# Video-Based Crowd Counting Using a Multi-Scale Optical Flow Pyramid Network

Mohammad Asiful Hossain[1], Kevin Cannons[2], Daesik Jang[3], Fabio Cuzzolin[4], and Zhan Xu[5]

Huawei Technologies Canada Co. Ltd.[1,2,3,5], Oxford Brookes University[4]

**Abstract.** This paper presents a novel approach to the task of video-based crowd counting, which can be formalized as the regression problem of learning a mapping from an input image to an output crowd density map. Convolutional neural networks (CNNs) have demonstrated striking accuracy gains in a range of computer vision tasks, including crowd counting. However, the dominant focus within the crowd counting literature has been on the single-frame case or applying CNNs to videos in a frame-by-frame fashion without leveraging motion information. This paper proposes a novel architecture that exploits the spatiotemporal information captured in a video stream by combining an optical flow pyramid with an appearance-based CNN. Extensive empirical evaluation on five public datasets comparing against numerous state-of-the-art approaches demonstrates the efficacy of the proposed architecture, with our methods reporting best results on all datasets.

## 1 Introduction

Crowd counting is a well-studied area in computer vision, with several real-world applications including urban planning, traffic monitoring, and emergency response preparation [1]. Despite these strong, application-driven motivations, crowd counting remains an unsolved problem. Critical challenges that remain in this area include severe occlusion, diverse crowd densities, perspective effects, and differing illumination conditions.

The task of crowd counting is well understood: Given an arbitrary image of a scene without any prior knowledge (i.e., unknown camera position, camera parameters, scene layout, and crowd density), estimate the number of people in the image. In general, there are two methodologies for estimating the person count in an image: detection-based (e.g., [2–4]) and regression-based (e.g., [5–11]). Detection-based approaches leverage the rapid advancements of convolutional neural network (CNN) object detectors, applying them to the specialized task of identifying human bodies/heads. Although significant progress has been made recently with detection-based approaches, they still perform better at lower crowd densities, with accuracies degrading on challenging images with very high densities, low resolution faces, and significant occlusions. In contrast, regression-based approaches typically employ a CNN to produce a density map, representing the estimated locations of persons within the image. With regression-based

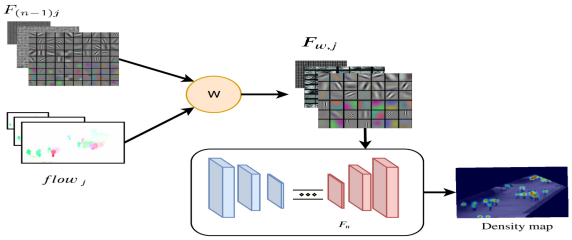

**Fig. 1.** Overview of the proposed approach to video-based crowd counting. Motion information is incorporated via a pyramid of optical flow that is computed from consecutive frames of the input video. The flow field is applied to multi-scale feature maps extracted from the previous frame via an image warp, $W$, and injected as an additional source of information into the decoder portion of the baseline network, which is described in Section 3.1.

methods, the overall person count can be attained by integrating over the entire density map. Thus, with regression-based approaches, the detection challenge is bypassed completely and the problem is transformed to that of training a CNN to learn the correspondence between an input image and a crowd-density map.

Although most prior work on crowd counting has focused on determining the number of people in a static image (e.g., [12, 8, 13–15, 6]), in most real-world settings, a video-stream is available. In such settings, it is natural to consider what techniques can leverage this additional temporal information and improve count accuracies. Intuitively, motion information can effectively remove false positives and negatives by combining information from neighboring frames, thus producing more temporally-coherent density maps. Moreover, temporal information can benefit occlusion scenarios where people are blocked from view in a specific frame, but are visible in surrounding frames.

One of the most well-studied representations of motion information in computer vision is optical flow, which can be computed using traditional (e.g., [16]) or deep learning (e.g., [17]) techniques. The fundamental idea explored in this paper is to improve crowd counting estimates in video by utilizing the motion information provided by explicitly-computed optical flow.

Figure 1 shows a conceptual overview of the proposed approach. The foundation of the method is a baseline CNN that receives a single image as input and produces a crowd density map as the output. In this work, a novel CNN is used that consists of two sub-sections: a feature extractor and a decoder. As shown in Figure 1, motion information is incorporated into the full system by computing a pyramid of optical flow from consecutive video frames. The multi-scale pyramid of flow is used to warp the previous frame's feature maps (i.e., feature embeddings) from the decoder sub-network toward the current frame. These warped feature maps are concatenated with the corresponding maps from the current frame. By complementing the decoder sub-network with motion information, the overall system is able to produce more temporally coherent density maps and achieve state-of-the-art accuracies.

There are four contributions of this paper:

- A novel video-based crowd counting system that incorporates motion information via a multi-scale embedding warp based on optical flow. To the best of our knowledge, integrating optical flow with a deep neural network has not been attempted previously for region of interest (ROI) crowd counting.
- An extensive evaluation on three video-based crowd counting datasets (UCSD [18], Mall [19] and Fudan-ShanghaiTech [9]) showing the proposed model outperforms all state-of-the-art algorithms.
- An illustration of the transfer learning abilities of the proposed approach, whereby knowledge learned in a source domain is shown to effectively transfer over to a target domain, using a small amount of training data. Here, the source and target domains correspond to two different scenes/environments observed in video datasets.
- Although not the primary focus, a secondary contribution is a new coarse-to-fine baseline CNN architecture for image-based crowd counting. This customized network is an extension of CSRNet [7], with a novel decoder sub-network. In an extensive evaluation on two challenging image datasets (UCF_CC_50 [20] and UCF-QNRF [20]), as well as the abovementioned three video datasets, this enhanced network meets or exceeds alternative state-of-the-art methods.

## 2 Related Work

### 2.1 Counting in static images

In recent years, most crowd counting systems are based on convolutional neural networks (CNNs). An early example of such an approach was that by Zhang et al. [12], which introduced a cross-scene crowd counting method by fine-tuning a CNN model to the target scene.

One of the major research directions within crowd counting is addressing the challenge of scale variation (e.g., [8, 13, 15, 6, 21, 22]). Specifically, a crowd counting system should produce accurate results regardless of the size of the people within the image. One such work that addresses this challenge proposed a multi-column architecture (MCNN) [8]. Other approaches have taken a different tack whereby coarse-to-fine architectures are used to produce high-resolution density maps (e.g., [15, 6]).

One work on image-based crowd counting by Li et al. [7] proposed a novel architecture called CSRNet that provides accurate estimates in crowded environments. CSRNet shares a similar network architecture to the baseline proposed here; however, their decoder sub-network uses dilated convolution to produce density maps that are $1/8^{th}$ of the input image size. In contrast, the proposed decoder has a deeper network structure and employs transposed convolution to attain density maps at the full image resolution.

Recently, PGCNet proposed a single column architecture to resolve intra-scene scale variation with the help of an autoencoder-based perspective estimation branch [23]. S-DCNet [22], which is another recent algorithm, operates in a divide-and-conquer fashion where feature maps are split until the person count within any one division is no greater than a set value. The system then classifies person counts into a set of intervals to determine the best count for each division. In contrast to S-DCNet, our baseline does not require any classification stages and is independent of assumptions regarding person counts within a division, such as interval ranges.

### 2.2 Video-based counting methods

Most previous works in crowd counting focus on the single image setting; there are much fewer examples of video-based crowd counting in the literature. Within the video domain, two sub-problems have emerged for crowd counting: region of interest (ROI) [11, 10, 9] and line of interest (LOI) [24–26]. For ROI counting, the number of people within a certain image region (or, the entire image) is estimated; whereas, for LOI counting, a virtual line in the image is specified and the task is to determine the number of individuals that cross this line.

Several LOI works extract temporal slices from the line of interest to detect the transient crossing events [27, 28, 26]. Challenges for these approaches include foreground blob detection and processing, as well as disentangling confounding variables (e.g., blob widths are affected by number of people as well as velocity). More recent LOI counting work has considered using deep neural networks, including one system that included an ROI counting sub-module [24]. Although ROI and LOI counting share common challenges (e.g., perspective distortion, scale variation, occlusions), the specialized problem definition tends to drive different technical approaches, which are not typically directly transferable. The methods proposed in the current work focus on ROI counting, which will be referred to simply as crowd counting in the remainder of the paper.

For video-based crowd counting, a significant open problem is how to best leverage temporal information to improve count estimates. In one such work, ConvLSTMs were used to integrate image features from the current frame with those from previous frames for improved temporal coherency [11]. Further, Zhang et al. [10] proposed the use of LSTMs for vehicle counting in videos. Most of the LSTM-based approaches suffer from the drawback that they require a predefined number of frames to use as 'history' and, depending on dataset, some of these frames may provide irrelevant or contradictory information.

Fang et al. [9] updated their model parameters using information based on dependencies among neighbouring frames, rather than via an LSTM. However, in their approach, a density regression module was first used in a frame-by-frame fashion to produce regression maps, upon which a spatial transformer was applied to post-process and improve the estimates. Although focusing on LOI counting, Zhao et al. used a convolutional neural network that processed pairs of video frames to jointly estimate crowd density and velocity maps [24]. The estimated velocity maps differ from dense optical flow in that they only have non-zero values in the locations of pedestrians.

The sole work that we are aware of that has incorporated optical flow for ROI crowd counting is a classical approach using traditional computer vision techniques (e.g., background subtraction and clustering the flow vectors) [29]. Their proposed system includes numerous hand-tuned parameters and employed the assumption that the only moving objects in the scene are pedestrians, which is not realistic in most scenarios. Differing from the above, our proposed approach integrates optical flow-based motion information directly, by warping deep neural network feature maps from the previous frame to the next.

### 2.3 Optical flow pyramid

Many recent works applying CNNs to video data have demonstrated the benefit of including optical flow. Two-stream and multi-stream networks have already shown effectiveness for action recognition [30, 31] and action detection [32–35]. These approaches mostly use optical flow as an additional, parallel source of information which is fused prior to prediction. Other work has utilized optical flow to warp intermediate network features to achieve performance speed-ups for video-based semantic segmentation [36, 37] and object detection [36].

Most similar to the current work is an approach to semantic segmentation in video that introduces a "NetWarp" module [38]. This module utilizes optical flow, modified by a learned transformation, to warp feature maps between consecutive frames and subsequently combine them with maps from the current frame, resulting in more stable and consistent segmentation results. In contrast, our proposed solution adopts an optical flow pyramid to capture motion at multiple scales and applies the unmodified flow to the feature maps directly for the task of crowd counting. To the best of our knowledge, no prior work has made use of optical flow-based feature map warping for video-based crowd counting, as proposed here.

## 3 Technical Approach

### 3.1 Crowd counting baseline network

The baseline network serves as a single-frame crowd density estimator and contains two sub-modules: a feature extractor and a decoder. Although it is not the primary technical contribution of this work, the baseline network extends CSRNet [7], yielding a significantly more accurate density estimator. These extensions will be highlighted in the following.
**Feature extractor:** A customized VGG-16 network [39], initialized with ImageNet [40] weights was selected as the feature extractor in order to perform fair comparison with other methods [7, 6] using the same backbone network. To avoid feature maps with small spatial extent, three maxpool layers were used, which results in feature maps of $1/8^{th}$ of the input image size at the bottleneck. Differing from [7], ReLU activation functions were replaced with PReLU [41] for each layer to avoid the 'dying ReLU' problem.
**Decoder network:** The decoder of CSRNet consists of six dilated convolution layers followed by a $1 \times 1$ convolution, resulting in an output density map that is

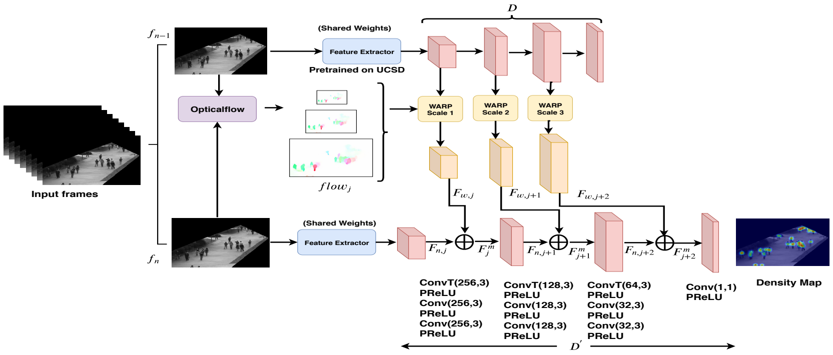

**Fig. 2.** System diagram for MOPN. The input image is passed through the feature extractor and optical flow is computed between the previous and current frame. Multi-scale feature maps from the previous frame are warped via the computed optical flow and concatenated with the corresponding feature maps in the current frame. This step combines complementary scale-aware motion based features with traditional, appearance-derived features in the proposed network. The crowd count can be obtained by summing over the entries in the predicted density map provided by the $1 \times 1$ convolution layer.

$1/8^{th}$ the size of the input image. In contrast, the proposed decoder is comprised of nine convolutional layers, three transposed convolution layers, followed by a final $1 \times 1$ convolution layer. This modified decoder design results in coarse-to-fine feature maps and a high-resolution (i.e., same as input size) density map as the final output. The ReLU activation functions in CSRNet were also replaced with PReLU throughout the decoder.

The main motivation for these architectural changes was three-fold: i) The proposed coarse-to-fine design eases the integration with the optical flow pyramid in the full, proposed model. ii) By using transposed convolution, the decoder output is full-resolution, making it more practical to resolve small humans within the image. iii) The additional learnable parameters introduced by the extra convolutional layers and PReLU activation functions empirically leads to significantly improved accuracies. In Section 4.2 and Section 4.3, the performance of the proposed baseline network is compared against state-of-the-art methods.

### 3.2 Multi-scale optical flow pyramid network (MOPN)

The general philosophy of the proposed full model is to leverage complementary appearance and motion information to improve counting accuracies.

One challenge with optical flow is effectively capturing large object displacements while simultaneously maintaining accuracies for small movements. Video camera configurations for crowd density estimation are varied: Some cameras are high resolution and have frame rates of 30 fps (e.g., FDST Dataset [9]), while others may be low resolution with frame rates of 2 fps or lower (e.g., Mall Dataset [19]). For a 30 fps video, the inter-frame motion of objects tends to be small, but for cameras running at 2 fps, scene objects can move significantly between consecutive frames.

To accommodate the range of inter-frame motion that may be encountered in crowd counting scenarios, an image pyramid is utilized when computing optical flow. With this method, large pixel displacements will be captured by the coarse scales, which are subsequently refined to have higher precision at the finer scales. This pyramid of multi-resolution optical flow maps is then applied to the corresponding feature maps from the decoder network. With this approach, both large and small displacements are modeled and addressed.

In detail, let $f_n$ and $f_{n-1}$ represent the current and previous input video frames, respectively. The proposed approach computes optical flow between $f_n$ and $f_{n-1}$ at three scales in an image pyramid, using a pixel subsampling factor of two between pyramid layers. As shown in Fig. 2, Scale 1 ($S_1$) captures large inter-frame displacements found in the video, while Scale 3 ($S_3$) effectively captures small motions that would typically be found in 30 fps video. The middle scale, $S_2$, describes mid-range optical flow bounded by $S_1$ and $S_3$. FlowNet 2.0 [17] is employed for computing the flow in the current work, although the overall approach is agnostic to the specific optical flow algorithm adopted.

As shown in Fig. 2, once the multi-scale pyramid of optical flow is computed, each flow map is applied as a warping transformation to the feature maps at the corresponding pixel resolution from the previous frame. The warped feature map is then concatenated with the corresponding embedding computed for the current frame. By including the motion information via the previous frame's warped feature maps, MOPN achieves improved temporal consistency and robustness when appearance information is unreliable (e.g., partial occlusions, objects with human-like appearances).

### 3.3 Training details

The training method for the proposed MOPN system consists of two steps: baseline network training and full model training. Baseline network training proceeds by initializing the network with ImageNet weights, from which it is subsequently updated. During this stage, a dataset is selected (e.g., UCSD [18]) and the network is trained using samples from that dataset. Based on the validation samples, the best model is selected and evaluated on the test samples. All images and corresponding ground truth are resized to 952 × 632. In Fig. 2, the upper portion of the network depicts the baseline network.

For the full MOPN model, the parameters of the feature extractor portion of the network, $\theta_z$, are initialized with the corresponding baseline pretrained weights, $\theta_P$, and frozen. To incorporate motion information into MOPN, the baseline decoder, $D$, is replaced with a trainable, motion-based decoder, $D^{'}$. For every frame, the image is first downsampled to create a three-level image pyramid from which optical flow is calculated to yield $flow_j$, where $j$ is the pyramid level.

For each epoch, $i$, training of the MOPN motion-based decoder proceeds as follows. The feature maps for the previous frame, $n-1$, and current frame, $n$, are computed using, $F_{(n-1)j} = \mathcal{Z}_j^{'}(f_{n-1}, \theta_{D'_{n-1}})$ and $F_{nj} = \mathcal{Z}_j^{'}(f_n, \theta_{D'_{n-1}})$, respectively. The term $\mathcal{Z}_j^{'}$ denotes the nonlinear network function that produces the feature maps at network layer $j$ for an input image. Warped versions of the feature maps from the previous frame are calculated according to $F_{wj} = \text{WARP}(F_{(n-1)j}, flow_j)$ which are then concatenated with $F_{nj}$, the feature maps of the current frame. Feature map concatenation results in the formation of higher dimensional maps, $F_j^m$, which are subsequently used to update the motion decoder and obtain a new set of parameters $\theta_{D'_n}$. Intuitively, the intermediate layer outputs from every frame are propagated forward to the next frame in order to train the decoder of MOPN. Note that for the special case of $n = 2$, the baseline decoder network is used for feature map generation, as the shared parameters within the MOPN decoder are initialized by the frozen baseline decoder parameters, $\theta_D$.

Regarding the loss function, the difference between the predicted density map and ground truth is measured by Frobenius Norm. Namely, the loss function is:

$$L(\theta) = \frac{1}{2N} \sum_{n=1}^{N} ||M(f_n, \theta) - M_n^{GT}||_2^2, \tag{1}$$

where, $N$ is the number of training frames, $M(f_n, \theta)$ is the predicted density map and $M_n^{GT}$ is the corresponding ground truth.

For all experiments in the paper, we use the following hyperparameter settings across all the datasets: learning rate = 0.00001, number of epochs = 2000, batch size = 2 (two consecutive frames at a time) with the Adam optimizer [42]. A summary of the training procedure for updating the MOPN decoder is provided in Algorithm 1.

**Ground truth generation:** For crowd density estimation, ground truth generation is very important in order to ensure fair comparison. To remain consistent with previous research, the same approaches described in [6, 7, 11, 9] were used to generate the ground truth density maps in the current paper. For the datasets in which a ROI mask is provided, the ROI was multiplied with each frame to allow density maps to be generated based on the masked input images.

## 4 Experiments

### 4.1 Evaluation metric

Following previous works [8, 5–7], Mean Absolute Error (MAE) and Mean Square Error (MSE) are used as evaluation metrics. Let $N$ be the number of test images, $\mathrm{C}_{gt}^{(n)}$ the ground truth count, and $\mathrm{C}^{(n)}$ be the predicted count for the $n$-th test image. These two evaluation metrics are defined as follows: MAE $= \frac{1}{N}\sum_{n=1}^{N} |\mathrm{C}^{(n)} - \mathrm{C}gt^{(n)}|$ and MSE $=\sqrt{\frac{1}{N}\sum_{n=1}^{N} |\mathrm{C}^{(n)} - \mathrm{C}gt^{(n)}|^2}$.

**Algorithm 1:** MOPN training procedure.

```
Input: Frame sequence $\{f_n\}_{n=1}^{N}$ with ground truth density maps $\{M_n^{GT}\}$
Output: Trained parameters $\theta_{D'}$
/* $\theta_z$ denotes parameters of the MOPN feature extractor */
/* $\theta_{D'}$ denotes parameters of the MOPN decoder */
/* $\theta_P$ denotes parameters of base network */
Initialize $\theta_z$ and $\theta_{D'}$ with $\theta_P$
Freeze $\theta_z$
/* $T$ denotes the maximum number of epochs. */
for $i = 1$ to $T$ do
    for $n = 2$ to $N$ do
        Extract $\{F_{(n-1)j}\}_{j=1}^{3}$ from $f_{(n-1)}$
        Extract $\{F_{nj}\}_{j=1}^{3}$ from $f_n$
        /* $\{F_{nj}\}_{j=1}^{3}$ denotes $F$ as the feature map output for the $n^{th}$
           frame with $j^{th}$ scale */
        for $j = 1$ to $3$ do
            $flow_j = Optical\_flow(f_{(n-1)j}, f_{nj})$
            $F_{wj} = WARP(F_{(n-1)j}, flow_j)$
            $F_j^m = F_{wj} \oplus F_{nj}$
        /* From Eq. 1 */
        $loss^{best} = argmin[L(\theta)]$
        Backpropagate and update $\theta_{D'}$
```

### 4.2 Crowd Counting in Images

**UCF_CC_50:** The UCF_CC_50 dataset [20] is a benchmark for crowd counting in static images focusing on dense crowds captured from a wide-range of locations around the world. The images in this dataset do not come from a video camera, meaning that it can not be used to test the full, proposed MOPN model; however, the proposed baseline model is evaluated on this dataset. To ensure a fair comparison, 5-fold cross validation was performed, as was done for S-DCNet [22]. As shown in Table 2, the propoed baseline attains the best MAE and second best MSE scores against the alternative approaches. Only DRSAN [43] slightly outperforms our baseline under the MSE mertic.

**UCF-QNRF:** UCF-QNRF [44] is a large crowd counting dataset consisting of 1535 high-resolution images and 1.25 million head annotations. This dataset focuses primarily on dense crowds, with an average of roughly 815 persons per image. The training split is comprised of 1201 images, with the remaining left for testing. During training, we follow the data augmentation techniques described in [22]. Also, we resized the images to $1/4^{th}$ of their original size.

The results on this dataset from the proposed baseline are impressive, attaining the best result for both MAE and MSE. This result clearly indicates the

**Table 1.** Performance comparisons on UCF_CC_50 [20] and UCF-QNRF [44] datasets. For this and subsequent tables throughout the paper, **blue** numbers refer to the best result in each column, while **red** numbers indicate second best.

| Methods | UCF_CC_50 | | UCF-QNRF | |
|---|---|---|---|---|
| | MAE | MSE | MAE | MSE |
| Idrees *et al.* [3] | 468.0 | 590.3 | 315 | 508 |
| Context-Aware Counting [45] | 212.2 | 243.7 | 107 | 183 |
| ADCrowdNet [46] | 257.1 | 363.5 | - | - |
| MCNN [8] | - | - | 277 | 426 |
| CMTL [15] | - | - | 252 | 514 |
| Switching-CNN [6] | - | - | 228 | 445 |
| Cross Scene [12] | 467.0 | 498.5 | - | - |
| IG-CNN [47] | 291.4 | 349.4 | - | - |
| D-ConvNet [48] | 288.4 | 404.7 | - | - |
| CSRNet [7] | 266.1 | 397.5 | - | - |
| SANet [49] | 258.4 | 334.9 | - | - |
| DRSAN [43] | 219.2 | **250.2** | - | - |
| PGC [23] | 244.6 | 361.2 | - | - |
| TEDnet [50] | 249.4 | 354.5 | 113 | 188 |
| MBTTBF-SCFB [51] | 233.1 | 300.9 | **97.5** | **165.2** |
| S-DCNet [22] | **204.2** | 301.3 | 104.4 | 176.1 |
| Proposed baseline (w/o optical flow) | **181.8** | **260.4** | **78.65** | **140.63** |

effectiveness of the proposed baseline network, as it is able to outperform the latest state-of-the-art methods on large-scale datasets with dense crowds.

**Table 2.** Comparative performance of the proposed baseline and full model (MOPN) against state-of-the-art alternatives on three standard datasets.

| Methods | UCSD | | MALL | | FDST | |
|---|---|---|---|---|---|---|
| | MAE | MSE | MAE | MSE | MAE | MSE |
| Switching CNN [6] | 1.62 | 2.10 | — | — | — | — |
| CSRNet [7] | 1.16 | 1.47 | — | — | — | — |
| MCNN [8] | 1.07 | **1.35** | — | — | 3.77 | 4.88 |
| Count Forest [52] | 1.60 | 4.40 | 2.50 | 10.0 | — | — |
| Weighted VLAD [53] | 2.86 | 13.0 | 2.41 | 9.12 | — | — |
| Random Forest [54] | 1.90 | 6.01 | 3.22 | 15.5 | — | — |
| LSTN [9] | 1.07 | 1.39 | 2.03 | 2.60 | **3.35** | **4.45** |
| FCN-rLSTM [10] | 1.54 | 3.02 | — | — | — | — |
| Bidirectional ConvLSTM [11] | 1.13 | 1.43 | 2.10 | 7.60 | 4.48 | 5.82 |
| Proposed baseline (w/o optical flow) | **1.05** | 1.74 | **1.79** | **2.25** | 3.70 | 4.80 |
| Full proposed model (MOPN) | **0.97** | **1.22** | **1.78** | **2.25** | **1.76** | **2.25** |
| % Improvement: MOPN over Baseline | 7.6% | 29.9% | 0.6% | 0.0% | 52.4% | 53.1% |

### 4.3 Crowd Counting in Videos

**UCSD Dataset:** The UCSD dataset consists of a single 2,000 frame video taken with a stationary camera overlooking a pedestrian walkway. The video was captured at 10 fps and has a resolution of 238 × 158. The provided ground truth denotes the centroid of each pedestrian. Following the common evaluation protocol for this dataset (e.g., [18]), Frames 601–1,400 are used for training, while the remaining images are used during testing.

The MAE and MSE results for the baseline (without optical flow) and MOPN are shown in Table 2. The full proposed model, MOPN, attains second-best MAE and MSE, slightly behind while the baseline has second-based MAE results. For MAE, MOPN offers a 9% improvement over the third best result (MCNN [8] and LSTN [9]), while a 10% decrease in MSE is observed compared to the second-best results (of MCNN [8]). Compared to the baseline, the full proposed model provides a 7.6% and 29.9% improvement for MAE and MSE, respectively. This final result demonstrates clearly the benefits of incorporating motion information to complement the appearance cues that are traditionally used for crowd counting.

**Mall Dataset** The mall dataset is comprised of a 2,000 frame video sequence captured in a shopping mall via a publicly accessible camera. The video was captured at a resolution of 640 × 480 and with a framerate of less than 2 fps. As was done in [19], Frames $1 - 800$ were used for training, while the final 1,200 frames were considered for evaluation.

As Table 2 indicates, MOPN and the proposed baseline achieve the best and second best results on this dataset, respectively. Although the MAE with MOPN is better than the baseline, in this case the improvement from motion-related information is marginal. This result is expected, as the frame rate for the Mall Dataset is low. With such a low frame rate, the inter-frame motion of people in the scene can be quite large (e.g., one quarter of the image), meaning that only the scales of the optical flow pyramid corresponding to large displacements are contributing to the full network. The results from the Mall Dataset are encouraging, as they indicate that even in low framerate settings when motion cues are less effective, the full model can rely on the appearance information provided by the baseline network to still achieve state-of-the-art accuracies.

**Fudan-ShanghaiTech Dataset** The Fudan-ShanghaiTech (FDST) dataset [9] is currently the most extensive video crowd counting dataset available with a total of 15,000 frames and 394,081 annotated heads. The dataset captures 100 videos from 13 different scenes at resolutions of 1920 × 1080 and 1280 × 720.

Following the evaluation protocol defined by the dataset authors, 60 of the videos are used for training while the remaining 40 videos are reserved for testing. Table 2 shows the results for the FDST dataset. Since this dataset is new, only three alternative state-of-the-art approaches have reported results for comparison. MOPN has the lowest MAE and MSE, while the proposed baseline was third-best. MOPN achieves a 47% and 49% improvement over the second-best performer, LSTN [9], for MAE and MSE, respectively. To attain this significant of an accuracy increase on the largest video-based crowd counting dataset illustrates the importance of combining both appearance and motion cues.

### 4.4 Qualitative Results

To demonstrate the qualitative performance of the proposed system, Fig. 3 shows a zoomed image from the FDST dataset along with superimposed density maps corresponding to ground truth, proposed baseline, and MOPN. The qualitative results show that MOPN produces much more accurate count estimates than the baseline. It can be seen that the baseline model (third column) does not detect three individuals (denoted by red circles); whereas, MOPN (fourth column) is able to detect these individuals (highlighted with green circles).

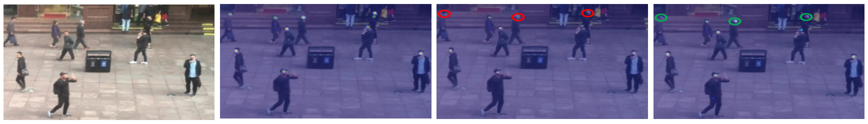

**Fig. 3.** Qualitative example of density maps. From left to right, the columns correspond to a cropped input video frame from the FDST dataset [9], ground truth density map, density map from the proposed baseline (without optical flow), and the density map from the full MOPN model. Superimposed red and green circles highlight certain false negatives and true positives, respectively. Best viewed in color and with magnification.

### 4.5 Transfer Learning

The goal of this experiment is to consider the performance tradeoffs when only a portion of the network is fine-tuned on a target domain dataset. This scenario can be relevant in situations in which the amount of data in the target domain is limited and therefore it may be more effective to train only a specific portion of the network. The transfer learning experiment is setup as follows. First, the baseline model is trained on a source domain dataset. Once this source domain baseline is in place, the trained model is evaluated on a target domain test dataset. In the finetuning setting, we simply update the decoder of our baseline model. Table 3 shows the results for this evaluation, where alternative methods that have considered such transfer learning experiments have been included. In addition to several deep learning-based approaches detailed earlier, some methods that do not involve deep learning are also included, as follows: Feature Alignment (FA) [55], Learning Gaussian Process (LGP) [56], Gaussian process (GP) [57], Gaussian Process with Transfer Learning (GPTL) [57]. The proposed fine-tuned baseline model achieves the best MAE compared to the other models on the transfer learning experiment.

**Table 3.** Results from the transfer learning experiment using the Mall and UCSD datasets. The finetuned baseline model attains best results when completing the transfer learning task from UCSD to MALL, as well as from MALL to UCSD.

| Methods | UCSD to MALL | MALL to UCSD |
|---|---|---|
| | MAE | MAE |
| FA [55] | 7.47 | 4.44 |
| LGP [56] | 4.36 | 3.32 |
| GPA [57] | 4.18 | 2.79 |
| GPTL [57] | 3.55 | 2.91 |
| MCNN [8] | 24.25 | 11.26 |
| CSRNet [7] | 14.01 | 13.96 |
| Bidirectional ConvLSTM [11] | **2.63** | **1.82** |
| Proposed baseline (w/o optical flow) | 6.18 | 12.21 |
| Finetuned baseline model | **2.36** | **1.55** |

### 4.6 Ablation Studies

**Component analysis:** Table 4 shows a study regarding the performance gains due to the individual extensions of the proposed baseline over CSRNet. The first row from Table 4 corresponds to a network comparable to CSRNet, while the fourth row is the proposed baseline. Rows two and three show the individual contributions of transposed convolution and PReLU, when integrated into the decoder portion of the baseline network. As shown in the table, both modifications contribute evenly to the accuracy gains. Also, the alterations are complementary, leading to further improved results when combined (Row 4).

**Table 4.** Individual contributions of network components in the baseline network.

| Methods | UCSD MAE |
|---|---|
| ReLU (w/o transposed convolution) | 1.26 |
| ReLU (with transposed convolution) | 1.18 |
| PReLU (w/o transposed convolution) | 1.14 |
| PReLU (with transposed convolution) | **1.05** |

**Multi-scale pyramid:** One of the main parameters of MOPN is the number of layers in the optical flow pyramid for warping the feature maps. Table 5 shows the proposed method's performance on UCSD as a function of the number of levels in the optical flow pyramid. With only a single pyramid level, the warping and feature concatenation can be performed at low, mid, or high resolution, corresponding to specialization in capturing large, medium, and small-scale motions.

The table shows that the multi-scale optical flow pyramid indeed yields best accuracies. When using just a single scale of optical flow, Scale 3 (small inter-frame displacements) performs slightly better than Scale 1 (large inter-frame displacements), but the difference is minimal.

**Table 5.** The effect of modifying the number of optical flow pyramid levels.

| Methods | UCSD | |
|---|---|---|
| | MAE | MSE |
| Proposed with Scale-1 | 1.07 | 1.34 |
| Proposed with Scale-3 | 1.04 | 1.30 |
| Proposed multi-scale | **0.97** | **1.22** |

**Effect of optical flow warping:** Another ablation study considers providing the full proposed network with two frames of input images without any explicit optical flow. This experiment was performed by concatenating the unwarped feature maps from the previous frame with those of the current frame. For the UCSD dataset, this configuration yielded MAE/MSE = 1.12/1.97 compared to 0.97/1.22 for MOPN (Table 2). Also note that this two-frame configuration is worse than the proposed baseline (1.05/1.74 from Table 2). This finding exemplifies the importance of optical flow to the proposed approach. Without warping, features from previous and current frames are misaligned, which confuses the network, as it is not provided with the necessary motion information to resolve correspondences across the feature maps. With MOPN, optical flow removes this ambiguity, constraining the solution space and yielding less localization error.

## 5 Conclusion

In this paper, a novel video-based crowd density estimation technique is proposed that combines a pyramid of optical flow features with a convolutional neural network. The proposed video-based approach was evaluated on three challenging, publicly available datasets and universally achieved best MAE and MSE when compared against nine recent and competitive approaches. Accuracy improvements of the full proposed MOPN model were as high as 49% when compared to the second-best performer on the recent and challenging FDST video dataset. These results indicate the importance of using all spatiotemporal information available in a video sequence to achieve highest accuracies rather than employing a frame-by-frame approach. Additionally, results on the UCF_CC_50 and UCF-QNRF datasets, which focus on images of dense crowds, show that the proposed baseline network (without optical flow) achieves state-of-the-art performance for crowd counting in static images.

## References

1. Sindagi, V.A., Patel, V.M.: A survey of recent advances in CNN-based single image crowd counting and density estimation. Pattern Recognition Letters **107** (2018) 3–16
2. Ge, W., Collins, R.T.: Marked point processes for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2009) 2913–2920
3. Idrees, H., Soomro, K., Shah, M.: Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015) 1986–1998
4. Hu, P., Ramanan, D.: Finding tiny faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1522–1530
5. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid cnns. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1861–1870
6. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 4031–4039
7. Li, Y., Zhang, X., Chen, D.: CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1091–1100
8. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 589–597
9. Fang, Y., Zhan, B., Cai, W., Gao, S., Hu, B.: Locality-constrained spatial transformer network for video crowd counting. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE (2019) 814–819
10. Zhang, S., Wu, G., Costeira, J.P., Moura, J.M.: FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3667–3676
11. Xiong, F., Shi, X., Yeung, D.Y.: Spatiotemporal modeling for crowd counting in videos. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 5151–5159
12. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 833–841
13. Onoro-Rubio, D., López-Sastre, R.J.: Towards perspective-free object counting with deep learning. In: European Conference on Computer Vision, Springer (2016) 615–629
14. Boominathan, L., Kruthiventi, S.S., Babu, R.V.: Crowdnet: A deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM International Conference on Multimedia, ACM (2016) 640–644
15. Sindagi, V.A., Patel, V.M.: CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE (2017) 1–6
16. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artifical Intelligence **17** (1981) 185–203

17. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1647–1655
18. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–7
19. Chen, K., Loy, C.C., Gong, S., Xiang, T.: Feature mining for localised crowd counting. In: Proceedings of the British Machine Vision Conference. (2012) 1–7
20. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 2547–2554
21. Hossain, M., Hosseinzadeh, M., Chanda, O., Wang, Y.: Crowd counting using scale-aware attention networks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2019) 1280–1288
22. Xiong, H., Lu, H., Liu, C., Liu, L., Cao, Z., Shen, C.: From open set to closed set: Counting objects by spatial divide-and-conquer. In: The IEEE International Conference on Computer Vision (ICCV). (2019)
23. Yan, Z., Yuan, Y., Zuo, W., Tan, X., Wang, Y., Wen, S., Ding, E.: Perspective-guided convolution networks for crowd counting. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 952–961
24. Zhao, Z., Li, H., Zhao, R., Wang, X.: Crossing-line crowd counting with two-phase deep neural networks. In: European Conference on Computer Vision, Springer (2016) 712–726
25. Ma, Z., Chan, A.B.: Crossing the line: Crowd counting by integer programming with local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 2539–2546
26. Ma, Z., Chan, A.B.: Counting people crossing a line using integer programming and local features. IEEE Transactions on Circuits and Systems for Video Technology **26** (2015) 1955–1969
27. Cong, Y., Gong, H., Zhu, S.C., Tang, Y.: Flow mosaicking: Real-time pedestrian counting without scene-specific learning. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2009) 1093–1100
28. Cao, L., Zhang, X., Ren, W., Huang, K.: Large scale crowd analysis based on convolutional neural network. Pattern Recognition **48** (2015) 3016–3024
29. Fujisawa, S., Hasegawa, G., Taniguchi, Y., Nakano, H.: Pedestrian counting in video sequences based on optical flow clustering. International Journal of Image Processing **7** (2013) 1–16
30. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1933–1941
31. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4724–4733
32. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. (2014) 568–576
33. Peng, X., Schmid, C.: Multi-region two-stream R-CNN for action detection. In: European Conference on Computer Vision, Springer (2016) 744–759
34. Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M.: A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Pro-

ceedings of the 27th International Conference on Neural Information Processing Systems. (2016) 1961–1970

35. Singh, G., Saha, S., Sapienza, M., Torr, P., Cuzzolin, F.: Online real-time multiple spatiotemporal action localisation and prediction. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3657–3666
36. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4141–4150
37. Li, M., Sun, L., Huo, Q.: Flow-guided feature propagation with occlusion aware detail enhancement for hand segmentation in egocentric videos. Computer Vision and Image Understanding **187** (2019) 102785
38. Gadde, R., Jampani, V., Gehler, P.V.: Semantic video CNNs through representation warping. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 4463–4472
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
40. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. (2012) 1097–1105
41. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1026–1034
42. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations. (2014)
43. Liu, L., Wang, H., Li, G., Ouyang, W., Lin, L.: Crowd counting using deep recurrent spatial-aware network. arXiv preprint arXiv:1807.00601 (2018)
44. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 532–546
45. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
46. Wan, J., Chan, A.: Adaptive density map generation for crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019)
47. Babu Sam, D., Sajjan, N.N., Venkatesh Babu, R., Srinivasan, M.: 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. (2018) 3618–3626
48. Shi, Z., Zhang, L., Liu, Y., Cao, X., Ye, Y., Cheng, M.M., Zheng, G.: Crowd counting with deep negative correlation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5382–5390
49. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 734–750
50. Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D., Shao, L.: Crowd counting and density estimation by trellis encoder-decoder networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6133–6142

51. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019)
52. Pham, V.Q., Kozakaya, T., Yamaguchi, O., Okada, R.: Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3253–3261
53. Sheng, B., Shen, C., Lin, G., Li, J., Yang, W., Sun, C.: Crowd counting via weighted VLAD on a dense attribute feature map. IEEE Transactions on Circuits and Systems for Video Technology **28** (2016) 1788–1797
54. Xu, B., Qiu, G.: Crowd density estimation based on rich features and random projection forest. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2016) 1–8
55. Change Loy, C., Gong, S., Xiang, T.: From semi-supervised to transfer counting of crowds. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 2256–2263
56. Yu, K., Tresp, V., Schwaighofer, A.: Learning gaussian processes from multiple tasks. In: Proceedings of the 22nd International Conference on Machine Learning (ICML-05). (2005) 1012–1019
57. Liu, B., Vasconcelos, N.: Bayesian model adaptation for crowd counts. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 4175–4183