FusionCount: Crowd Counting via Multiscale Feature Fusion

Yiming Ma†, Victor Sanchez†, Tanaya Guha§

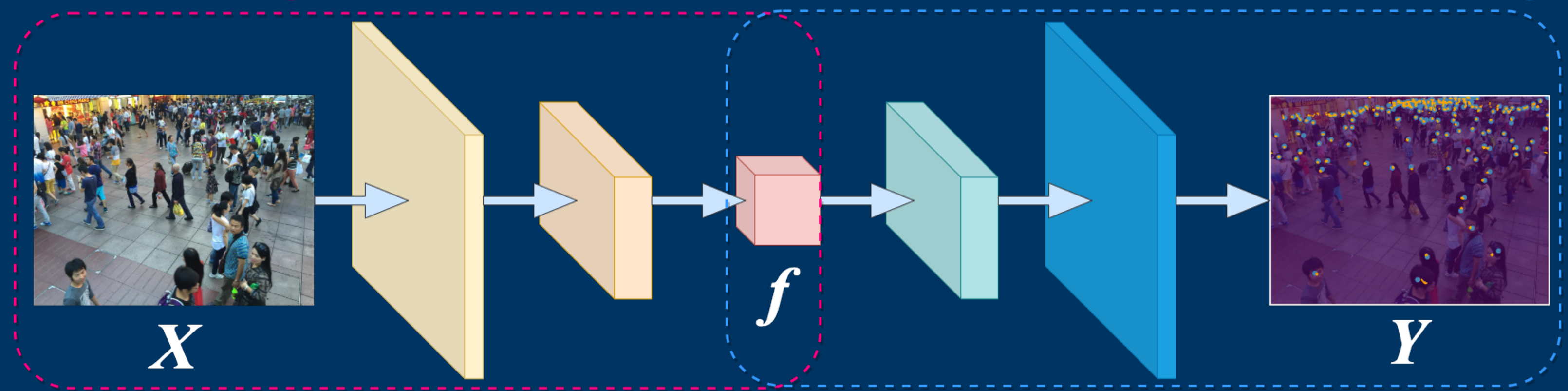
†: University of Warwick
§: University of Glasgow

Introduction

- **Crowd counting** aims to automatically estimate the number of individuals present in a scene from an image or video.

An example from ShanghaiTech A [1].

- State-of-the-art methods follow an encoder-decoder approach.



An example of encoder-decoder structures in crowd counting. Given an image X, the encoder extracts the feature map f, from which the decoder generates the predicted density map Y.

- The feature maps f should be multiscale to cover different sizes of people depicted in the image.

People of similar scales (from [1]).

People of disparate scales (from [1]).

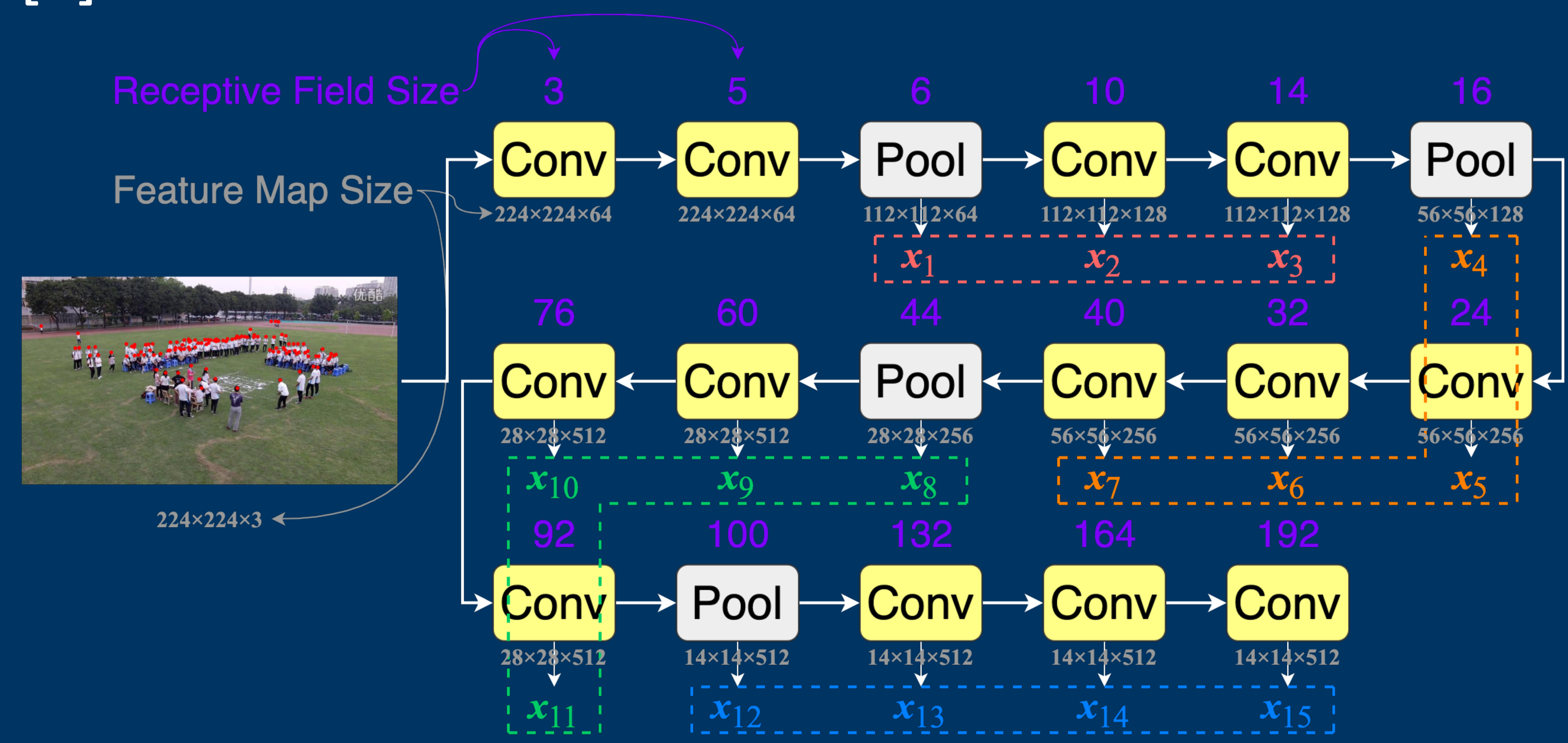
- The latest algorithms [2, 3] exploit multiscale modules after encoding to further process the embeddings f.
  - In these approaches, filters of different sizes are leveraged, and the outputs are fused adaptively:
  $$f' = W_1 f_1 + \cdots + W_n f_n.$$
  - Using these modules to introduce multiscale information can lead to extra computation.

Our Model: FusionCount

Features extracted by different encoding layers already have different receptive field sizes.

Encoding

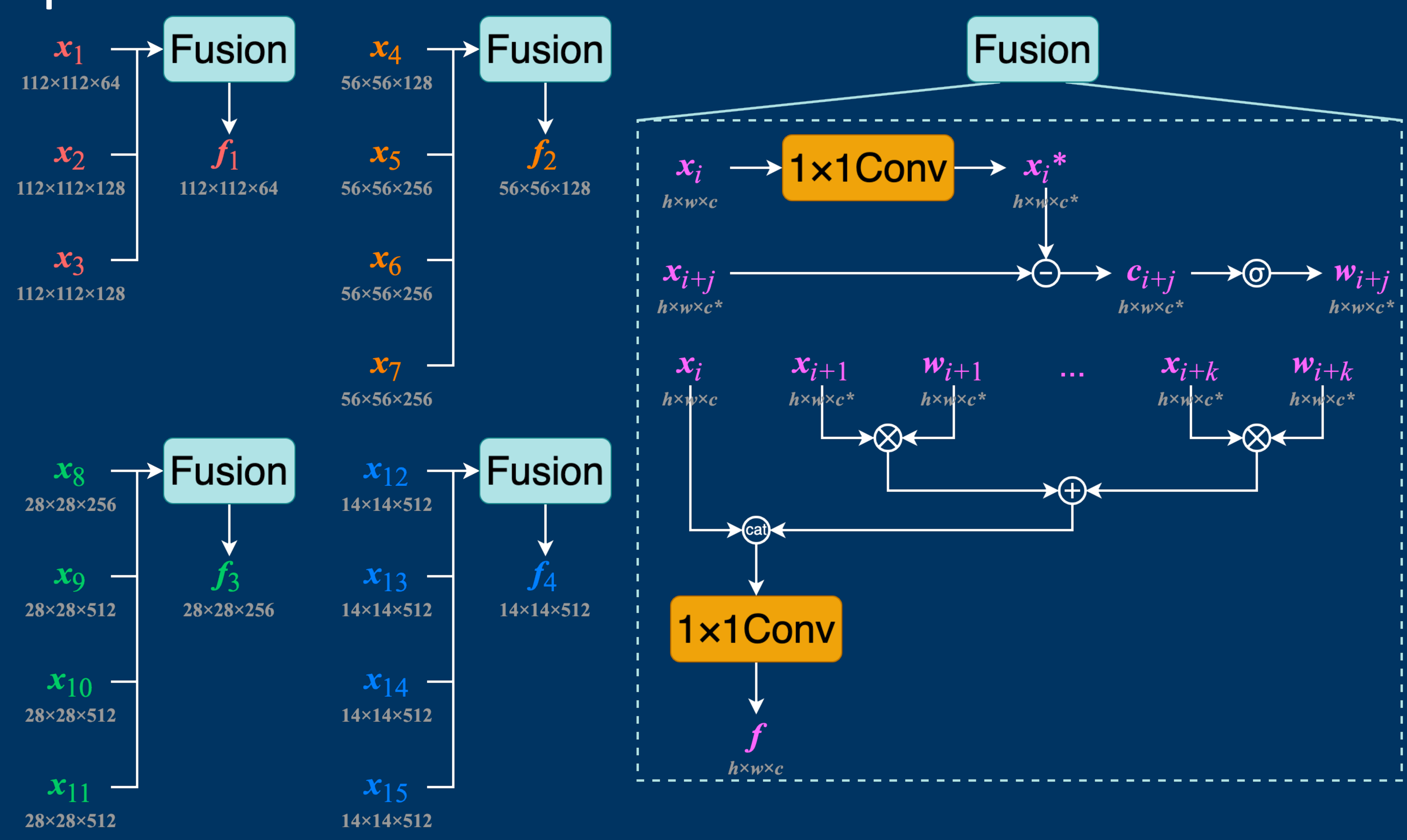
Following previous work [2, 3], we leverage VGG-16 [4] as the encoder.



The encoder of our proposed model FusionCount: Only the first 17 layers of the original VGG-16 are leveraged, and feature maps are collected starting from the third layer. Numbers in purple are features' receptive field sizes and those in grey ($h \times w \times c$) indicate their sizes, assuming the input image has the size of $224 \times 224 \times 3$. Features with the same spatial resolution are grouped together for the first-phase fusion.

Feature Fusion

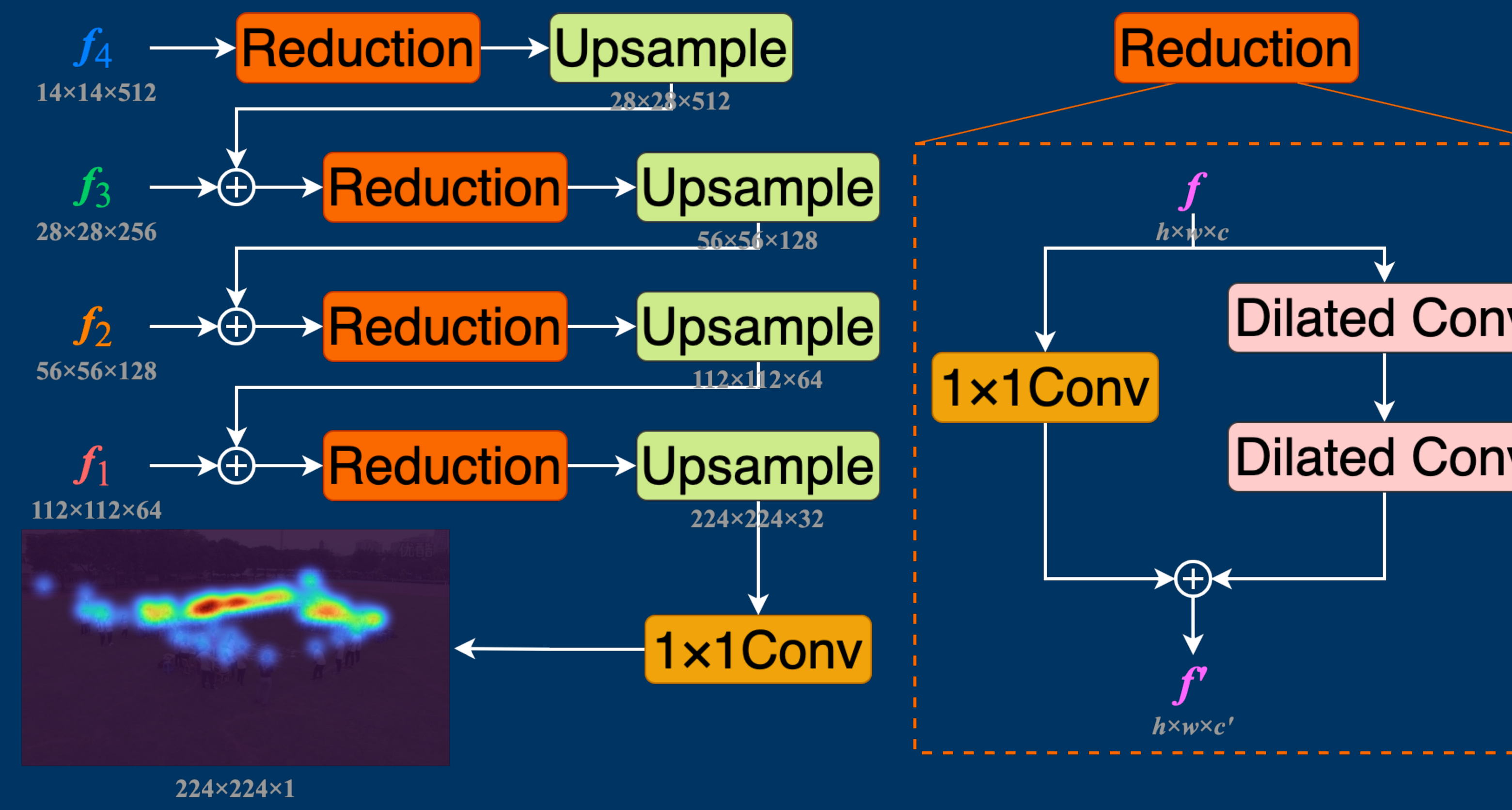
We exploit the conception of contrast features proposed in [2] to fuse features with the same spatial size.



The feature fusion modules of FusionCount: In each group, weights are computed from contrast features c_{i+j}. Then features from convolutional layers are averaged by using these weights and subsequently concatenated with the feature map from the pooling layer.

Decoding

We propose a novel channel reduction module by combining point-wise convolution with dilated convolution.

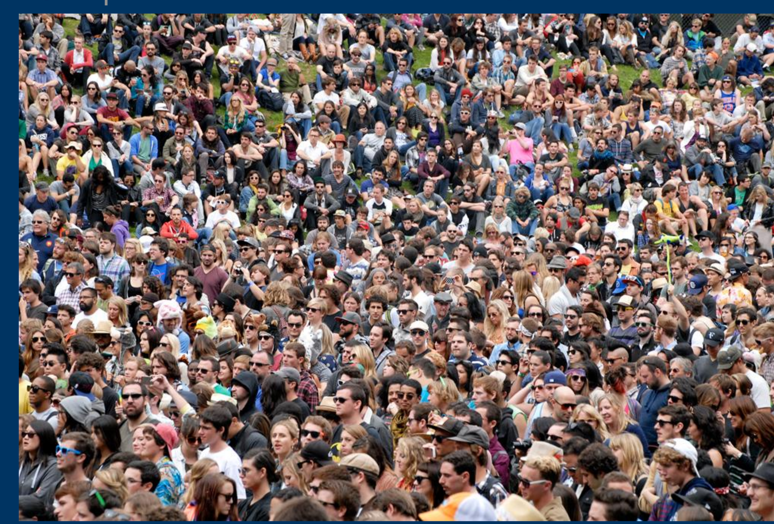


The decoding process of FusionCount: Starting from f_4, the proposed channel reduction module first decreas its number of channels. The result is then upsampled and fused with another first-phase multiscale feature f_3

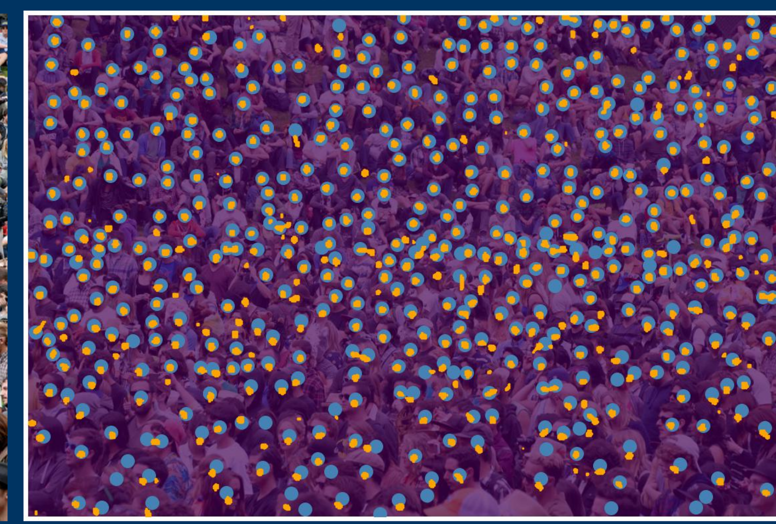
Experiments

Model	Mult-Adds	ShanghaiTech A		ShanghaiTech B	
		MAE	RMSE	MAE	RMSE
CSRNet [5]	859.99G	68.2	115.0	10.6	16.0
CAN [2]	908.05G	62.3	<u>100.0</u>	7.8	<u>12.2</u>
BL [6]	<u>853.70G</u>	62.8	101.8	7.7	12.7
DM-Count [7]	<u>853.70G</u>	**59.7**	**95.7**	<u>7.4</u>	**11.8**
FusionCount (ours)	**815.00G**	<u>62.2</u>	101.2	**6.9**	**11.8**

Comparison of our model FusionCount with state-of-the-art models of similar sizes.



Ground Truth: 521



Prediction: 525.97; Relative Error: 0.95%

Ground Truth: 176

Prediction: 176.08; Relative Error: 0.05%

References:

[1] Zhang, Yingying, et al. "Single-image crowd counting via multi-column convolutional neural network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[2] Liu, Weizhe, Mathieu Salzmann, and Pascal Fua. "Context-aware crowd counting." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[3] Thanasutives, Pongpisit, et al. "Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting." *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.
[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *International Conference on Learning Representations, 2015.*
[5] Li, Yuhong, Xiaofan Zhang, and Deming Chen. "CSRnet: Dilated convolutional neural networks for understanding the highly congested scenes." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[6] Ma, Zhiheng, et al. "Bayesian loss for crowd count estimation with point supervision." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
[7] Wang, Boyu, et al. "Distribution matching for crowd counting." *Advances in neural information processing systems* 33 (2020): 1595-1607.