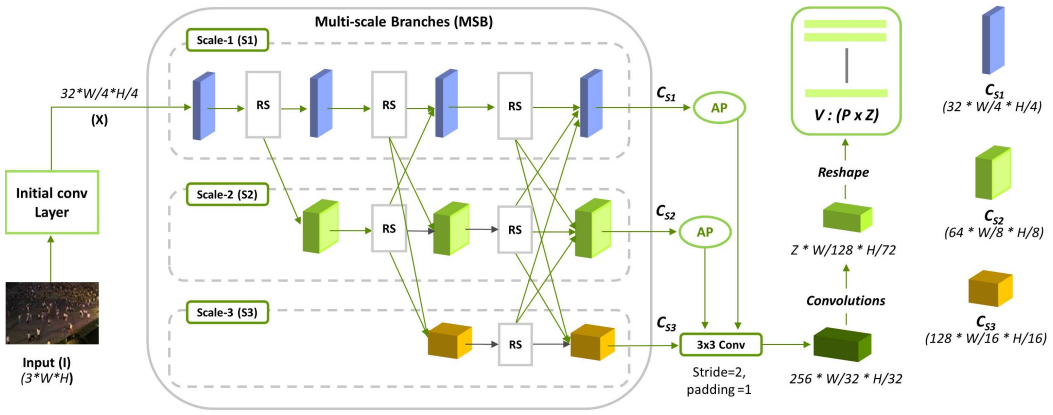


Figure 3. The framework of Visual Feature Extraction (VFE) block.

the valuable density-map based computation. The DMap methods [17, 60, 42, 32, 46] estimate crowd density-map, where each pixel indicates crowd-density. The pixel values are then summed-up to obtain final count. Switch-CNN [42] uses CCN-based switch that routes the image to one of three specialized regressors, each dealing with specific crowd-density. Li *et al.* [29] employed dilation layers for better contextual information retrieval. Liu *et al.* [32] used rank-based system for unsupervised learning. Idrees *et al.* [17] deployed composition loss to jointly learn the count, localization and density-map. HA-CCN [46] 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.* [15] 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. Therefore, the proposed DMap-based network focuses on solving these major challenges amid dealing with severe conditions as discussed above.

3. Proposed Approach

The proposed multi-task model, as shown in Fig. 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.

3.1. Audio Feature Extraction (AFE)

To extract the audio features embedding, we deploy the ResNet-like CNN structure [13] (pretrained on the AudioSet dataset [25]) and apply it on the conventionally computed [15] 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:

$$A = AFE(A_{LMS}) \tag{1}$$

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

3.2. Visual Feature Extraction (VFE)

The VFE component, as shown in Fig. 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 [11]. Similar to the high-resolution networks [49, 53], 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 ($C_{S1} \in \mathbb{R}^{32*\frac{W}{4}*\frac{H}{4}}, C_{S2} \in \mathbb{R}^{64*\frac{W}{8}*\frac{H}{8}}, C_{S3} \in \mathbb{R}^{128*\frac{W}{16}*\frac{H}{16}}$).

3.2.1 Inter-Branch Fusion

The purpose of inter-branch fusion is to develop coordinated knowledge in-between the multi-scale branches. We

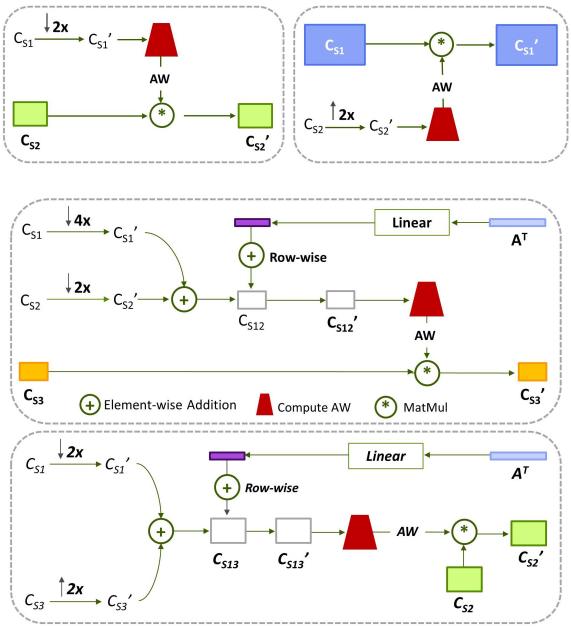


Figure 4. Illustration of different transformer-inspired inter-branch fusion cases. (S1 → S2 only (top-left), S2 → S1 only (top-right), {S1, S2} → S3 (middle), {S1, S3}, → S2 (bottom))

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. 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. 4. Mathematically, it is defined as:

$$C'_{S2} = AW * C_{S2} = softmax(C'_{S1} * C'^{T}_{S1}) * C_{S2} \quad (2)$$

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. 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 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. 4. It is defined as:

$$C'_{S3} = AW * C_{S3} = softmax(C'_{S12} * C'^{T}_{S12}) * C_{S3} \quad (3)$$

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 $2\times$ 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. 4. We can define it as:

$$C'_{S2} = AW * C_{S2} = softmax(C'_{S13} * C'^{T}_{S13}) * C_{S2} \quad (4)$$

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

3.2.2 Visual Features Generation

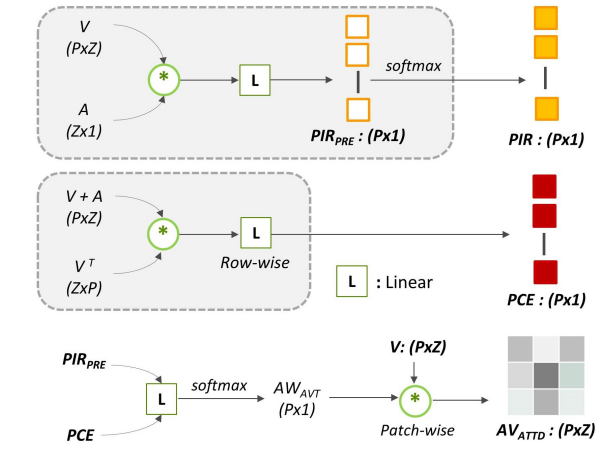
The MSB higher-scales outputs (C_{S1}, C_{S2}) are merged together with the lowest-branch output channels (C_{S3}) through (3×3) convolution after down-scaling higher features via required average pooling (AP). The generated channels ($\in \mathbb{R}^{256 * \frac{W}{32} * \frac{H}{32}}$) employ several 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 [18] and ReLU [35] 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:

$$V = VFE(X), \quad (5)$$

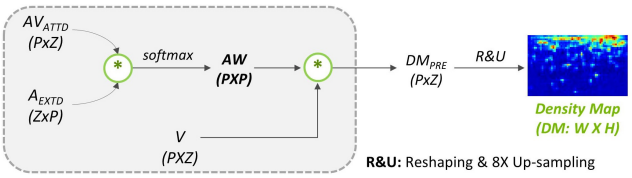
where $V \in \mathbb{R}^{P * Z}$, and P ($= \frac{W}{128} * \frac{H}{72}$) represents the total patches/regions in the input image. Intuitively, the V matrix can be perceived as containing the Z-dimensional embedding for each image-patch, with P total patches.

3.3. Audio-Visual Transformer (AVT)

The purpose of the AVT module is twofold: 1) Calculate and output auxiliary Patch-Importance Ranking (PIR) and



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



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

Figure 5. Illustration of PIR, PCE, AV_{ATTD}, and DM computations.

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(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:

$$PIR = softmax(PIR_{PRE}) \tag{6}$$

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:

$$PIR_{GT}(j) = \frac{CC_{GT}(j)}{CC_{GT}(image)} \tag{7}$$

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}):

$$Loss_{PIR} = \frac{1}{\sqrt{P}} \sum_{i=1}^{P} PIR_{GT}(i) \log(\frac{PIR_{GT}(i)}{PIR(i)}) \tag{8}$$

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

$$PCE = Linear_{Row-wise}((V + A) * V^T) \tag{9}$$

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(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:

$$Loss_{PCE} = \sum_{i=1}^{P} (\frac{PCE_{GT}(i) - PCE(i)}{\sum_{j=1}^{P} PCE_{GT}(j)})^2 \tag{10}$$

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.

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(b) and defined as:

$$DM_{PRE} = softmax(AV_{ATTD} * A_{EXTD}) * V \tag{11}$$

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

3.5. Final Crowd Estimate (CE_{FINAL})

The DM_{PRE} gets re-shaped and up-sampled $8\times$ to output the final crowd Density-Map ($DM \in \mathbb{R}^{W*H}$) as shown in Fig. 5(b). We sum all DM pixel-values to obtain the final crowd estimate (CE_{FINAL}) for the input image-audio.

Method	Regular Images		Low Resolution		Gaussian Noise				Low Illumination & Gaussian Noise				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
MCNN [60]	53.40	84.10	60.17	89.35	53.47	84.04	53.92	84.04	70.72	96.11	70.58	96.11	60.38	88.96
CANNet [31]	15.41	28.96	22.16	39.60	13.31	27.23	14.20	28.04	26.03	49.11	33.14	58.27	20.71	38.54
CSRNet [29]	13.88	28.79	17.14	30.64	13.79	28.01	14.55	29.15	35.78	62.76	45.88	75.40	23.50	42.46
AudioCSRNet [15]	14.24	28.07	16.88	31.46	13.07	27.45	13.70	28.67	25.06	51.58	27.33	45.16	18.38	35.40
CC-V (Ours)	12.97	25.76	16.91	32.82	13.31	28.79	13.92	29.01	26.03	55.72	27.59	58.69	18.46	38.47
CC-AV (Ours)	**9.24**	**19.81**	**11.18**	**26.25**	**10.15**	**19.76**	**10.39**	**19.79**	**20.14**	**44.58**	**21.17**	**40.86**	**13.71**	**28.51**
CC-AV 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 1. Quantitative Evaluation on the DISCO Benchmark [15] 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 [15])

We deploy L_2-norm as the DM loss-function, given as:

$$Loss_{DM} = \sum_{m=1}^{W}\sum_{n=1}^{H}(DM_{mn} - DM'_{mn})^2 \quad (12)$$

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:

$$Loss_{TOTAL} = Loss_{PIR} + Loss_{PCE} + Loss_{DM} \quad (13)$$

Unlike other existing audio-visual mechanisms [15], 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.

4. Training and Evaluation Details

The only available audio-visual crowd counting dataset to-date (DISCO) [15] 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 [24] 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 overfitting, 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:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|E_n - C_n|, RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(E_n - C_n)^2} \quad (14)$$

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. Experiments

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

5.1. Experiments on Audio-Visual Dataset

DISCO [15] 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 5.2. 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 1, indicate that both proposed network versions outperform the state-of-the-arts under all evaluation metrics with CC-AV giving

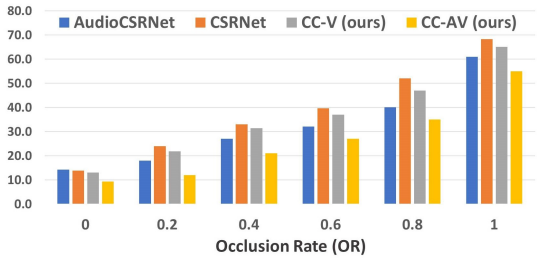


Figure 6. Occluded images setting based evaluation using MAE metric.

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-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 [15]. Lastly, the Gaussian noise has been added in the strong noise case as given in [15]. Observing the results for all three settings, as shown in Table 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. 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[15] and CSRNet [29]) 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), CC-AV still performs better, indicating its robustness and better utilization of the audio-modality as compared to the best audio-visual models.

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.

	ShanghaiTech [60]		UCF-QNRF [17]	
Method	MAE	RMSE	MAE	RMSE
MCNN [60]	110.2	173.2	277	426
Switch-CNN [42]	90.4	135.0	228	445
CSRNet [29]	68.2	115.0	-	-
CL[17]	-	-	132	191
CAN [31]	62.3	100.0	107	183
RRP [5]	63.2	105.7	93	156
HA-CCN [46]	62.9	94.9	118.1	180.4
ADSCNet [2]	**55.4**	97.7	**71.3**	132.5
RPNet [57]	61.2	96.9	-	-
PRM-based[41]	67.8	86.2	94.5	141.9
CC-V (Ours)	58.7	**81.3**	75.4	**125.6**

Table 2. MAE and RMSE based evaluation on image-only datasets.

We compare our CC-V model on two image-only diverse benchmark datasets: UCF-QNRF [17] and ShanghaiTech Part-A [60]. 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 $1,024 \times 576$ with zero-padding if required. The results on both datasets are shown in Table 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.

5.3. Ablation Study

In addition to the previous sub-section 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}.

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

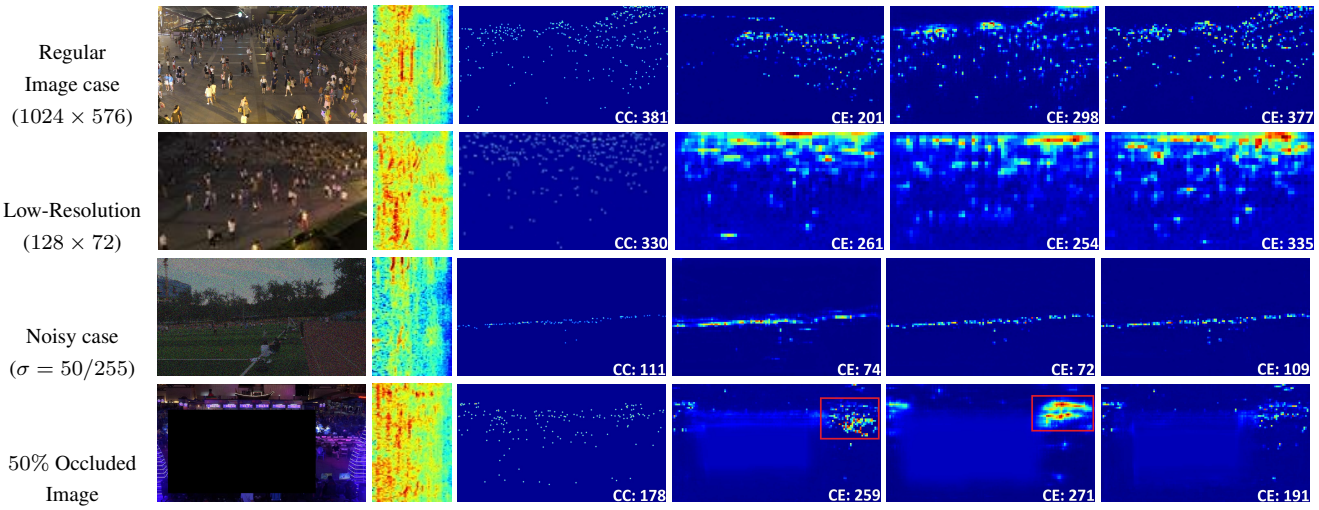


Figure 7. Ground truth (GT) density-map and crowd-count (CC) based qualitative comparison. (From Left to Right Column: Input Image, Audio Log Mel-Spectogram, GT density-map, AudioCSRNet network [15] estimated density-map, CC-V model (ours) density-map, CC-AV (ours) density-map)

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

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 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.

5.4. Qualitative Analysis

We present a few visual results as shown in Fig 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 [15]. 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 [15] 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. 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 [15] 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 in such extreme cases with reasonable estimates for both occluded and non-occluded regions.

6. Conclusion

In this paper, we have presented a new audio-visual multi-task network for effective people counting. To address severe challenging situations, we have introduced explicit PIR and PCE information for better modalities association, and also produce 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 indicates the superior performance of our networks in most cases.

References

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019. 2

[2] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2020. 7

[3] Feng Cen, Xiaoyu Zhao, Wuzhuang Li, and Guanghui Wang. Deep feature augmentation for occluded image classification. *Pattern Recognition*, 111:107737, 2021. 1

[4] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 508–513, 2014. 2

[5] Xinya Chen, Yanrui Bin, Changxin Gao, Nong Sang, and Hao Tang. Relevant region prediction for crowd counting. *Neurocomputing*, 407:399–408, 2020. 7

[6] Xiangyu Chen and Guanghui Wang. Few-shot learning by integrating spatial and frequency representation. *arXiv preprint arXiv:2105.05348*, 2021. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 2

[9] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[10] Mihai Gurban, Jean-Philippe Thiran, Thomas Drugman, and Thierry Dutoit. Dynamic modality weighting for multi-stream hmms in audio-visual speech recognition. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 237–240, 2008. 1, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[12] John Hershey, Hagai Attias, Nebojsa Jojic, and Trausti Kristjansson. Audio-visual graphical models for speech processing. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–649. IEEE, 2004. 2

[13] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 3

[14] Di Hu, Xuelong Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582, 2016. 1

[15] Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, and Xiao Xiang Zhu. Ambient sound helps: Audiovisual crowd counting in extreme conditions. *arXiv preprint arXiv:2005.07097*, 2020. 1, 2, 3, 6, 7, 8

[16] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE, 2020. 2

[17] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018. 1, 3, 7

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[19] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2019. 1

[20] Qin Jin and Junwei Liang. Video description generation using audio and visual cues. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 239–242, 2016. 2

[21] Shichao Kan, Linna Zhang, Zhihai He, Yigang Cen, Shiming Chen, and Jikun Zhou. Metric learning-based kernel transformer with triplets and label constraints for feature fusion. *Pattern Recognition*, 99:107086, 2020. 2

[22] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1

[23] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 3687–3691. IEEE, 2013. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley. Audio set classification with attention model: A probabilistic perspective. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320. IEEE, 2018. 3

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 2

[27] Chenda Li and Yanmin Qian. Deep audio-visual speech separation with attention mechanism. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7314–7318. IEEE, 2020. 2

[28] Wei Li, Hongliang Li, Qingbo Wu, Fanman Meng, Linfeng Xu, and King Ngi Ngan. Headnet: An end-to-end adaptive relational network for head detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 2

[29] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. 3, 6, 7

[30] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[31] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019. 6, 7

[32] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018. 1, 3

[33] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976. 2

[34] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE, 2015. 1, 2

[35] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 4

[36] Stavros Petridis, Yujiang Wang, Zuwei Li, and Maja Pantic. End-to-end audiovisual fusion with lstms. *arXiv preprint arXiv:1709.04343*, 2017. 2

[37] Wasifur Rahman, Md Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, Ehsan Hoque, et al. Integrating multimodal information in large pretrained transformers. *arXiv preprint arXiv:1908.05787*, 2019. 2

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[39] Usman Sajid, Wenchi Ma, and Guanghui Wang. Multi-resolution fusion and multi-scale input priors based crowd counting. *arXiv preprint arXiv:2010.01664*, 2020. 1, 2

[40] Usman Sajid, Hasan Sajid, Hongcheng Wang, and Guanghui Wang. Zoomcount: A zooming mechanism for crowd counting in static images. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3499–3512, 2020. 2

[41] Usman Sajid and Guanghui Wang. Plug-and-play rescaling based crowd counting in static images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2287–2296, 2020. 1, 2, 7

[42] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039. IEEE, 2017. 1, 3, 7

[43] Mamoona Shami, Salman Maqbool, Hasan Sajid, Yasar Ayaz, and Sen-Ching Samson Cheung. People counting in dense crowd images using sparse head detections. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018. 2

[44] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 517–524, 2013. 2

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[46] Vishwanath A Sindagi and Vishal M Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2019. 3, 7

[47] George Sterpu, Christian Saam, and Naomi Harte. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 111–115, 2018. 1, 2

[48] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 2

[49] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3

[50] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

[52] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1299–1302, 2015. 2

[53] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[54] Jiamei Wei, Ercheng Pei, Dongmei Jiang, Hichem Sahli, Lei Xie, and Zhonghua Fu. 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*, pages 1–4. IEEE, 2014. 2

[55] Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 167–176, 2014. 1, 2

[56] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 279–286, 2020. 1

[57] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Reverse perspective network for perspective-aware object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4374–4383, 2020. 7

[58] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE, 2020. 2

[59] Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71, 1989. 1, 2

[60] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 1, 3, 6, 7