TASK-GUIDED BIASED DIFFUSION MODELS FOR POINT LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We hypothesize that diffusion models can be used to enhance the performance of deep learning methods for predictive tasks involving sparse outputs, such as point-localization tasks. However, this approach presents two challenges: slow inference and the stochastic nature of sampling, resulting in varying predictions based on different initialization seeds. To improve inference efficiency, we propose the introduction of task bias in the forward diffusion process, replacing the standard convergence to zero-mean Gaussian noise by convergence to a noise distribution closer to that of the target sparse point localization data. This simplifies the reverse diffusion process and is shown to decrease the number of necessary denoising steps, while improving prediction quality. To decrease prediction variance due to seed stochasticity, we propose a task-guided loss that is shown to decrease the average distance between predictions from different noise realizations. The two contributions are combined into the Task-Guided Biased Diffusion Model (TGBDM), which maps an initial prediction from a classical localization method into a refined localization map. This is shown to achieve state-of-the-art performance for crowd localization, pose estimation, and cell localization.

1 INTRODUCTION

Computer vision tasks such as crowd localization and human pose estimation require the prediction of sparse localization maps (Wan et al., 2021). However, due to the overlap between receptive fields of nearby visual features, existing deep learning solutions tend to predict smooth maps. For crowd localization, these make it hard to recover individual heads in crowded regions. For human pose estimation, a smooth heat map makes the boundary between joints ambiguous.

Diffusion models have shown remarkable performance for the synthesis of images (Ho et al., 2020) and videos (Ho et al., 2022) from random noise (Song et al., 2020), natural language (Rombach et al., 2022) or sketches (Wang et al., 2022). They have also been shown to benefit discriminative tasks including object detection (Chen et al., 2022), segmentation (Baranchuk et al., 2021), and classification (Han et al., 2022). In this work, we hypothesize that they can be used to increase the sharpness of the predictions produced by classical approaches to localization tasks. This can greatly benefit problems like crowd localization, human pose estimation, or cell localization.

However, there are two challenges to the application of diffusion models to point localization tasks. First, their inference is slow. Typical diffusion models require 1000 steps to generate high-quality images. Second, the predictions can vary substantially with the noise seed used to initialize the sampling chain at inference. This is desirable for generic image and video synthesis, where diversity is highly desirable, but not for the refinement of an initial prediction for a conditional task, such as the location of heads, body joints, or cells in a a given image.

To address these issues, we propose the Task-Guided Biased Diffusion Model (TGBDM) for crowd localization, human pose estimation, and cell localization. This has two main contributions. First, to improve inference speed, the initial conditioning prediction, produced by the classical method, is used as a bias term in the forward diffusion process. This makes the process converge to a Gaussian random variable whose mean is identical to that of the conditioning prediction, rather than zero as is usual for standard diffusion processes. In result, the distribution from which the random seed is drawn is centered around the initial prediction produced by the classical method. For example, in the crowd localization task, the density map produced by an existing crowd localization approach

is used to create this bias. As shown in Figure 3, this makes the distribution of noisy images closer to this original prediction, which makes the denoising process easier. We have found that the addition of this bias substantially reduces the number of diffusion steps required to generate accurate localization maps. Second, taking inspiration from classifier-guided diffusion models (Dhariwal & Nichol, 2021), we propose a task-guided loss that encourages the denoising network to predict similar results for different noise seeds. This is shown to significantly reduce the dependency of the predicted localization maps on the seed, leading to overall better localization performance. We show that, with these two contributions, the TGBDM is able to improve the localization performance of state of the art methods to crowd localization, human pose estimation, and cell localization.

The contribution of the paper is three-fold: 1) We show how to apply diffusion models to point localization tasks requiring sparse outputs, such as crowd localization and human pose estimation, and how this improves on the performance of classical solutions to these problems. 2) A task-guided loss function that both decreases the variance of diffusion model predictions and improves their accuracy. 3) A biased diffusion process, where the introduction of task-guided bias in the random seed of the sampling chain simplifies the denoising process for predictive tasks, decreasing the number of denoising steps and improving inference speed.

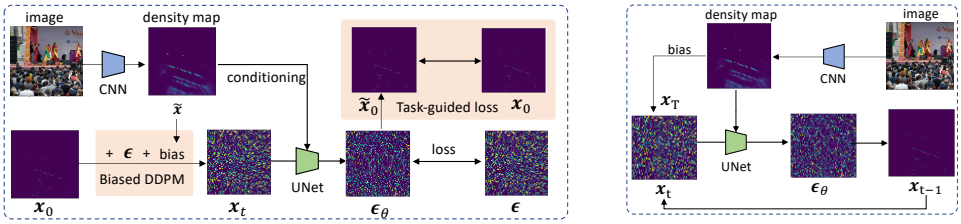
2 Related Works

2.1 Diffusion Models

Motivated by Langevin Dynamics (Welling & Teh, 2011), Denoising Diffusion Probabilistic Models (DDPM) are proposed to generate images from random Gaussian noise based on a Markov chain (Ho et al., 2020). Denoising Diffusion Implicit Models (DDIM) are then proposed with a non-Markov chain (Song et al., 2020). Nichol & Dhariwal (2021) propose classifier guidance to generate images from different classes. To avoid explicit classifier training, Ho & Salimans (2022) propose to sample from a conditional and an unconditional model during inference. Liu et al. (2023b) proposes novel conditional diffusion models that directly learn the diffusion processes between two distributions. Latent diffusion (Rombach et al., 2022) is proposed to improve the inference efficiency by applying diffusion models in the latent space. Except for generative tasks, diffusion models have also been applied to discriminative tasks, such as classification (Han et al., 2022), segmentation (Baranchuk et al., 2021), and object detection (Chen et al., 2022). However, these methods suffer from intensive computation complexity and stochastic predictions from different random noises. In this paper, we propose a biased DDPM and task-guided loss to address these challenges. DiffusionDet (Chen et al., 2022) also localizes object bounding boxes from images, but their method is based on a standard diffusion of the bounding box coordinates (a 4-dim space), whereas we modify the diffusion model so that it can be directly applied to generate the localization map (image space). Our point localization tasks also do not require scale information which is easier for annotation.

2.2 Crowd Localization

Detection-based crowd localization methods first generate pseudo bounding boxes from point annotations (Liu et al., 2019b) and use object detection algorithms to detect individuals (Ren et al., 2015). However, the performance of detection-based methods is limited since the bounding boxes are not accurate. Density map-based methods are proposed with point annotations only during training (Wan et al., 2021). Idrees et al. (2018) and Gao et al. (2019) propose to predict sharp density maps for better localization. However, the predicted maps are still blurry for high-density regions. Therefore, postprocessing is required to find the local maximum. Moreover, the postprocessing is sensitive to the hyperparameter. In recent years, point prediction-based methods have been proposed to directly predict points (Song et al., 2021). Liang et al. (2022) propose a Transformer-based architecture to predict points motivated by DETR (Carion et al., 2020). An improved initial query selection method is proposed to further improve the performance in Liu et al. (2023a). Our proposed method can effectively predict very sharp output which does not require complex postprocessing and is robust to the hyperparameter.



(a) training (b) inference

Figure 1: (a) and (b) are training and inference pipelines respectively of the proposed method.

2.3 Human Pose Estimation

Human pose estimation methods can be divided into two categories: top-down and bottom-up. Top-down methods first detect individuals in the image and then localize joints from the detected result (He et al., 2017). Stacked hourglass networks (Newell et al., 2016) are proposed to extract multi-scale features. HRNet (Sun et al., 2019) is designed for high-resolution representations. Bottom-up methods detect individuals and their joints simultaneously (Cheng et al., 2020). Openpose (Cao et al., 2021) is proposed to associate joints from the same person with part affinity fields. Besides, Associate Embedding (Newell et al., 2017) is proposed to learn a specific embedding for each individual. The proposed method is effective in further improving the performance by predicting sharper joint heatmaps.

2.4 Cell Localization

Detection-based methods are also used in cell localization (Paulauskaite-Taraseviciene et al., 2019). However, the detection-based methods tend to miss cells in high-density regions. Therefore, the counting performance is limited. Regression-based methods (Li et al., 2018) are proposed to address this issue by predicting density maps instead of bounding boxes. However, the localization performance is worse than detection-based methods since the predicted density maps are too blurry to distinguish individual cells (Ciampi et al., 2022). Our proposed method can effectively improve both cell counting and localization performance by utilizing diffusion models for sharp outputs.

3 Methodology

In this section, we briefly review diffusion models. and introduce enhancements for their application to point localization tasks. Finally, the biased DDPM and task-guided loss are proposed to improve inference speed and decrease stochastic effects of generative models. The pipeline of the proposed method is shown in Figure 1.

3.1 Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) are generative models that exploit Langevin Dynamics to generate images from a Gaussian noise seed.

3.1.1 Forward Diffusion Process

Consider an image sampled from a real data distribution $\mathrm{x}_0 \sim q(\mathrm{x})$. At step t this is mapped to noisy image x_t by addition of Gaussian noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \beta_t \mathbf{I})$ to x_{t-1}, where β_t is the variance for step $t \in [1, T]$. The noisy image produced at step t is a sample from

$$\mathrm{x}_t = \sqrt{\bar{\alpha}_t} \cdot \mathrm{x}_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \boldsymbol{\epsilon}, \tag{1}$$

where $\bar{\alpha} = \prod_{i=1}^{t} \alpha_i$, $\alpha_i = 1 - \beta_i$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

3.1.2 REVERSE DIFFUSION PROCESS

Given noisy images x_T, a denoising network is trained to iteratively recover the clean images x_0. The reverse Markov chain is defined as

$$p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t) = \mathcal{N}(\mathrm{x}_{t-1}; \mu_\theta(\mathrm{x}_t, t), \Sigma_\theta(\mathrm{x}_t, t)), \tag{2}$$

where $\mu_\theta(\mathrm{x}_t, t)$ and $\Sigma_\theta(\mathrm{x}_t, t)$ are approximated by a denoising neural network. Instead of directly predicting $\mu_\theta(\mathrm{x}_t, t)$ and $\Sigma_\theta(\mathrm{x}_t, t)$, the network is trained to predict the noise at each time step, using the loss

$$L_n = \mathbb{E}_{t,\mathrm{x}_0,\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathrm{x}_t, t)\|^2], \tag{3}$$

where x_t is generated with Equation 1. Once the denoising network is trained, the Markov chain mean is computed with

$$\mu_\theta(\mathrm{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathrm{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathrm{x}_t, t)\right), \tag{4}$$

while the covariance $\Sigma_\theta(\mathrm{x}_t, t)$ is assumed fixed (Ho et al., 2020). Finally, x_{t-1} is sampled using Equation 2. See (Ho et al., 2020) for details.

3.2 DIFFUSION MODELS FOR POINT LOCALIZATION

While diffusion models are commonly used for image synthesis, we have found that they are also able to generate conditioned sparse output, such as the map of head locations in the image of a crowd. However, with standard diffusion, this has two challenges. First, around 1,000 steps are needed to ensure the synthesis of images with sufficient quality for localization. We propose the *biased DDPM*, which leverages biased Gaussian noise to significantly lower this requirement. Second, as illustrated in Figure 4(b), the synthesized localization maps vary with the noise seed x_T used to initialize inference. To address this issue, we propose to decrease the stochasticity of the synthesized maps, using additional task-guided regularization during the training of the DDPM network. We refer to the model combining these contributions as the Task-Guided Biased Diffusion Model (TGBDM).

3.2.1 MOTIVATION AND OVERALL APPROACH

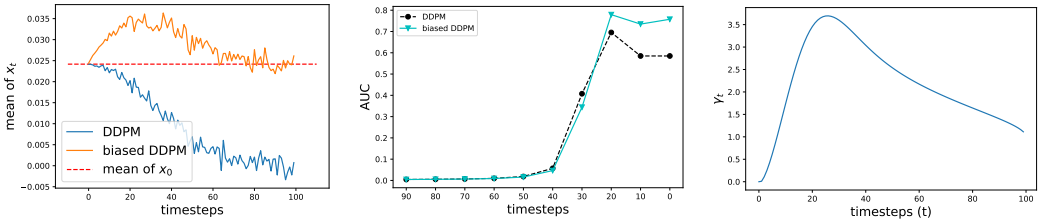
Given a crowd image, the goal of crowd localization is to produce a density map composed by a delta function at each head location. This is usually done by a deep network that takes images as input and outputs density maps of head locations (Wan et al., 2021). Crowd localization is a challenging task because there are usually many heads in a crowd image, see e.g. Figure 2, frequently resulting in low accuracy maps. We hypothesize that the problem can be addressed by using a DDPM to produce a higher quality density map, conditioned on the lower quality one output by the classic approach. This, however, poses two challenges. First, DDPMs require many iterations to produce a denoised image, particularly for very sparse images like density maps. Second, the outcome of the DDPM denoising process usually varies depending on the random seed used to initialize it. In the context of localization, this means that the output density map has some stochasticity around the localization ground truth. Ideally, this stochasticity should be small.

In this work, we propose to overcome these two problems with the approach of Figure 1. An initial density map is produced by a conventional crowd localization network. This initial prediction is used to condition the denoising DDPM. To improve inference efficiency, it is also used as a bias term in the forward diffusion process, leading to the biased DDPM presented in Section 3.2.2. This uses the binary ground truth localization map as x_0, gradually generating noisy maps x_t. A task-guided loss, defined in Section 3.2.3, is used to train the denoising network so as to be less sensitive to the stochasticity of the random seed. At inference, the biased Gaussian noise is used as input and the position map prediction is gradually generated by the reverse diffusion process.

3.2.2 BIASED DDPM

To increase the inference speed of diffusion models for localization tasks, we propose to add biased Gaussian noise during the forward diffusion process. Rather than Gaussian noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \beta_t\mathbf{I})$, we propose to sample Gaussian noise from $\tilde{\boldsymbol{\epsilon}}_t \sim \mathcal{N}(\tilde{\mathrm{x}}, \beta_t\mathbf{I})$ where $\tilde{\mathrm{x}}$ is an initial prediction, namely the low-quality localization map produced by the classic crowd localization approach. Figure 3

Figure 2: The effectiveness of using diffusion models for localization. The top row is the baseline prediction and the bottom row is our method. The threshold is increasing from left to right. The baseline predictions change dramatically while our prediction is more robust to the threshold. Note that the red crosses indicate the ground-truth and the white circles are predictions.



(a) mean of x_t v.s. timestep (b) AUC v.s. timestep (c) γ v.s. timestep

Figure 3: The comparison between DDPM and biased DDPM.

shows the mean of the noisy image x_t over diffusion steps. The traditional DDPM produces samples of decreasing mean, converging to zero mean Gaussian noise. The introduction of the bias term $\tilde{\mathrm{x}}$ constrains the mean of x_t to be closer to that of x_0, simplifying the denoising process and decreasing the number of timesteps required.

Forward Diffusion Process in Biased DDPM In the biased DDPM, the noise is defined as $\tilde{\boldsymbol{\epsilon}}_t \sim \mathcal{N}(\tilde{\mathrm{x}}, \beta_t \mathbf{I})$. Using the standard reparameterization trick leads to the forward diffusion process

$$q(\mathrm{x}_t|\mathrm{x}_{t-1}) = \mathcal{N}(\mathrm{x}_t; \sqrt{1-\beta_t}\mathrm{x}_{t-1} + \sqrt{\beta_t}\tilde{\mathrm{x}}, \beta_t\mathbf{I}), \tag{5}$$

where $\tilde{\mathrm{x}}$ is the initial density map prediction. This results in the noisy image at timestep t

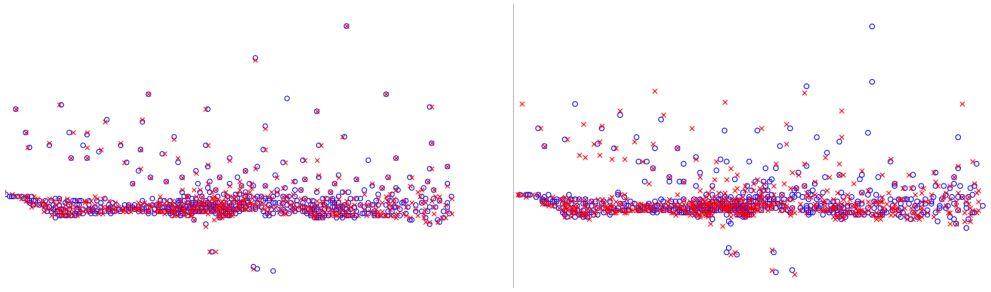
$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\tilde{\mathrm{x}} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\tilde{\mathrm{x}} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}) + \sqrt{1-\alpha_t}\tilde{\mathrm{x}} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + (\sqrt{\alpha_t(1-\alpha_{t-1})} + \sqrt{1-\alpha_t})\tilde{\mathrm{x}} + \sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + (\sqrt{\alpha_t(1-\alpha_{t-1})} + \sqrt{1-\alpha_t})\tilde{\mathrm{x}} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} \\
&= \dots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \gamma_t\tilde{\mathrm{x}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}.
\end{aligned} \tag{6}$$

where $\bar{\alpha} = \prod_{i=1}^{t}\alpha_i$, $\alpha_i = 1-\beta_i$, $\boldsymbol{\epsilon}_i, \bar{\boldsymbol{\epsilon}}_i, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and

$$\gamma_t = \sqrt{1-\alpha_t} + \sqrt{\alpha_t(1-\alpha_{t-1})} + \dots + \sqrt{\alpha_t\cdots\alpha_2(1-\alpha_1)} = \sum_{j=1}^{t}\sqrt{(1-\alpha_j)\prod_{i=j+1}^{t}\alpha_i}. \tag{7}$$

Please refer to the appendix for the full derivation. It follows that it is possible to sample x_t at an arbitrary timestep with

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \gamma_t\tilde{\mathrm{x}}, (1-\bar{\alpha}_t)\mathbf{I}). \tag{8}$$



(a) with task loss (b) without task loss

Figure 4: The comparison of stochastic sampling with or without task loss.

Figure 3(c), shows a visualization of γ_t. When compared to the standard DDPM, the mean of the noisy samples is influenced by $\tilde{\mathrm{x}}$. This has initially zero weight but quickly becomes predominant. Note that the weight $\sqrt{\bar{\alpha}_t}$ of x_0 decreases to zero. Hence, while in the early iterations, the mean of the diffusion process is determined by the groundtruth image x_0, in the later iterations it is determined by the initial prediction $\tilde{\mathrm{x}}$.

Reverse Diffusion Process of the Biased DDPM The reverse diffusion process of the biased DDPM is similar to that of the standard DDPM, with the exception that random seeds are sampled from $\mathrm{x}_T \sim \mathcal{N}(\gamma_T \tilde{\mathrm{x}}, \mathbf{I})$ instead of $\mathcal{N}(\mathbf{0}, \mathbf{I})$, as shown in Figure 1(b).

3.2.3 Task-guided Loss

We propose a task-guided loss to reduce the variability of the reconstructed localization maps with the noise seed x_T. This leverages the fact that, given the predicted noise $\boldsymbol{\epsilon}_\theta(\mathrm{x}_t, t)$, the groundtruth image can be reconstructed with

$$\tilde{\mathrm{x}}_0 = \frac{1}{\sqrt{1-\bar{\alpha}_t}}(\mathrm{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathrm{x}_t, t)). \tag{9}$$

The task-guided loss $L_t(\tilde{\mathrm{x}}_0, \mathrm{x}_0)$ is added to minimize the stochasticity of the reconstructed samples. This is a task-specific loss. For crowd localization and cell localization, we use the Generalized Loss of Wan et al. (2021) while for human pose estimation, the MSE loss is used (Ding et al., 2022). The overall loss function is $L = L_n + \lambda L_t$. When the this task-guided loss is used to train the biased DDPM, we refer to the model as Task-Guided Biased Diffusion Model (TGBDM).

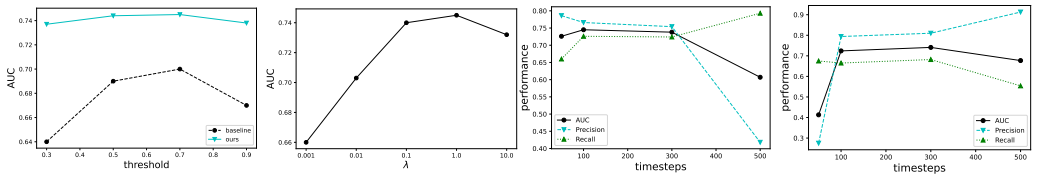
4 Experiments

In this section, we present the results of an experimental evaluation of the TGBDM. We first discuss the experimental setting and datasets used for evaluation. Then, we ablate the efficacy of the proposed components: biased DDPM and task-guided loss. Finally, we visualize the crowd localization maps produced by the TGBDM and compare to state-of-the-art algorithms for three localization tasks: crowd localization, human pose estimation, and cell localization.

4.1 Settings

Datasets: All crowd localization experiments use the NWPU-Crowd (Wang et al., 2020b) and UCF-QNRF (Idrees et al., 2018) datasets. NWPU-Crowd is the largest benchmark for crowd counting and localization. It contains 3,106 training images, 500 validation images, and 1,500 testing images. UCF-QNRF is the most widely used dataset for crowd localization with 1,535 high-resolution images (1,201/334 for training/testing). For human pose estimation, we use the popular CrowdPose (Li et al., 2019) dataset, which contains 20,000 images. For cell localization, the most recent and dense dataset, the Nuclei dataset of Sirinukunwattana et al. (2016), is used.

Metrics: Following Wan et al. (2021), Precision, Recall, and F-measure are used for NWPU-Crowd and Precision, Recall, and AUC for UCF-QNRF. Following Li et al. (2019), Average Precision (AP) and Average Recall (AR) are used to evaluate human pose estimation. For cell counting, evaluation is based on Mean Absolute Error (MAE), Grid Average Mean Error (GAME), and mean Average Precision (mAP) (Ciampi et al., 2022).



(a) AUC v.s. threshold (b) AUC v.s. λ (c) bias DDPM (d) DDPM

Figure 5: (a) The effectiveness of using diffusion models for localization. (b) AUC v.s. λ on NWPU-Crowd validation set. (c) (d) The performance with different timesteps with/without bias. Biased DDPM achieves better performance with a small timestep while standard DDPM requires more timesteps.

Table 1: Comparison with state-of-the-art crowd localization algorithms on NWPU-Crowd dataset.

Annotation	Method	Validation			Test		
		Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑
Box	Faster RCNN (Ren et al., 2015)	**0.964**	0.038	0.073	**0.958**	0.035	0.067
	TinyFaces (Hu & Ramanan, 2017)	0.543	**0.666**	**0.598**	0.529	0.611	0.567
	TopoCount (Abousamra et al., 2021)	-	-	-	0.695	**0.687**	**0.691**
Point	GPR (Gao et al., 2019)	0.610	0.522	0.563	0.558	0.496	0.525
	RAZ_Loc (Liu et al., 2019a)	0.692	0.569	0.625	0.666	0.543	0.598
	AutoScale_loc (Xu et al., 2022)	0.701	0.638	0.668	0.673	0.574	0.620
	Crowd-SDNet (Wang et al., 2021b)	-	-	-	0.651	0.624	0.637
	CLTR (Liang et al., 2022)	0.739	0.713	0.726	0.694	**0.676**	0.685
	GL (Wan et al., 2021)	-	-	-	0.800	0.562	0.660
	GL + TGBDM	**0.766**	**0.726**	**0.745**	**0.805**	0.670	**0.731**

Network architecture and training: We adopt the UNet commonly used to implement DDPMs and follow the guided diffusion model (Dhariwal & Nichol, 2021). We set the number of channels 32 for acceleration. We use Adam optimizer with a learning rate of $1e$-4. λ is set to 1 according to the experiment shown in 5(b). The initial crowd localization map is produced by the generalized loss (GL) method of Wan et al. (2021). This is a state-of-the-art approach for localization, as shown in Tables 1 and 2. We will refer to this method as the baseline in what follows. The initial density map is concatenated with the noisy image for conditioning (Dhariwal & Nichol, 2021).

4.2 Ablation Studies

4.2.1 Effectiveness of using diffusion models for localization

We start by verifying the effectiveness of diffusion models for localization. The main limitation of classical methods is the smoothness of the predicted localization maps. Two post-processing steps are required to generate the final locations: a search for local maximum responses and a filtering of the predicted points below a threshold. Most methods are quite sensitive to this threshold. Figure 2 compares the crowd locations of the proposed TGBDM (bottom row) to those of the GL baseline (top row), for various thresholds. The maps predicted by the TGBDM are much more robust to the threshold value. Figure 5(a) shows that the AUC of the TGBDM is significantly superior to that of the GL method for all threshold values and much more insensitive to the choice of threshold. Its performance is between 4 and 10 points better than that of the baseline.

4.2.2 Model components

We next evaluate the effectiveness of the TGBDM components. Table 3 compares various methods. The baseline is the GL method used to produce the initial predictions. GL + DDPM uses a standard DDPM to refine the GL predictions. This uses the common approach to condition DDPMs,

Table 2: Comparison with state-of-the-art crowd localization algorithms on UCF-QNRF dataset.

Method	Precision ↑	Recall ↑	F1 ↑
CL (Idrees et al., 2018)	0.758	0.598	0.668
LCFCN (Laradji et al., 2018)	0.779	0.524	0.627
LSC-CNN (Sam et al., 2020)	0.758	0.747	0.753
AutoScale_loc (Xu et al., 2022)	0.813	0.758	0.784
TopoCount (Abousamra et al., 2021)	0.818	0.790	0.803
CLTR (Liang et al., 2022)	0.822	0.798	0.810
GL (Wan et al., 2021)	0.782	0.748	0.764
GL + TGBDM	**0.833**	**0.826**	**0.835**

Table 3: The effectiveness of different components on NWPU-Crowd validation set.

Method	NWPU-Crowd			
	Precision ↑	Recall ↑	F1 ↑	Avg. dist. ↓
GL	**0.821**	0.605	0.697	-
GL + DDPM	0.722 (9e-4)	0.448 (9e-4)	0.553 (6e-4)	27.4
GL + DDPM + task bias	0.650 (8e-4)	0.593 (1e-3)	0.620 (1e-3)	23.5
GL + DDPM + task loss	0.810 (8e-4)	0.682 (7e-4)	0.741 (6e-4)	8.5
GL + TGBDM (ours)	0.766 (7e-4)	**0.726** (5e-4)	**0.745** (4e-4)	**7.9**

Table 4: Comparison with state-of-the-art human pose estimation algorithms on CrowdPose dataset.

Methods	AP ↑	AP_{50} ↑	AP_{75} ↑	AR ↑	AR_{50} ↑	AR_{75} ↑	AP_{easy} ↑	AP_{medium} ↑	AP_{hard} ↑
Bottom-up									
OpenPose (Cao et al., 2021)	-	-	-	-	-	-	62.7	58.7	32.3
HigherHRNet (Cheng et al., 2020)	65.9	86.4	70.6	-	-	-	73.3	66.5	57.9
HigherHRNet Multi-scale (Cheng et al., 2020)	67.6	87.4	72.6	-	-	-	75.8	68.1	58.9
SPM (Nie et al., 2019)	63.7	85.9	68.7	-	-	-	70.3	64.5	55.7
DEKR (Geng et al., 2021)	65.7	85.7	70.4	-	-	-	73.0	66.4	57.5
DEKR Multi-scale (Geng et al., 2021)	67.0	85.4	72.4	-	-	-	75.5	68.0	56.9
PINet (Wang et al., 2021a)	68.9	88.7	74.7	-	-	-	75.4	69.6	61.5
PINet Multi-scale (Wang et al., 2021a)	69.9	89.1	75.6	-	-	-	76.4	70.5	62.2
Top-down									
Mask R-CNN (He et al., 2017)	57.2	83.5	60.3	65.9	89.5	69.4	69.4	57.9	45.8
AlphaPose(Fang et al., 2017)	61.0	81.3	66.0	67.6	86.7	71.8	71.2	61.4	51.1
Simple baseline (Xiao et al., 2018)	60.8	81.4	65.7	67.3	86.3	71.8	71.4	61.2	51.2
CrowdPose (Li et al., 2019)	66.0	84.2	71.5	72.7	89.5	77.5	75.5	66.3	57.4
OPEC-Net (Qiu et al., 2020)	70.6	86.8	75.6	-	-	-	-	-	-
HRNet (Sun et al., 2019)	71.3	91.1	77.5	-	-	-	80.5	71.4	62.5
HRNet† (Sun et al., 2019)	72.8	92.1	78.7	-	-	-	81.3	73.3	65.5
TransPose-H (Yang et al., 2021)	71.8	91.5	77.8	75.2	92.7	80.4	79.5	72.9	62.2
HRFormer-B (Yuan et al., 2021)	72.4	91.5	77.9	75.6	92.7	81.0	80.0	73.5	62.4
HRFormer-B + TGBDM	73.6	92.5	79.8	76.7	93.1	82.2	81.1	74.4	63.6
I^2R-Net (Ding et al., 2022)	77.4	**93.6**	83.3	80.3	94.5	85.5	83.8	78.1	69.3
I^2R-Net + TGBDM	**77.8**	93.5	**84.2**	**80.7**	**94.7**	**86.0**	**84.2**	**78.6**	**69.7**

where the conditioning GL localization map is simply concatenated with the noisy image. The table shows that standard conditioning does not work for localization, even degrading the performance of GL. Introducing the bias of Section 3.2.2 (GL+DDPM+task bias) substantially improves the DDPM refinement but not enough to recover the original GL performance. Adding the task loss of Section 3.2.3 without bias (GL+DDPM+task loss) produces even larger gains, outperforming GL. However, the best results are obtained when bias and task loss are combined (TGBDM), leading to a significant gain over GL (F-1 score of 0.745 v.s. 0.697). In addition, the variances (in braces) are smaller with task loss, showing that the latter reduces the stochastic effect of diffusion models.

To better understand the difference between predictions obtained with different seeds, we compute the average distance (Avg. dist.) between two predicted point sets. The correspondence between the point sets is computed with bipartite graph matching and the average distance between corresponding points used to measure the difference between predictions. Table 3 shows that the average distance decreases dramatically when the task-guided loss is used. Figure 4 visualizes predictions with and without task-guided loss. Red crosses and blue circles are generated with two different seeds. The predictions generated with task-guided loss are much more consistent, illustrating how the loss reduces the effects of seed stochasticity.

4.2.3 Convergence Speed

Table 3 shows that methods with task bias usually have higher recall and F1 score, indicating they detect more heads in the crowd. In addition, as shown in Figures 5(c) and 5(d), the biased DDPM only requires 100 time steps to achieve its best performance, far less than GL + DDPM , which requires 300 steps and never reaches similar performance. The biased DDPM mostly increases recall over time. Further increasing the time steps fails to improve performance because precision decreases. The localization maps have more peaks above threshold, but their localization is less precise. This is likely because as the head density increases, there are more interactions between adjacent peaks and it is difficult to maintain the same precision. Interestingly, the same does not happen with the standard DDPM, for which recall never increases. It seems that, with standard conditioning, the DDPM is not able to create more complex crowd scenes, just to refine the precision of the initial head locations. It localizes much fewer heads with higher precision. These observations are also consistent with the comparison of the DDPM and biased DDPM in Figure 3. For the biased

Table 5: Comparison with state-of-the-art cell localization algorithms on Nuclei dataset.

Methods	MAE ↓	GAME (L=1) ↓	GAME (L=2) ↓	GAME (L=3) ↓	GAME (L=4) ↓	mAP ↑
S-UNet (Falk et al., 2019)	62.4	66.9	75.1	95.3	138.4	66.8
D-CSRNet (Li et al., 2018)	37.3	45.7	58.2	77.6	100.5	27.7
FRCNN (Ren et al., 2015)	96.5	103.8	112.6	133.9	168.2	57.9
GL (Wan et al., 2021)	32.4	36.7	50.8	77.2	120.9	71.1
GL + TGBDM	**31.8**	**34.2**	**46.3**	**69.0**	**111.4**	**74.5**

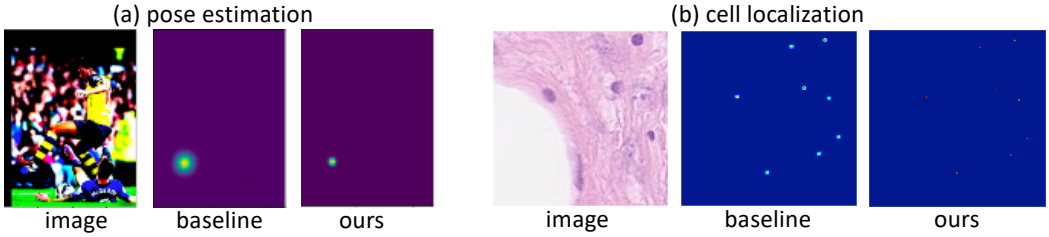


Figure 6: Visualization of (a) pose estimation and (b) cell localization. The baseline predictions are smooth while our predictions are sharper. Please zoom in for more details.

DDPM, the mean of x_t approaches the mean of x_0 for larger time steps, making the prediction easier. For the DDPM the mean decreases to 0, which makes the prediction harder.

4.3 Comparison with state-of-the-art

4.3.1 Crowd localization

For crowd localization, we compare the TGBDM with state-of-the-art algorithms on two datasets. Since the GL + TGBDM method already improves on the state-of-the-art method GL (Wan et al., 2021), which is used to produce the initial predictions, it is not suprising that it achieves state-of-the-art results on this task. This is shown in Table 1 for NPWU-Crowd and Table 2 for UCF-QNRF.

4.3.2 Human Pose Estimation

Table 4 compares the localization performance of the TGBDM to previous methods for pose estimation on the CrowdPose dataset. We apply the TGBDM with initial predictions from two state-of-the-art methods: HRFormer-B Yuan et al. (2021) and I^2R-Net Ding et al. (2022). Performance improves in both cases, establishing a new state-of-the-art for the combination of I^2R-Net and TGBDM. We further visualize the prediction in Figure 6 (a) where our prediction is sharper than the baseline.

4.4 Cell Localization

Cell localization is a challenging localization task since the background can contain similar textures to those of foreground objects. Moreover, datasets are small, e.g. Nuclei only contains 100 images, usually not enough to train an effective diffusion model. Nevertheless, as shown in Table 5, the combination of GL + TGBDM achieves a new state-of-the-art on this task, significantly outperforming classical approaches like the segmentation-based S-UNet, density-based D-CSRNet (Li et al., 2018), detection-based FRCNN (Ren et al., 2015), and GL itself, which is based on distribution matching (Wang et al., 2020a). As shown in Table 5, GL + TGBDM achieves the best MAE, GAME, and MAP, confirming its effectiveness for challenging localization tasks even with limited training examples. A visualization is shown in Figure 6 (b).

5 Conclusion

In this work, we proposed the TGBDM model for accurate and efficient point localization. This combines a biased DDPM, which improves inference speed and accuracy by simplifying the reverse diffusion process, and a task-guided loss, which decreases the variance of predictions from different sampling seeds. We demonstrated that, with these contributions, diffusion processes become a powerful mechanism to refine the predictions of state-of-the-art point localization approaches, producing predictions that are both sharper and can recover more points. The effectiveness of the TGBDM model has been demonstrated with experiments on three challenging point localization tasks. Future work will concentrate on enhancing efficiency since the primary limitation of our method is its computational demand.

REFERENCES

Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 872–881, 2021.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5386–5395, 2020.

Luca Ciampi, Fabio Carrara, Giuseppe Amato, Claudio Gennaro, et al. Counting or localizing? evaluating cell counting and detection in microscopy images. In *VISIGRAPP (4: VISAPP)*, pp. 887–897, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Yiwei Ding, Wenjin Deng, Yinglin Zheng, Pengfei Liu, Meihong Wang, Xuan Cheng, Jianmin Bao, Dong Chen, and Ming Zeng. Iˆ 2r-net: intra-and inter-human relation network for multi-person pose estimation. *arXiv preprint arXiv:2206.10892*, 2022.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2334–2343, 2017.

Junyu Gao, Tao Han, Qi Wang, and Yuan Yuan. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. *arXiv preprint arXiv:1912.03677*, 2(5), 2019.

Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14676–14686, 2021.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 951–959, 2017.

Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–546, 2018.

Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 547–562, 2018.

Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10863–10872, 2019.

Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1091–1100, 2018.

Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pp. 38–54. Springer, 2022.

Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1217–1226, 2019a.

Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. Point-query quadtree for crowd counting, localization, and more. *arXiv preprint arXiv:2308.13814*, 2023a.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I˙2 sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023b.

Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6469–6478, 2019b.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pp. 483–499. Springer, 2016.

Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6951–6960, 2019.

Agne Paulauskaite-Taraseviciene, Kristina Sutiene, Justas Valotka, Vidas Raudonis, and Tomas Iesmantas. Deep learning-based detection of overlapping cells. In *Proceedings of the 3rd International Conference on Advances in Artificial Intelligence*, pp. 217–220, 2019.

Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pp. 488–504. Springer, 2020.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R Venkatesh Babu. Locate, size, and count: accurately resolving people in dense crowds via detection. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2739–2751, 2020.

Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3365–3374, 2021.

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

Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1974–1983, 2021.

Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33:1595–1607, 2020a.

Dongkai Wang, Shiliang Zhang, and Gang Hua. Robust pose estimation in crowded scenes with direct pose-level inference. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6278–6289. Curran Associates, Inc., 2021a.

Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43 (6):2141–2149, 2020b.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

Yi Wang, Junhui Hou, Xinyu Hou, and Lap-Pui Chau. A self-training approach for point-supervised object detection and counting in crowds. *IEEE Transactions on Image Processing*, 30:2876–2887, 2021b.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.

Chenfeng Xu, Dingkang Liang, Yongchao Xu, Song Bai, Wei Zhan, Xiang Bai, and Masayoshi Tomizuka. Autoscale: learning to scale for crowd counting. *International Journal of Computer Vision*, 130(2):405–434, 2022.

Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11802–11812, 2021.

Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*, 2021.

A APPENDIX

Full derivation of x_t.

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\tilde{\mathrm{x}} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\ &= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\tilde{\mathrm{x}} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}) + \sqrt{1-\alpha_t}\tilde{\mathrm{x}} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\ &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + (\sqrt{\alpha_t(1-\alpha_{t-1})} + \sqrt{1-\alpha_t})\tilde{\mathrm{x}} + \sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1},\end{aligned} \tag{10}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For two independent Gaussians r.v.s, $\boldsymbol{\nu}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$ and $\boldsymbol{\nu}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2^2\mathbf{I})$, their sum is another Gaussian r.v. $\boldsymbol{\nu}_3 \sim \mathcal{N}(\mathbf{0}, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$. In other words, if $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \boldsymbol{\epsilon}_3$ are independent Gaussians $\mathcal{N}(\mathbf{0}, \mathbf{I})$, then $\sigma_1\boldsymbol{\epsilon}_1 + \sigma_2\boldsymbol{\epsilon}_2 = \sqrt{\sigma_1^2 + \sigma_2^2}\boldsymbol{\epsilon}_3$. Therefore, we have

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + (\sqrt{\alpha_t(1-\alpha_{t-1})} + \sqrt{1-\alpha_t})\tilde{\mathrm{x}} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} \\ &= \dots \\ &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \gamma_t\tilde{\mathrm{x}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\end{aligned} \tag{11}$$

where $\bar{\boldsymbol{\epsilon}}_i, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\bar{\alpha} = \prod_{i=1}^{t}\alpha_i$, and

$$\gamma_t = \sqrt{1-\alpha_t} + \sqrt{\alpha_t(1-\alpha_{t-1})} + ... + \sqrt{\alpha_t\cdots\alpha_2(1-\alpha_1)} = \sum_{j=1}^{t}\sqrt{(1-\alpha_j)\prod_{i=j+1}^{t}\alpha_i}. \tag{12}$$