Crowd Counting Using Diffusion-Based Latent Space

학 과 명 정보통신공학과

지도교수 홍성은

팀 구성원 안민혁/손기훈/이시원

Latent Diffusion Models를 활용한 Crowd Counting

안전사고의 예방을 위해 인구 밀집 지역의 군집 인원 파악은 중요한 문제로 자리잡고 있음. Computer Vision 분야에서도 영상 속의 정확한 인구수를 세기 위해 다양한 시도가 이루어짐에 따라, 우리는 딥러닝 기반 이미지 생성에서 우수한 성능을 보이는 Latent Diffusion Models의 원리를 인구수 추정에 적용하는 방법을 제안함.

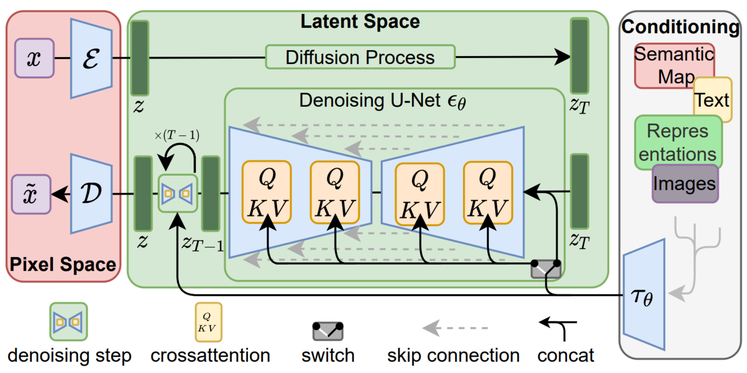

Figure 1. Latent Diffusion Models, which form the basis of our model.

Figure 2. The output of our model when given an RGB image input. Green dots are marked on the part where the human head is recognized.

Introduction

- ❑ 목표: Latent Diffusion Models를 활용한 인구 수 추정
- ❑ 주제선정배경: 도시 계획, 보안, 안전, 마케팅 등 다양한 분야에서 인구 밀집 지역의 인구수를 파악하는 것은 현대 사회의 중요한 문제임. 이에 따라 우리는 Computer Vision 기반의 기술을 활용하여 영상 내의 인구수를 추정하고자 함. 최근 이미지 생성 분야에서 주목 받고 있는 Diffusion Models의 확률적인 특성을 인원 수 추정 분야에서도 활용한다면 보다 정확한 인구수와 위치 측정이 가능할 것이라고 기대함.
- ❑ 앞으로의 연구방향: Crowd Counting의 정확도를 더욱 높이고, 모델을 경량화 할 예정

Model Arichitecture

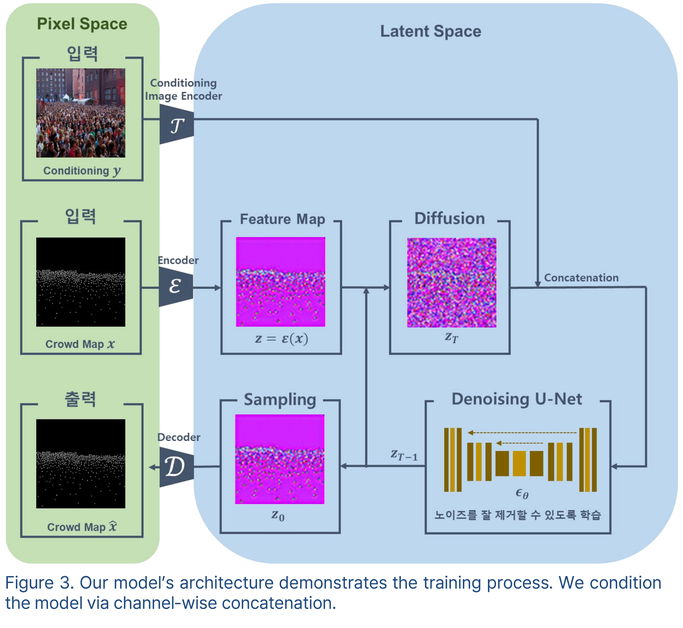

Figure 3. Our model's architecture demonstrates the training process. We condition the model via channel-wise concatenation.

Diffusion Models are a type of probablisitic models that are specifically designed to learn the data distribution $p(x)$ by iteratively denoising a normally distributed variable. This denoising process can be seen as learning the reverse process of a fixed Markov Chain of length T. Conditional distributions of the form $p(z|y)$ can be achieved by implementing a conditional denoising autoencoder with $\epsilon_\theta(z_t, t, y)$, where z_t is a noisy version of z. With the incorporation of conditional information, diffusion models can generate samples that align with the desired characteristics specified by the input y. A crowd map $x \in \mathbb{R}^{1\times1\times256\times256}$ is encoded and downsampled by a factor of 4 by $\mathcal{E}(x)$ into a vector-quantization-regularized latent representation $z = \mathcal{E}(x)$, where $z \in \mathbb{R}^{1\times1\times64\times64}$. On the other hand, RGB Image y is mapped into $\zeta \in \mathbb{R}^{1\times3\times64\times64}$ by a domain-specific encoder $\mathcal{T}(y)$ that adjusts the dimensions. These representations are then channel-wise concatenated and further mapped to the intermediate layers of the U-Net via a cross-attention layer. Based on crowd map-RGB image pairs, we then learn our LDMs(Latent Diffusion Models) via Eq. 1. where $\mathcal{E}_\theta$ is optimized.

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \mathcal{T}(y)\right)\right\|_2^2\right] \quad (1)$$

We can predict the distribution p and z_0 from the forward diffusion process q based on the current step z_t, It is specified as

$$p(z_0) = \int_z p(z_T)\prod_{t=1}^{T} q(z_{t-1}|z_t, z_\theta(z_t, t)) \quad (2)$$

Then, the Decoder $\mathcal{D}$ reconstructs the crowd map $\hat{x}$ from the latent space, giving $\hat{x} = \mathcal{D}(z_0) = \mathcal{D}(\mathcal{E}(x))$.

Sampling Process

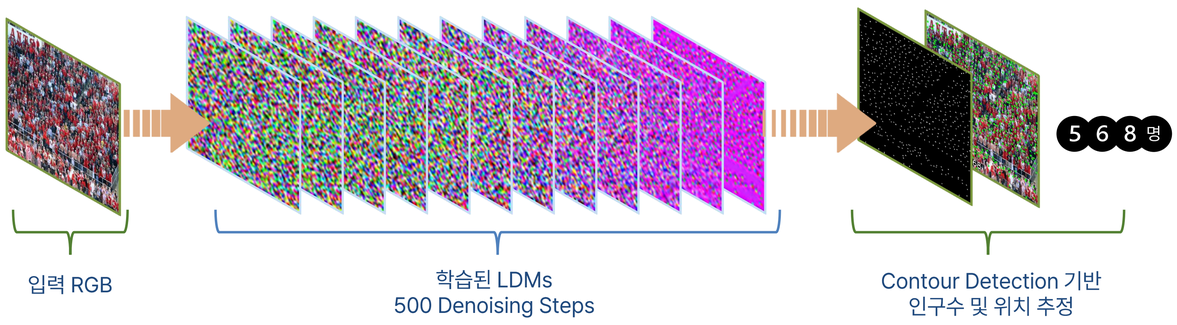

Figure 4. This diagram shows how an RGB image input is used to estimate the number of individuals. Feature maps($z_T, z_{T-1}, \ldots, z_0$) are obtained by performing 500 sampling steps using DDIM (Differentiable Diffusion Models), followed by the generation of a crowd map through contour detection.

Technical Specifications

① Diffusion Models
- 이미지 생성 기법 중에서 최근 가장 주목받는 Diffusion Models를 활용
- Diffusion Models의 장점을 실생활에 도움이 될 수 있는 정보를 생성할 수 있는 수단으로 활용

② Latent Space
- 고차원 데이터를 저차원으로 인코딩하여 복잡성을 줄이면서도 데이터의 중요한 특성을 보존
- 데이터를 낮은 차원으로 압축하여 연산량 감소 및 컴퓨팅 자원 절약 효과

③ Single RGB Input
- 열화상 카메라 등의 높은 비용이 발생하는 추가적인 센서 없이 인구수 및 위치 예측 가능

④ RGB Conditioning via Channel-wise Concatenation
- 너비와 높이가 동일한 $\mathcal{E}(x)$와 $\mathcal{T}(y)$를 채널 축으로 연결하는 방식
- RGB Image가 사람 위치에 대한 가이드 역할을 하여 위치 정보를 학습

⑤ Contour Detection
- 이미지 내의 객체나 영역의 경계를 개별적으로 식별하고 추출하는 기법

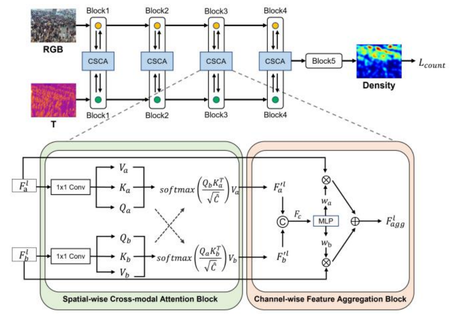

Figure 5. RGB-T crowd counting models use both RGB color and thermal information.

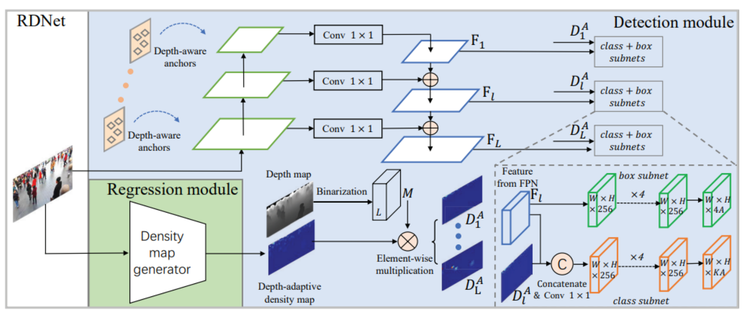

Figure 6. Similarly, RGB-D crowd counting models use both RGB information and depth information.

Comparison with Other Service

현재까지도 정확한 인구수 측정을 위한 연구는 지속적으로 이루어지고 있음. 주된 연구는 인구수 측정의 정확도를 높이기 위해 깊이 정보 또는 열화상 정보 등을 RGB 이미지와 함께 활용함. 따라서 높은 정확도를 가지지만, 깊이 및 열화상 정보 취득을 위한 추가적인 센서를 사용해야 하기 때문에 높은 Cost가 발생한다는 단점이 있음. 반면에 우리의 연구는 저가의 카메라 센서로도 쉽게 취득할 수 있는 RGB 이미지만으로 보다 정확한 인구수와 위치 측정을 가능하도록 함.

Analysis

현재 인구 밀집 지역의 인원 수를 파악하는 일에 대한 관심이 더욱 커지는 추세. 이에 따라 정확한 인구 수를 측정하기 위한 연구가 지속되고 있음. 그러나 고가의 센서를 함께 사용하는 기법은 실생활에서 일반인들이 사용하기 어렵다는 한계가 있음. 앞으로는 저가의 카메라 센서와 적은 컴퓨팅 자원만으로 누구나 쉽고 정확하게 인원 수를 측정할 수 있는 연구들이 늘어날 것이라 예상함.

Evaluation

Method	ShanghaiTech A		ShanghaiTech B	
	MAE	RMSE	MAE	RMSE
Zhang et al.	<u>181.8</u>	<u>277.7</u>	32.0	49.8
Marsden et al.	126.5	173.5	23.8	33.1
MCNN	110.2	173.2	26.4	41.3
Cascaded-MTL	101.3	152.4	20.0	31.1
Switching-CNN	90.4	135.0	21.6	33.4
Ours	181.1	262.2	70.9	98.8

Table 1. Estimation errors on ShanghaiTech Part-A and ShanghaiTech Part-B dataset. We evaluated the performance of our method on publicly available crowd-counting datasets, and compared it with earlier methods. The results indicate that our method achieves highest metrics (Excluding the <u>underlined</u>), which is not desirable compared to the earlier methods. This demonstrates that out method does not meet the standards of state-of-the-art performance. However, while the reconstruction capability of LDMs may have limitations for tasks requiring precise accuracy in pixel space, such as crowd counting, our results offer fresh evidence of the potential benefits of utilizing Diffusion Models and Latent Space for location-based crowd counting. Additionally, due to the nature of diffusion models, they typically require large amounts of data, and if trained in a more powerful GPU environment, one can expect improved generalization performance.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{C}_i - C_i\right| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\widehat{C}_i - C_i\right|^2} \quad (4)$$

$$C_i = \sum_{h=0}^{H}\sum_{w=0}^{W} x_{h,w} \quad (5)$$

MAE in Eq. 3. is a metric that quantifies the average magnitude of errors. RMSE in Eq. 4., on the other hand, is a metric that capture the square root of the average of squared errors. N is the number of images in one test sequence and C_i is the ground truth of counting. In Eq. 5., H and W show the height and width of the crowd map respectively while $x_{h,w}$ is the pixel at (h, w) of the generated crowd map. The above equations are evaluation metrics commonly used to evaluate the performance of crowd counting models. They generally indicate that smaller values represent superior models.

Experimental Result

현재 도시계획, 안전, 보안, 마케팅 등의 다양한 분야에서 인구 밀집 지역의 인원 수를 정확하게 파악 하는 기술을 필요로 하고 있음. 이와 더불어 다수의 센서 데이터를 활용해 보다 정확한 인원 수를 측정하기 위한 연구들이 많이 이루어지고 있음. 반면 우리는 단일 센서 입력과 적은 컴퓨팅 비용으로 인원수와 위치를 정확하게 측정하기 위한 방법으로 Latent Diffusion Models의 사용을 제안함.

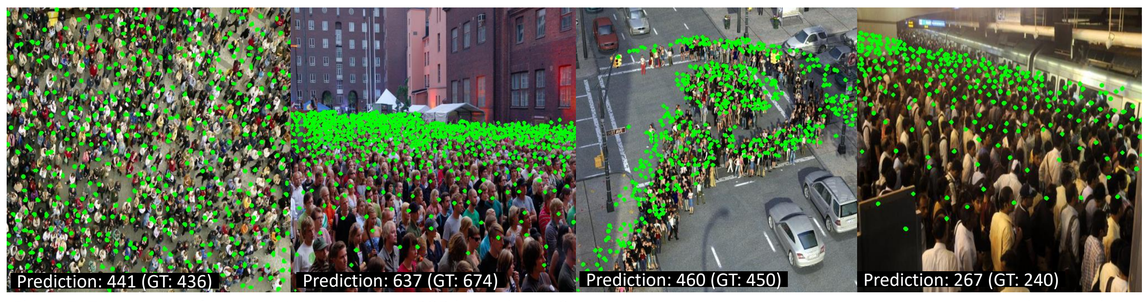

Figure 7. The outputs of our model when given an RGB image input. Green dots are marked on the part where the human head is recognized. Each of them shows the results in high-angle, low-angle, peculiar shape, and indoor scene.

References

[1] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." CVPR 2022.

[2] Zhang, Yingying, et al. "Single-image crowd counting via multi-column convolutional neural network." CVPR 2016.

[3] Zhang, Youjia, Soyun Choi, and Sungeun Hong. "Spatio-channel Attention Blocks for Cross-modal Crowd Counting." ACCV. 2022.

[4] Lian, Dongze, et al. "Density map regression guided detection network for rgb-d crowd counting and localization." CVPR 2019.

[5] Li, Yuhong, Xiaofan Zhang, and Deming Chen. "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes." CVPR 2018.

Hi Technology 정보통신공학과