Few-Shot Scene Adaptive Crowd Counting Using Meta-Learning

Mahesh Kumar Krishna Reddy†, Mohammed Asiful Hossain‡, Mrigank Rochan†, and Yang Wang†

†Department of Computer Science, University of Manitoba
‡ Huawei Technologies Co. Ltd.

Overview

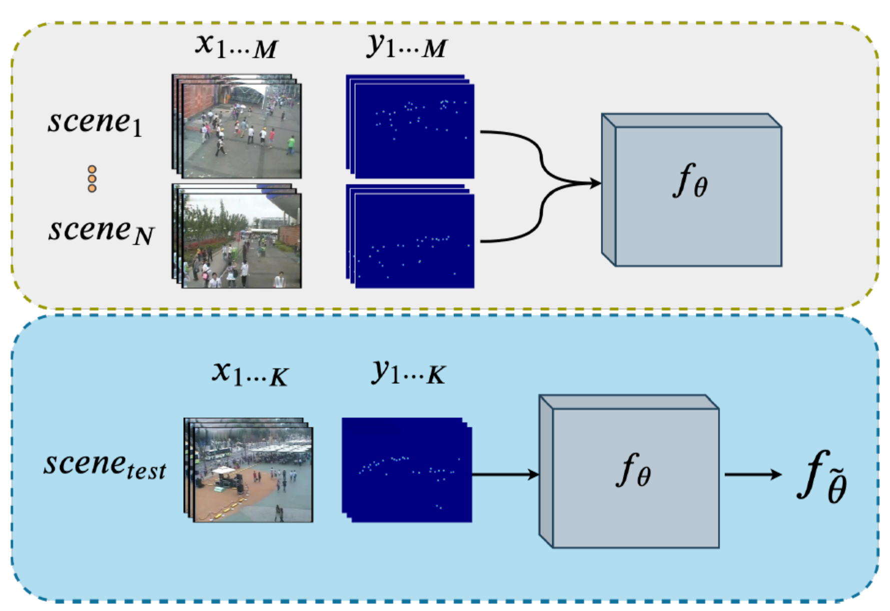
Traditional Models: They require large number of labeled data to achieve a successful model. However, for application like crowd counting, collecting large amount of labeled data or annotating every camera images is expensive, or cumbersome.

Meta-learning: It enables to exploit the adaptable scene representation to learn a new camera scene (task) with limited data.

Problem Setup

Top row: During training, we have access to a set of N different camera scenes where each scene comes with M labeled examples. From such training data, we learn the model parameters θ of a mapping function f_θ such that θ is generalizable across scenes in estimating the crowd count.

Bottom row: Given a test (or target) scene, we assume that we have a small number of K labeled images from this scene, where $K \ll M$ (e.g., $K \in \{1,5\}$) to learn the scene-specific parameters $\tilde{\theta}$. With the help of meta-learning guided approach we quickly adapt f_θ to test scene specific parameters $f_{\tilde{\theta}}$ that predicts more accurate crowd count than other alternative solutions.



Few-shot Scene Adaptive Crowd Counting

Inner update:

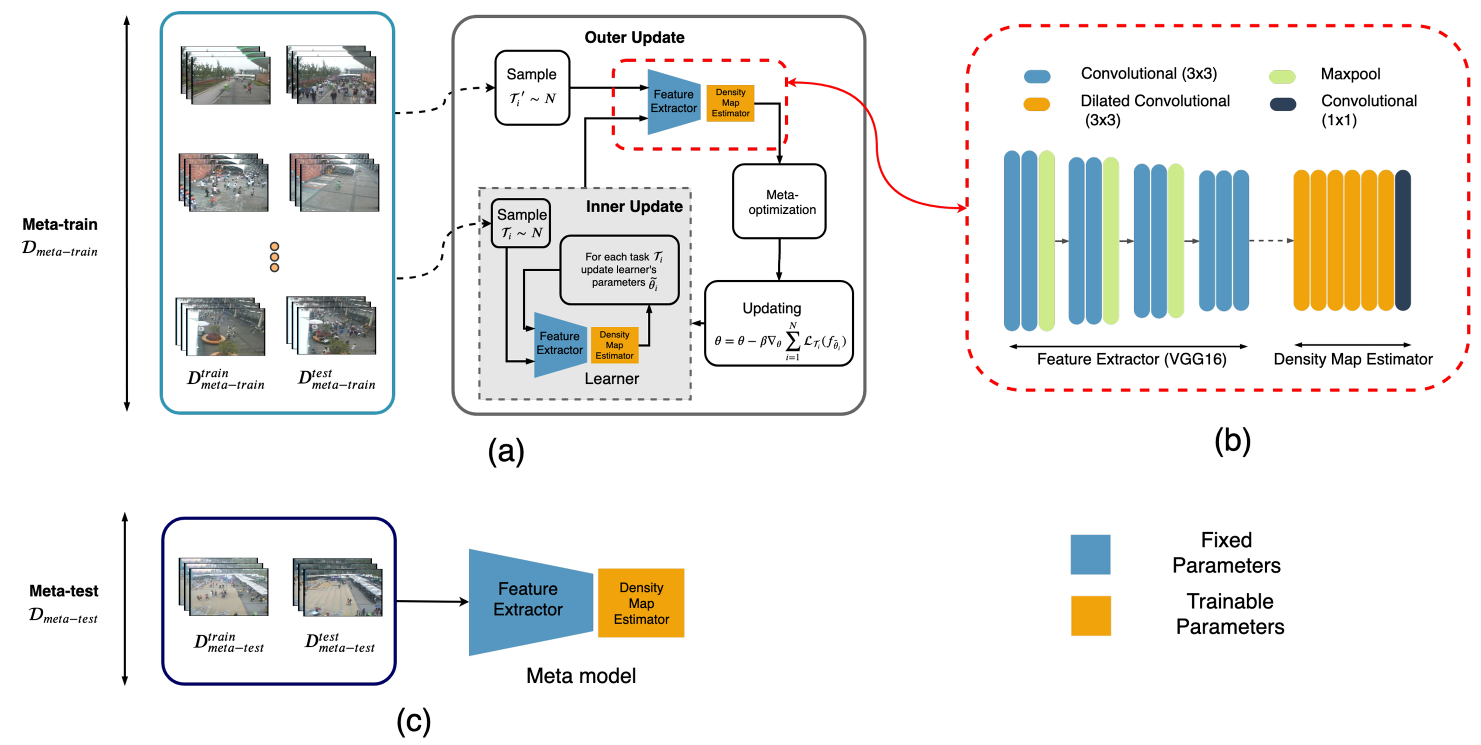
$$\tilde{\theta}_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$$

$$\text{where } \mathcal{L}_{\mathcal{T}_i}(f_\theta) = \sum_{(x^{(j)}, y^{(j)}) \in D_i^{train}} \| f_\theta(x^{(j)}) - y^{(j)} \|_F^2 \quad (1)$$

$$\mathcal{L}_{\mathcal{T}_i}(f_{\tilde{\theta}_i}) = \sum_{(x^{(j)}, y^{(j)}) \in D_i^{test}} \| f_{\tilde{\theta}_i}(x^{(j)}) - y^{(j)} \|_F^2 \quad (2)$$

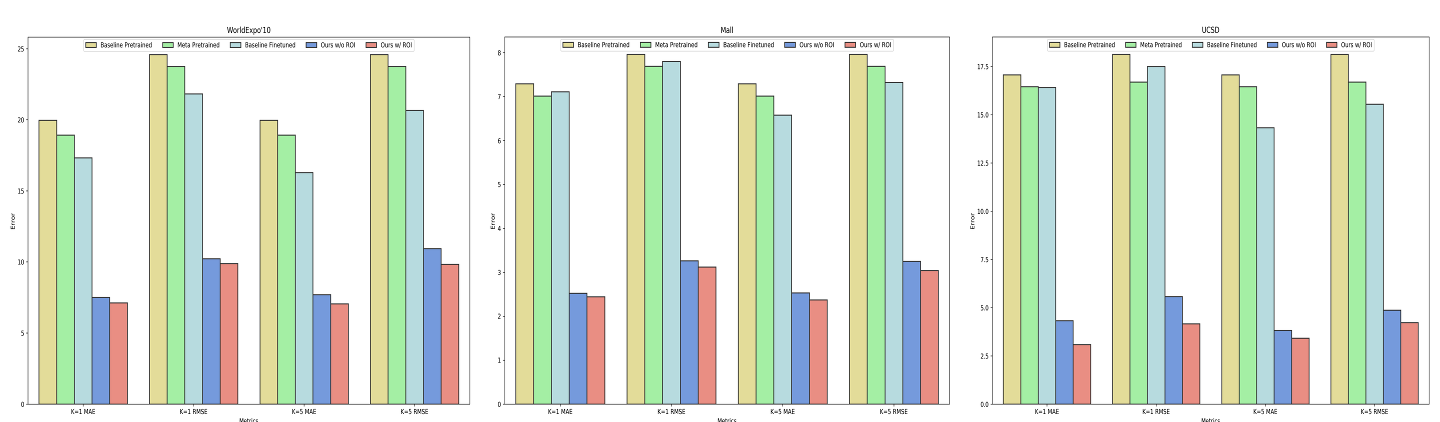
Outer update:

$$\theta = \theta - \beta \nabla_\theta \sum_{i=1}^{N} \mathcal{L}_{\mathcal{T}_i}(f_{\tilde{\theta}_i}) \quad (3)$$

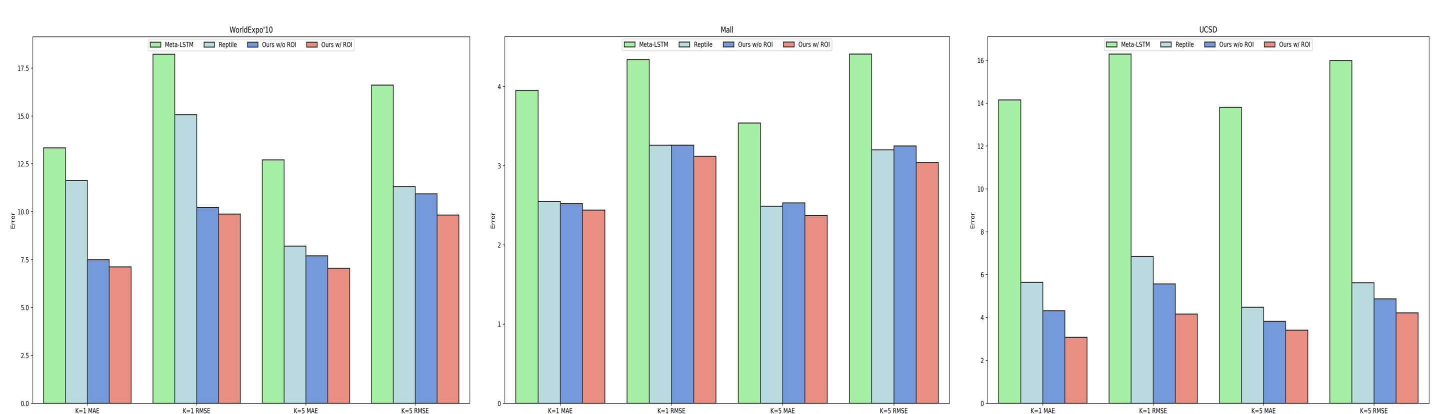


Experiments

- **Datasets**: WorldExpo'10 [Zhang et al.], UCSD [Chan et al.], and Mall [Loy et al.]
- **Metrics**: Mean Absolute Error, and Root Mean Squared Error
- **Crowd baselines**: Pre-trained, Fine-tuned, and Meta pre-trained
- **Meta baselines**: Meta-LSTM, and Reptile



Crowd counting results on WorldExpo'10, Mall, and UCSD datasets

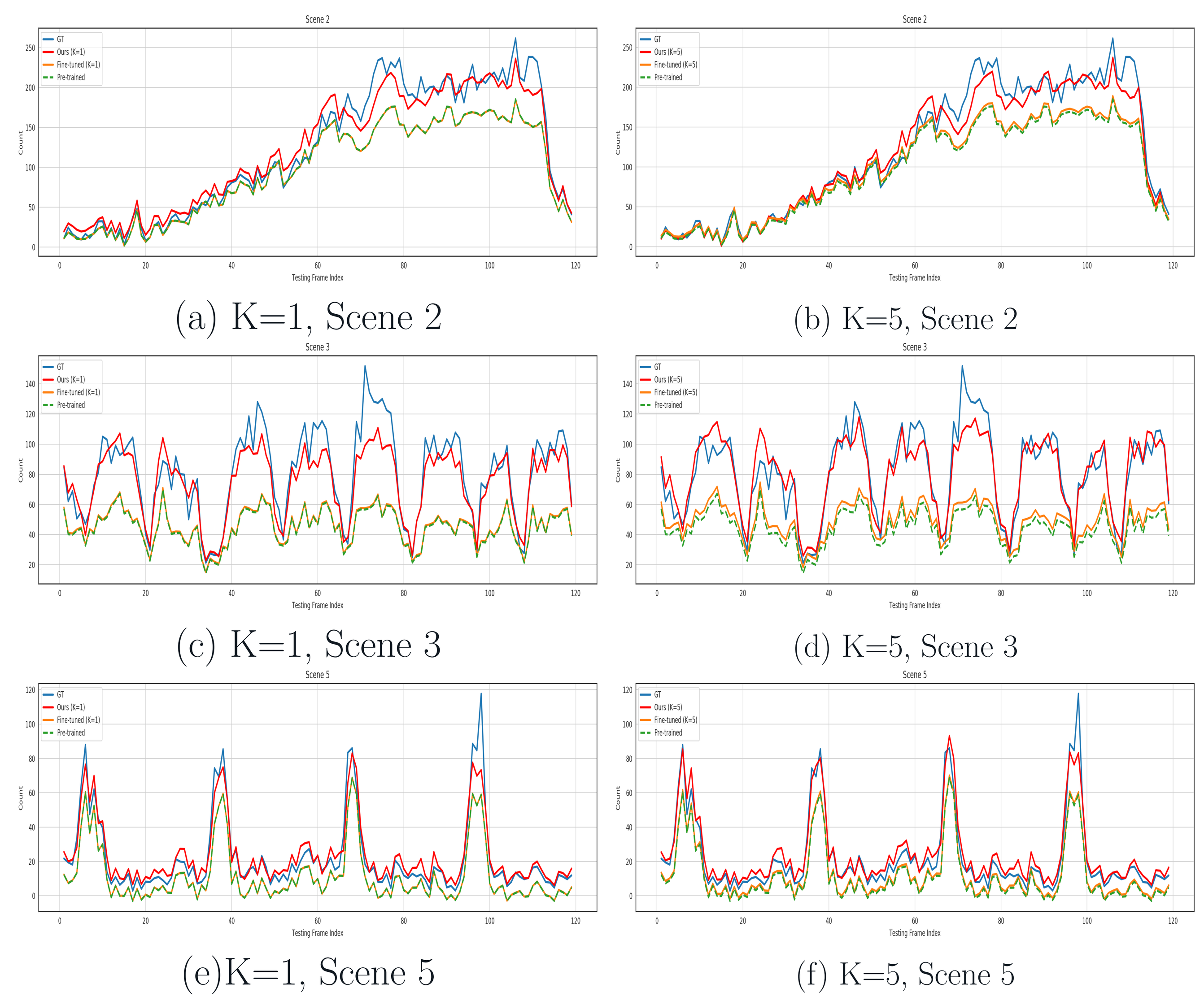


Meta-learning comparison results between different optimization based approaches [2, 3]

Methods	1-shot (K=1)	
	MAE	RMSE
Hossain et al. [1]	8.23	12.08
Ours w/o ROI	7.5	10.22
Ours w/ ROI	**7.12**	**9.88**

Comparison of results on the WorldExpo'10 dataset with $K = 1$ images in the target scene with Hossain *et al.* [1].

Analysis



(a) K=1, Scene 2 (b) K=5, Scene 2

(c) K=1, Scene 3 (d) K=5, Scene 3

(e)K=1, Scene 5 (f) K=5, Scene 5

Fig. 5: Crowd counting performance comparison between the baselines and our approaches in different scene-specific images from WorldExpo'10 dataset

References

[1] Mohammad Asiful Hossain et al. "One-Shot Scene-Specific Crowd Counting". In: *British Machine Vision Conference.* 2019.

[2] Alex Nichol, Joshua Achiam, and John Schulman. "On first-order meta-learning algorithms". In: *arXiv preprint arXiv:1803.02999* (2018).

[3] Sachin Ravi and Hugo Larochelle. "Optimization as a model for few-shot learning". In: *International Conference on Learning Representations.* 2017.