# Almost Unsupervised Learning for Dense Crowd Counting

**Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, R. Venkatesh Babu**
Video Analytics Lab, Indian Institute of Science,
Bangalore 560012, India
deepaksam@iisc.ac.in, nnsajjan@gmail.com, himanshu.mib@gmail.com, venky@iisc.ac.in

## Abstract

We present an unsupervised learning method for dense crowd count estimation. Marred by large variability in appearance of people and extreme overlap in crowds, enumerating people proves to be a difficult task even for humans. This implies creating large-scale annotated crowd data is expensive and directly takes a toll on the performance of existing CNN based counting models on account of small datasets. Motivated by these challenges, we develop Grid Winner-Take-All (GWTA) autoencoder to learn several layers of useful filters from unlabeled crowd images. Our GWTA approach divides a convolution layer spatially into a grid of cells. Within each cell, only the maximally activated neuron is allowed to update the filter. Almost 99.9% of the parameters of the proposed model are trained without any labeled data while the rest 0.1% are tuned with supervision. The model achieves superior results compared to other unsupervised methods and stays reasonably close to the accuracy of supervised baseline. Furthermore, we present comparisons and analyses regarding the quality of learned features across various models.

## Introduction

Counting people in crowds, though a necessity in many practical scenarios, is very challenging. Typical dense crowds with thousands of people lend any naive person detector fruitless. This is because of the absence of consistent observable features like face, body parts etc. owing to extreme occlusion, pose variations and background clutter. In acute crowding, detecting people appearing as just blobs is laborious and difficult even for humans (see Figure 1). The visual patterns that need to be learned for detecting people, vary drastically from sparse to extreme dense crowds. As a result, any crowd counting system has to model such a huge diversity of appearance of people, requiring large annotated datasets for training. The performance of models based on Convolutional Neural Networks (CNN), in general, is directly related to the availability of large datasets encompassing the entire diversity. However, due to annotation difficulty, the datasets available for dense crowd counting are small, with the current largest one having only 482 images with 0.2 million person annotations. This seriously limits the advances in annotation intensive problems like dense crowd counting. Hence we formulate the objective of this work as to train crowd counting models to the maximum extent with unlabeled data. To the best of our knowledge, there are no other works in this direction for dense crowd counting and is expected to fuel more research in the area.



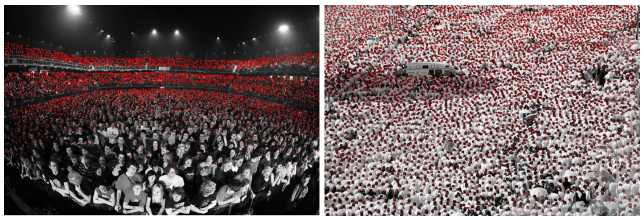

Figure 1: Typical dense crowd images with human annotations (from Part_A of Shanghaitech dataset (Zhang et al. 2016)). Sheer density of the crowd compounded with severe occlusions render crowd annotation challenging.

Existing unsupervised methodologies are mostly based on autoencoders. They learn features by training to predict its own input (Hinton and Salakhutdinov 2006; Vincent et al. 2008) or some function of the input (Larsson, Maire, and Shakhnarovich 2017; Agrawal, Carreira, and Malik 2015; Wang and Gupta 2015; Pathak et al. 2016; Doersch, Gupta, and Efros 2015; Noroozi and Favaro 2016). It has been shown that many autoencoder based approaches fail to learn useful features (Makhzani and Frey 2015). When applied on highly diverse dense crowd images, we show that current unsupervised methods do not learn enough useful features for density regression as evidenced from their performance scores. In order to improve feature learning from unlabeled crowd images, we consider winner-take-all (WTA) regularization for autoencoders. WTA autoencoder proposed by (Makhzani and Frey 2015), is inspired from the behavior of actual neuron adaptation in human brain. The basic idea of WTA approach is to selectively perform learning for neurons in the autoencoder. This means not all neurons are allowed to update their weights at a particular iteration, creating a race among neurons to learn a feature and get specialized. The "winner" neuron is the one which has the highest activation value. This loosely tries to model the inhibition mechanism seen in brain neurons. It has been shown

that WTA auto-encoders acquire better features than normal autoencoders (Makhzani and Frey 2015). Till now WTA models have only been evaluated on datasets like MNIST, CIFAR etc. and are not scalable to highly diverse scenarios like dense crowds. Hence we significantly modify the WTA training methodology and develop Grid Winner-Take-All (GWTA) convolutional autoencoders to handle huge diversity in crowd scenes.

In a nutshell, GWTA spatially divides each convolutional feature map into a grid of cells, where WTA is applied in each cell. This allows local winners in a fixed neighborhood rather than global ones as in WTA autoencoder. Hence, GWTA autoencoder is able to leverage diversity of features across space, allowing scalable and efficient training with diverse crowd data. Our crowd counting system is composed of a CNN regressor, for which we train several layers in an unsupervised manner, i.e. using only crowd images and no annotation. Each layer of the model is trained separately as an GWTA autoencoder to reconstruct its own input. This stacked autoencoder training progressively learns a hierarchy of discriminative features frequently appearing in crowd images. Majority of the parameters of the network, almost 99.9%, are trained in this manner without any labeled supervision. This is followed by supervised training of the remaining parameters to get the final crowd density regressor. But note that the layers trained in unsupervised manner are frozen and only the last two layers which take unsupervised representations as input are tuned with labeled data. This way our model leverages unlabeled data for training majority of its parameters and only require labeled examples to adjust very few parameters (less than 0.1%).

As a summary, this work contributes the following:

- A stacked convolutional autoencoder model based on grid winner-take-all (GWTA) paradigm for large-scale unsupervised feature learning.
- The first crowd counting system that can train almost 99.9% of its parameters without any annotated data.

## Previous Work

All learning based previous works in dense crowd counting require labeled data. The dominant methodology in the field is to learn a regressor to estimate crowd density map rather than predicting the crowd count directly. Some early works like (Wang et al. 2015) attempt to regress the count directly, but only to be outperformed by density regression based models as they can acquire better features. Zhang et al. (2015) optimize their counting CNN by back-propagating both crowd density loss as well as crowd count loss in an alternate fashion. In order to tackle large scale variations in crowd scenes, multi-scale models are introduced. Onoro et al. (2016) have a set of CNNs, each specific to one particular scale with their outputs fused at the end through learnable layers. Similar approach is employed with multi-column network by (Boominathan, Kruthiventi, and Babu 2016), where they use a combination of shallow and deep CNN. Zhang et al. (2016) fuse output from three CNN columns with different receptive fields to capture crowds at multiple scales. Further improvement on multi-column models is achieved by (Babu Sam, Surya, and Babu 2017; Babu Sam et al. 2018), where the CNN columns are forced to get specialized aggressively through a differential training procedure. Adding auxiliary information, low-level or scene-level, in the form of confidences over crowd density types to the regressor network is shown to improve performance (Sindagi and Patel 2017a; 2017b). In this case, separate networks need to be trained to classify crowd scenes based on predefined density classes (sparse, dense etc.). Different from these approaches, the top-down feedback mechanism of (Babu Sam and Babu 2018) iteratively improves initial density prediction of a CNN regressor. Recent work of Liu et al. (2018) leverage unlabeled data for training in a multitask framework. This method is fully supervised with an additional task of count ranking on unlabeled images. However, a similar VGG based model utilized by (Li, Zhang, and Chen 2018) achieves better performance without additional unlabeled data.

The importance of unsupervised learning has been realized long back, resulting in numerous works. While the traditional clustering based methods try to infer groups in the data, modern approaches effort to learn good features by training with reconstruction objective. An autoencoder (Hinton and Salakhutdinov 2006) consists of an encoder and a decoder. The encoder generates a latent representation for the input, which is constrained by the decoder to have enough information to reconstruct the input back. In order to avoid overfitting, several variations are proposed. Vincent et al. (2008) employ denoising autoencoders that force the network to learn random noise removal. Variational autoencoders by (Kingma and Welling 2013) model input distribution in a variational Bayesian approach. Restricted Boltzmann machines (RBMs) (Smolensky 1986) and deep Boltzmann machines (DBMs) (Salakhutdinov and Hinton 2009) are other generative models for the same. Convolutional neural network based approaches like Pixel-RNN (Oord, Kalchbrenner, and Kavukcuoglu 2016) and Pixel-CNN (Oord et al. 2016) learn image density models and can generate diverse scenes. Furthermore, generative adversarial training techniques are used for density modeling in (Donahue, Krähenbühl, and Darrell 2017) and (Dumoulin et al. 2017). More recent paradigm is that of self-supervision, where instead of reconstructing the input image, some label that can be computed from the input is used for supervision. In colorization works like (Zhang, Isola, and Efros 2016; Larsson, Maire, and Shakhnarovich 2016; 2017), the network is trained to output colored image from its gray-scale version, thereby hopefully learning representations useful for other tasks. Self-supervisory labels are computed from motion cues in (Agrawal, Carreira, and Malik 2015; Jayaraman and Grauman 2015; Pathak et al. 2017). Other works obtain self-supervision labels from videos (Wang and Gupta 2015; Misra, Zitnick, and Hebert 2016), inpainting (Pathak et al. 2016), co-occurrence (Isola et al. 2016), context (Noroozi and Favaro 2016; Doersch, Gupta, and Efros 2015), etc. Zhang et al. (2017) argue that cross-channel prediction of raw data itself outperforms other task based self-supervision. Recent work of (Jenni and Favaro 2018) formulate the task of spotting artifacts in images for learning

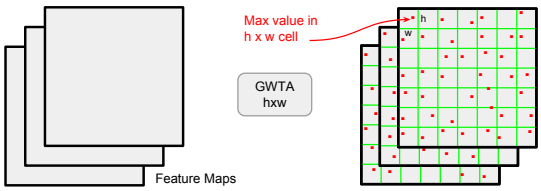


Figure 2: Grid Winner-Take-All architecture proposed in this work. Only the maximally activated neuron in a cell is allowed to pass its activation, creating sparse updates during backpropagation.

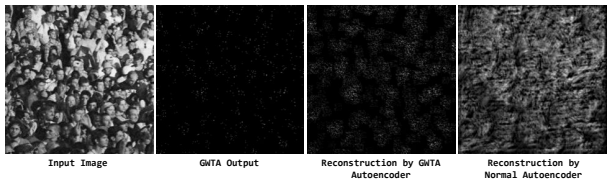


Figure 3: GWTA output of Conv1 layer for a sample image. Note that the reconstruction by GWTA autoencoder is very sparse compared to normal autoencoder.

useful features. One limitation of these self-supervised approaches is the need for defining certain pseudo label objectives compatible with the end task. If the objectives does not align, the final performance might suffer as we find in the case of density estimation. Hence in this work we prefer an unsupervised method for crowd counting. Especially, we leverage on winner-take-all (Makhzani and Frey 2015) paradigm, which we develop further to suite large-scale training with diverse crowd scenes.

## Our Approach

### Grid Winner-Take-All Autoencoders for Unsupervised Learning

Most of the unsupervised learning models are based on a reconstruction loss. Any normal autoencoder (Hinton and Salakhutdinov 2006; Vincent et al. 2008) learns features from unlabeled data in an attempt to reconstruct the input through a representational bottleneck. However, the representation acquired by the encoder is constrained to only have enough information for the decoder to reconstruct the input. This results in many cases, especially with convolutional neural networks, the encoder to learn delta or identity filters. These pass-through filters are degenerate and simply passes the input as such without applying any significant transformation (Makhzani and Frey 2015). Though these near identity filters causes trivial reduction in reconstruction objective, they are almost useless for any other tasks. It is hence apparent that normal reconstruction objective might not result in useful feature learning. One way to mitigate this effect is by increasing the task difficulty from input reconstruction to predict pseudo labels that can be easily obtained from the input (Larsson, Maire, and Shakhnarovich 2017; Agrawal, Carreira, and Malik 2015; Wang and Gupta 2015; Pathak et al. 2016; Doersch, Gupta, and Efros 2015; Noroozi and Favaro 2016). Another possible way is to constrain the encoder filters directly with some regularizers. In this paper, we follow the approach pioneered by (Makhzani and Frey 2015), where the encoder filters are constrained to fire only at the maximally activated locations. We make following crucial changes to WTA to create GWTA autoencoder:

- The WTA method is adapted for large scale training with highly diverse data. Instead of applying WTA sparsity over the entire spatial map, we apply only over fixed neighborhood. This helps in more efficient training and avoid extreme sparsity which is better for highly diverse crowd data.
- More model constraining. While Makhzani et al. (2015) use separate decoders with large filters, we show that for our task of interest, a tied decoder gives improved results.

Figure 2 illustrates our proposed GWTA architecture. GWTA is applied during the unsupervised training phase on the activation maps of the convolutional encoder. GWTA sparsity is applied independently over each channel. Any given feature map is divided into a grid of rectangular cells of pre-defined size $h \times w$. During forward propagation of the input, only the "winner" neuron in the $h \times w$ cell is allowed to pass the activation. The "winner" neuron is the one having the maximum value of activation in the cell and activations of all other neurons in the $h \times w$ cell are set to zero. Now the task of the decoder is to reconstruct the encoder input from such a sparse activation map, which is extremely hard. Hence, the encoder cannot simply learn near identity filters and get minimum reconstruction cost, but are forced to acquire useful features recurring frequently in the input data. Figure 3 shows an exemplar GWTA output and corresponding reconstruction. In GWTA, the weight update comes from few "winner" neurons in the entire feature map rather than receiving contribution from all the neurons in a normal autoencoder. This prevents filters from trying to reconstruct all parts of the input equally as in a normal autoencoder, but are forced to get specialized for certain patterns, resulting in more useful feature learning. Note that GWTA sparsity is applied only while training and is removed during testing. Since features learned are mostly non-trivial and not near identity, the encoder outputs carries significant abstractions.

The architecture for GWTA is motivated from unique characteristics of highly diverse crowd images. There exists severe variation in appearance of people even within a crowd image due to perspective changes, density gradients or occlusions. Hence the feature sets needed for faithful crowd density estimation mostly rely on local crowd patterns. Since GWTA is done in a grid fashion, we are allowing local winners to update themselves and better learn specific crowd patterns. Normal autoencoders or approaches like (Makhzani and Frey 2015) do not explicitly take into account this spatial locality, but learn features globally to reconstruct the entire input, which might not be very useful for density regression. At present, we do not have any theoretical measure of feature usefulness for density estimation, other than computing final regression performance.

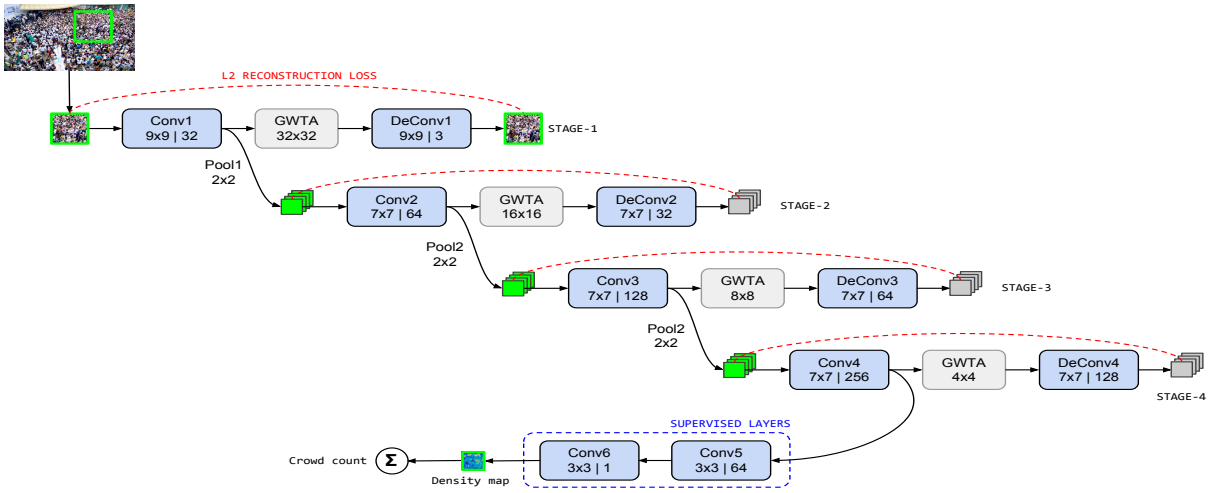


Figure 4: Architecture of GWTA based Crowd Counting CNN (GWTA-CCNN). Unsupervised training is done in stages, updating every layer by reconstructing its own input regularized by the GWTA sparsity. Last two layers are trained with supervision.

## Architecture of GWTA Counting CNN

To demonstrate the merit of the proposed architecture, we use a simple crowd counting CNN and train almost all parameters with unlabeled data followed by supervised training of the remaining parameters. We use a modified version of the CNN regressor introduced by (Zhang et al. 2016). The network consists of six convolutional layers with three pooling layers in-between. The first four layers accounting around 99.9% of the total parameters, are trained in an unsupervised manner and are then frozen. The remaining layers are trained with labeled data to regress crowd density map.

The unsupervised training is performed in stages, stacking a hierarchy of GWTA autoencoders as elucidated in Figure 4. For the first stage, random patches of size $224 \times 224$ are extracted from crowd images and are fed to the first GWTA autoencoder. This autoencoder has the convolutional layer Conv1 as enocoder followed by the GWTA regularizer layer. The GWTA cell size is chosen to be $32 \times 32$ and is subsequently halved after every pooling layer so that grid dimensions remain same across layers. The decoder DeConv1 is a transposed convolution with its weight tied with that of Conv1. Note that we do not have bias for the encoder and decoder, which we find to be empirically better. The parameters of Conv1 are updated by backpropagating the $l_2$ loss between the input and the DeConv1 output. In general, if $F^l_{X_i}(x;\Theta)$ denotes the output of layer $l$ for input $X_i$ and $\tilde{F}^l_{X_i}(x;\Theta)$ be the corresponding GWTA decoder reconstruction, then the loss function is given by,

$$L_{l_2}(\Theta) = \frac{1}{2N}\sum_{i=1}^{N} \|F^{l-1}_{X_i}(x;\Theta) - \tilde{F}^l_{X_i}(x;\Theta)\|_2^2, \qquad (1)$$

where $N$ is the number of training samples and $\Theta$ refers the learnable parameters. Parameters $\Theta$ are obtained by optimizing $L_{l_2}$ with stochastic gradient descent (SGD). The reconstruction loss tries to maximize the similarity between the reconstruction and the input, but is severely limited by the GWTA sparsity. This prevents the filters being learned from reaching near pass-through. The training is continued till loss $L_{l_2}$ on the validation set stops improving.

After the first stage encoder-decoder is trained, the Conv1 weights are frozen and the Conv1 output (without GWTA) after pooling is fed to the next stage encoder. The Conv1 activations are scaled for training stability to be in 0-1 range by dividing by the maximum response in every feature map. The maximum values are computed from the train set and are fixed for subsequent stages of training as well as for testing. Conv2 along with the corresponding deconvolution DeConv2 forms another GWTA autoencoder and is trained with the objective to reconstruct Conv1 output. This stage-wise training of GWTA autoencoders is continued till Conv4, each one learning useful representation for the output of previous layer. In this way, 99.9% of the parameters are trained without supervision and the feature representation of Conv4 is mapped to density map with supervision.

The supervised stage is required since the unsupervised training can result in some features not so useful for the end task of crowd counting. So, some level of supervision is needed to select appropriate features for density map estimation. There are many methods in the literature on how to generate density maps from head annotation available with the datasets. Most common method is to blur the head annotation with a Gaussian of fixed variance summing to one. In this work, we use a sigma of 8.0 for generating ground truth density maps. The supervised training is performed on the last two layers with simple $3 \times 3$ filters accounting for less than 0.1% of the total parameters (see Figure 4). These layers are trained to regress the density map by backpropagating $l_2$ loss between between the predicted and ground truth map. Here the $l_2$ loss function is defined as

$$L^D_{l_2}(\Theta_S) = \frac{1}{2N}\sum_{i=1}^{N} \|D_{X_i}(x;\Theta_S) - D^{GT}_{X_i}(x)\|_2^2, \qquad (2)$$

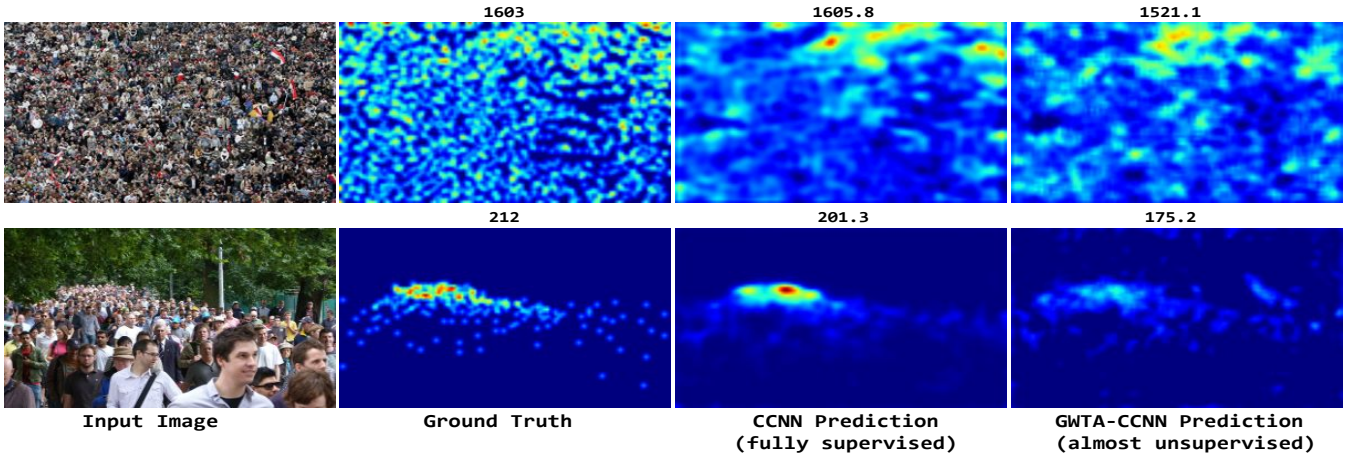


Figure 5: Sample predictions given by GWTA-CCNN on images from Shanghaitech dataset. The predicted density maps closely resemble that of the supervised CCNN model, emphasizing the ability of our unsupervised approach to learn useful features.

where $D_{X_i}(x;\Theta_S)$ stands for the output of the supervised layers with parameters $\Theta_S$ and $D^{GT}_{X_i}(x)$ is corresponding ground truth density map for the input image $X_i$. SGD is continued till the validation accuracy plateaus or does not improve. Note that none of the parameters in Conv1 to Conv4 are updated in the supervised stage.

For a given test image, overlapping patches (10% overlap) are obtained and evaluated on the trained model. The density map predictions of the overlapping areas are averaged to obtain the final density map. The crowd count is calculated by summing the density map.

## Experiments

### Evaluation Scheme

We evaluate the model performance on standard crowd counting datasets. Primarily two metrics are used by all supervised works on crowd counting. Count estimation accuracy of the model is inferred from the Mean Absolute Error or MAE metric. It is expressed as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|C_{X_i} - C^{GT}_{X_i}|, \tag{3}$$

where the count predicted by the model for image $X_i$ is $C_{X_i}$ while its actual count being $C^{GT}_{X_i}$. MAE is evaluated over the test set with $N$ images. Mean squared error or MSE is the second metric for model comparison. MSE is defined as,

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{X_i} - C^{GT}_{X_i})^2}, \tag{4}$$

a measure of variance of count estimation, indicating robustness of prediction.

### Shanghaitech dataset

The Shanghaitech dataset introduced by (Zhang et al. 2016) is the largest crowd counting dataset. Part_A set of the dataset has 300 training images and 182 images for testing. The images are collected from the Internet and the density of the crowds ranges from 33 to 3139.

We compare performance of GWTA-CCNN with that of other methods in Table 1. First important experiment is the random baseline where the unsupervised layers are not trained but randomly initialized. Subsequent supervised training is done on the feature representation obtained from this randomly initialized network. As expected, our GWTA based network achieves significantly higher count accuracy than the randomly initialized network. This suggests that the unsupervised training has resulted in learning of features useful for density estimation. Then we try end-to-end convolutional autoencoders (Hinton and Salakhutdinov 2006), where the CCNN is trained to predict the input image. This is followed by supervised training of last two layers to map features learned by Conv4 (in Figure 2) to crowd density. Denoising autoencoder (Vincent et al. 2008) is also evaluated where the objective is to reconstruct clean image from noisy input. Clearly, the proposed GWTA-CCNN achieves better MAE and MSE than these end-to-end autoencoders.

Another important baseline is with fully supervised training of the CCNN. The network is same as that in Figure 4 (Conv1 to Conv6). Obviously, the MAE for fully supervised

| Method | MAE | MSE |
|---|---|---|
| CCNN Supervised | 124.6 | 186.9 |
| CCNN Random | 367.6 | 510.1 |
| Autoencoder | 162.1 | 233.3 |
| Denoising Autoencoder | 181.9 | 254.1 |
| CCNN without WTA | 193.0 | 280.9 |
| GWTA-CCNN without tied decoder | 195.6 | 277.0 |
| GWTA-CCNN | **154.7** | **229.4** |

Table 1: Performance of GWTA-CCNN on Part_A of Shanghaitech dataset.

| Method | MAE | MSE |
|---|---|---|
| CCNN Supervised | 367.2 | 551.3 |
| CCNN Random | 903.2 | 1166.2 |
| Autoencoder | 1272.8 | 1562.3 |
| Denoising Autoencoder | 1080.9 | 1391.1 |
| CCNN without WTA | 448.3 | 633.7 |
| GWTA-CCNN without tied decoder | 500.3 | 697.8 |
| GWTA-CCNN | **433.7** | **583.3** |

Table 2: Comparison of GWTA-CCNN with other methods on UCF_CC_50 dataset (Idrees et al. 2013). Our model delivers superior performance than other unsupervised methods.

CCNN is lower than that of GWTA training, but is reasonably close, the difference in MAE being just 30.1. Further, we ablate our model by training without GWTA. The results evidence the significant improvement in performance contributed by the GWTA regularizer. Similarly, GWTA autoencoder with an untied decoder having larger filters as in (Makhzani and Frey 2015) performs worse, justifying our design choice.

Figure 5 presents density maps regressed by GWTA-CCNN and supervised network along with the corresponding ground truths. It is interesting to note that the density maps by GWTA-CCNN closely resemble the predictions by the supervised model. This emphasizes the ability of our approach to learn better features for crowd density estimation.

### UCF_CC_50 dataset

UCF_CC_50 dataset (Idrees et al. 2013) is one of the earliest and the smallest dataset for dense crowd counting with just 50 images. The dataset still remains very challenging because of the small size and the extreme variability of crowd density across the images. In fact, the crowd counts vary from 94 to 4543 amongst image. Since there is no train-test split made available, 5-fold cross-validation is adopted to evaluate counting models on the dataset (Idrees et al. 2013).

Again we see similar trend on UCF_CC_50 as with Shanghaitech dataset. In Table 2, GWTA-CCNN has better accuracy than other unsupervised baselines and is also close to the supervised baseline. We see that the end-to-end autoencoder methods have completely failed to learn useful feature for density regression. This may be possibly due to less training data (just 40 images) available from the dataset. But note that, despite having very less training images, our GWTA based model has significantly better results.

## Analysis and Ablations

### Supervised Vs Unsupervised Features

It is important to compare features obtained through unsupervised learning to that of its supervised counterpart. This would give valuable insights on how GWTA model works as well as help future researches to bridge the performance gap between the two training paradigms. Figures 6 and 7 display features maps from supervised CCNN model, autoencoder with and without GWTA. Only some of the feature maps are

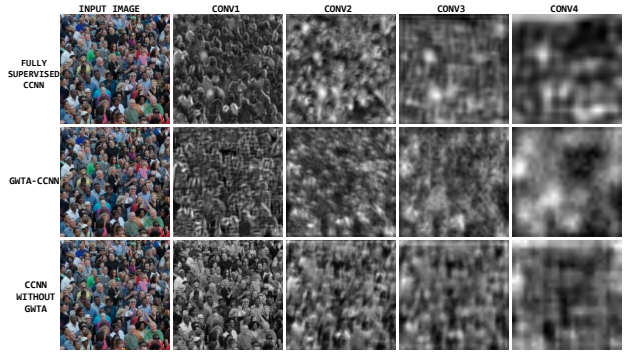


Figure 6: Qualitative comparison of features learned by GWTA autoencoder with that of the fully supervised CCNN. The images are sum maps of all the features in a layer.

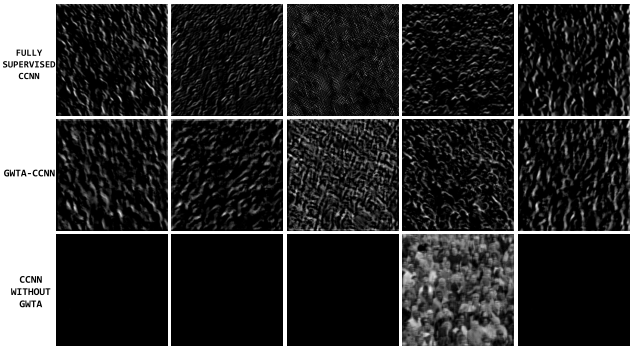


Figure 7: Some of the individual feature maps of Conv1 for GWTA and supervised CCNN.

shown due to space constraints, but the sum maps in Figure 6, which are sum of all the feature maps gives a general idea about all the feature maps in a particular layer. It is clear from Conv1 maps (Figure 7) that supervised and GWTA unsupervised are close in terms of the features learned, subject to different value ranges. Moreover, the sum maps of the features are also close indicating that most of the filters are similar as that of supervised. Note that the feature maps of autoencoder without GWTA are significantly different from the maps of supervised and are mostly passing the input with minimal transformation or are dead filters (blank output). Similarly for Conv2, the GWTA features look close to supervised than normal autoencoder and are more related to abstracting various types of edges to form compound patterns like shoulders, head etc. It starts to diverge at Conv3, where the supervised features show more aggregation to become like density maps. But the GWTA unsupervised feature maps, though not visually very different, still combines previous layer features to form more abstractions. This is due to the absence of any task oriented supervisory signal. Coming to Conv4, the supervised layer activations almost look like density maps. In contrast, the Conv4 unsupervised features look very different and still creates many abstractions which may or may not be useful for the task of crowd

| Method | MAE | MSE |
|---|---|---|
| Colorization | 168.4 | 244.5 |
| Inpainting | 166.3 | 252.8 |
| Count Consistency | 188.8 | 282.3 |
| GWTA-CCNN | **154.7** | **229.4** |

Table 3: Performance of GWTA-CCNN on Part_A of Shanghaitech dataset (Zhang et al. 2016) compared with self-supervised methods.

counting. This observation that initial layers of neural network have general features, with deeper layers tuned for task specific features is in line with existing findings in the literature. Also note that, many feature maps of the autoencoder without GWTA still have dense information about the input in order to reduce the reconstruction loss and hence differ significantly from that of the supervised. This proves that the GWTA stacked autoencoder learns features close to that of the supervised model compared to other competing models.

### Comparison with Self-Supervised Methods

In this section, we compare the performance of our model with some self-supervised methods, where the features are learned by training the model to predict pseudo labels that are computed from the input image. For example, in self-supervision with colorization, the CCNN model is trained as an autoencoder to regress colored image from its gray scale version. We see from Table 3 that the proposed GWTA-CCNN works better than self-supervision with colorization task. Inpainting (Pathak et al. 2016) is another task for self-supervision. A rectangular portion of the input image is removed and filled with the mean value. The surrounding context image is used to train CCNN with the task of predicting the missing region of the input. This task also does not surpass the performance of GWTA unsupervised learning. Further, to suite the end task of density regression, we employ the count ranking loss formulation of (Liu, van de Weijer, and Bagdanov 2018). We train CCNN to be consistent by enforcing the count estimation of the interior region of a crowd image to be less than the overall count of the crowd. Though the ranking loss provide count consistency, it seems to be incapable of providing enough good features. This might be because of the fact that the ranking loss could be satisfied without learning any crowd discriminative features. This indicates one drawback of self-supervised methods, the need for certain task (like colorization etc.) suitable for the end task to be defined. If the task is not compatible with the end task, performance might suffer.

### Effect of labeled data on performance

It is important to examine the dependence of count estimation quality on the amount of labeled data used for final supervision. Figure 8 shows performance of our GWTA counting model with different levels of supervision compared against fully supervised CCNN. On Part_A Shanghaitech dataset, we vary the number of labeled training images from 50% of the entire dataset to the extremity of just one image. We repeat the experiments eight times with different randomly drawn subset of the labeled data and report the average MAE. Interestingly, we see that the performance of GWTA-CNN at extreme less data case is clearly superior to fully supervised model. Some amount of training data is required for satisfactory accuracy for fully supervised case and outperforms GWTA-CCNN at around 40% data. With more data, though the accuracy of both approaches increases, the MAE for unsupervised method drastically decreases than fully supervised and saturates near the 100% data performance with less data (50%). This is primarily because the few parameters being updated with supervision require only limited data for training. Hence the suitability of our approach at extremely less labeled data scenario is well emphasized.

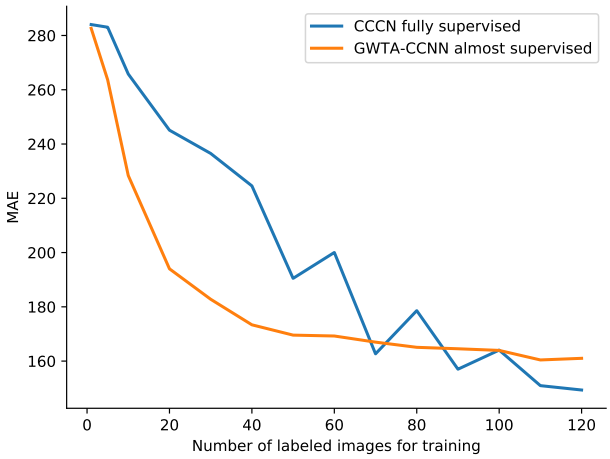


Figure 8: Amount of labeled data vs MAE. CCNN is trained in fully and almost supervised fashion with different amounts of labeled data of Part_A Shanghaitech dataset. We see that at less data scenarios our almost unsupervised approach performs better than fully supervised.

## Conclusion

Our proposed architecture attempts to train a crowd counting CNN in an almost unsupervised manner. Since it is difficult to obtain large-scale annotated data for dense crowds, this problem deserves prime attention. We develop Grid Winner-Take-All (GWTA) autoencoder to learn useful features from unlabeled images. The basic idea is to restrict weight update of neurons in convolutional output maps to the maximally activated neuron in a fixed spatial cell. Almost 99.9% of the parameters of the network are trained as stacked WTA autoencoders using unlabeled crowd images, while remaining parameters are updated with supervision. We evaluate our model on standard benchmark datasets and demonstrate better performance compared to other unsupervised methods. In fact, the count performance is reasonably close to the supervised baseline, with a performance gap of 25%. Future works should address this performance gap. Additional analysis reveals that our unsupervised approach outperforms fully supervised training when available labeled data is less.

## Acknowledgements

This work was supported by SERB, Dept. of Science and Technology, Govt. of India (Proj: SB/S3/EECE/0127/2015).

## References

Agrawal, P.; Carreira, J.; and Malik, J. 2015. Learning to see by moving. In *IEEE International Conference on Computer Vision*.

Babu Sam, D., and Babu, R. V. 2018. Top-down feedback for crowd counting convolutional neural network. In *AAAI Conference on Artificial Intelligence*.

Babu Sam, D.; Sajjan, N. N.; Babu, R. V.; and Srinivasan, M. 2018. Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Babu Sam, D.; Surya, S.; and Babu, R. V. 2017. Switching convolutional neural network for crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Boominathan, L.; Kruthiventi, S. S.; and Babu, R. V. 2016. Crowdnet: A deep convolutional network for dense crowd counting. In *ACM International Conference on Multimedia*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision*.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2017. Adversarial feature learning. In *International Conference on Learning Representations*.

Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; and Courville, A. 2017. Adversarially learned inference. In *International Conference on Learning Representations*.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

Idrees, H.; Saleemi, I.; Seibert, C.; and Shah, M. 2013. Multi-source multi-scale counting in extremely dense crowd images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Isola, P.; Zoran, D.; Krishnan, D.; and Adelson, E. H. 2016. Learning visual groups from co-occurrences in space and time. *IEEE Conference on Computer Vision and Pattern Recognition*.

Jayaraman, D., and Grauman, K. 2015. Learning image representations tied to ego-motion. In *IEEE International Conference on Computer Vision*.

Jenni, S., and Favaro, P. 2018. Self-supervised feature learning by learning to spot artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2016. Learning representations for automatic colorization. In *European Conference on Computer Vision*, 577–593. Springer.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2017. Colorization as a proxy task for visual understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Y.; Zhang, X.; and Chen, D. 2018. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, X.; van de Weijer, J.; and Bagdanov, A. D. 2018. Leveraging unlabeled data for crowd counting by learning to rank. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Makhzani, A., and Frey, B. J. 2015. Winner-take-all autoencoders. In *Advances in Neural Information Processing Systems*.

Misra, I.; Zitnick, C. L.; and Hebert, M. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, 527–544. Springer.

Noroozi, M., and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 69–84. Springer.

Onoro-Rubio, D., and López-Sastre, R. J. 2016. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*, 615–629. Springer.

Oord, A. v. d.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with PixelCNN decoders. In *Advances in Neural Information Processing Systems*.

Oord, A. v. d.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *International Conference on Machine Learning*, 1747–1756.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Pathak, D.; Girshick, R.; Dollár, P.; Darrell, T.; and Hariharan, B. 2017. Learning features by watching objects move. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Salakhutdinov, R., and Hinton, G. E. 2009. Deep boltzmann machines. In *international conference on artificial intelligence and statistics*.

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

Sindagi, V. A., and Patel, V. M. 2017b. Generating high-quality crowd density maps using contextual pyramid CNNs. In *IEEE International Conference on Computer Vision*.

Smolensky, P. 1986. Information processing in dynamical systems: foundations of harmony theory. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, 194–281. MIT Press.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *International conference on Machine learning*.

Wang, X., and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision*.

Wang, C.; Zhang, H.; Yang, L.; Liu, S.; and Cao, X. 2015. Deep people counting in extremely dense crowds. In *ACM international conference on Multimedia*, 1299–1302.

Zhang, C.; Li, H.; Wang, X.; and Yang, X. 2015. Cross-scene crowd counting via deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; and Ma, Y. 2016. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European Conference on Computer Vision*, 649–666. Springer.

Zhang, R.; Isola, P.; and Efros, A. 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*.