Latent Diffusion Models를 이용한 Crowd Counting

Crowd Counting Using Diffusion-Based Latent Space

손기훈, 안민혁, 이시원
(Gihun Son, Minhyuk An, Siwon Lee)
지도교수: 홍성은 **(Advisor: Sungeun Hong) (인)**

Abstract: Crowd Counting has been a continuously researched topic in the field of Computer Vision, aiming to accurately estimate the number of people in images. This research breaks the conventional approach of utilizing additional data such as depth or thermal data and instead introduces a new perspective. By leveraging the advantages of Diffusion Models (DMs), known for their outstanding performance in image generation, our Latent Diffusion Model-based approach is developed to generate high-quality Crowd Maps and estimate the number of people without the need for expensive sensors. This research showcases the potential of incorporating DMs into crowd counting and contributes to the development of modern society by combining the strengths of different fields.

Keywords: Crowd Counting, Latent Diffusion Models, Latent Space

Ⅰ. 서론

Crowd Counting은 영상 속의 인원 수를 정확하게 추정하기 위해 Computer Vision 분야에서 꾸준히 연구되어 온 주제다. 최근 연구들은 인원 수 추정의 정확도를 높이기 위한 다양한 방법론을 제시한다.

현대 사회가 발전해 나감에 따라 밀집 지역에 대한 인원 수는 매우 중요한 정보가 되고 있으며 관련 연구에 대한 관심과 중요도가 증가하고 있는 추세다. 특히 도시를 중심으로 인구가 밀집되면서 항상 과밀 위험이 존재하는 지역에는 안전사고의 위험이 언제나 존재한다. 따라서 인구 밀집도를 파악하고 사고를 예방하는 것은 우리 사회의 매우 중요한 문제가 되었다. 이에 따라 정확한 인구 수를 추정하기 위한 연구가 지속되고 있다.

기존의 연구는 RGB 영상과 함께 깊이 데이터 또는 열화상 데이터 등을 추가로 사용하는 Cross-Modal Crowd Counting을 중심으로 발전해왔다. 이러한 방법은 RGB 영상만으로는 얻을 수 없는 정보를 추가적으로 활용하여 우수한 성능을 보인다. 그러나 일반적으로 깊이 데이터와 열화상 데이터 확보에는 고비용의 특수한 센서가 필요하다는 단점이 있다. 우리 연구는, 성능 향상을 위해 추가적인 센서 데이터를 활용하는 기존 연구의 틀을 깨고 새로운 관점을 제시하고자 한다. 구체적으로, 고비용의 센서 데이터를 사용하지 않음과 동시에 높은 정확도의 Crowd Map을 생성하기 위한 연구를 진행했다.

우리는 최근 이미지 생성 분야에서 우수한 성능을 보여 전세계의 주목을 받고 있는 Diffusion Models(DMs)[6]의 장점에 주목했다. Text-to-Image, Image-to-Image 등의 방식으로 높은 품질의 이미지를 생성할 수 있는 DMs의 특성을 이용해 인구 밀집 정보를 포함하는 유의미한 데이터인 Crowd Map를 생성해 인구 수 추정을 수행하고자 했다.

특히, 우리는 정확한 인구 수 정보를 대중화하는 것에도 집중했기 때문에, 학습시 컴퓨팅 자원의 소모가 많은 DMs의 단점을 보완하고자 Latent Space라는 압축된 저차원의 공간에서 Diffusion 과정을 진행하여 한정된 컴퓨팅 자원으로도 우수한 성능을 보이는 Latent Diffusion Models(LDMs)[1]를 기반 모델로 선정하였다. 모델 학습시에는 Ground Truth 입력 데이터로써 Crowd Map을, LDMs의 Conditioning을 위한 입력 데이터로는 RGB 이미지를 사용하였다. 학습과정에서 Crowd Map과 RGB 이미지 사이의 연관성을 학습한 LDMs가, Sampling 과정에서는 단일 RGB 이미지만을 입력으로 사용하여 고품질의 Crowd Map을 생성할 수 있도록 모델을 구성하였다.

DMs는 주로 이미지 생성 연구의 모델로 사용된다. 또한, Crowd counting에 대한 기존 연구들을 살펴보면, DMs를 활용하여 제안하는 기법들은 찾아볼 수 없을 만큼 우리의 연구는 도전적인 시도다. 기존 연구와 비교한 성능지표만을 본다면 아직 개선해야 할 부분이 많다. 그러나 본 논문이 제안하는 방법은 Crowd Counting 연구의 새로운 방법론을 제시했을 뿐만 아니라, 생성형 모델의 활용 가능성, 앞으로의 연구를 위한 기반 마련, 더 나아가 서로 다른 분야에서 사용된 기술의 장점을 융합하여 현대 사회의 발전에 기여할 수 있다는 사례가 될 것이라 기대한다.

Latent Diffusion Models를 이용한 Crowd Counting

Crowd Counting Using Diffusion-Based Latent Space

손기훈, 안민혁, 이시원
(Gihun Son, Minhyuk An, Siwon Lee)
지도교수: 홍성은 **(Advisor: Sungeun Hong)**

Abstract: Crowd Counting has been a continuously researched topic in the field of Computer Vision, aiming to accurately estimate the number of people in images. This research breaks the conventional approach of utilizing additional data such as depth or thermal data and instead introduces a new perspective. By leveraging the advantages of Diffusion Models (DMs), known for their outstanding performance in image generation, our Latent Diffusion Model-based approach is developed to generate high-quality Crowd Maps and estimate the number of people without the need for expensive sensors. This research showcases the potential of incorporating DMs into crowd counting and contributes to the development of modern society by combining the strengths of different fields.

Keywords: Crowd Counting, Latent Diffusion Models, Latent Space

Ⅰ. 서론

Crowd Counting은 영상 속의 인원 수를 정확하게 추정하기 위해 Computer Vision 분야에서 꾸준히 연구되어 온 주제다. 최근 연구들은 인원 수 추정의 정확도를 높이기 위한 다양한 방법론을 제시한다.

현대 사회가 발전해 나감에 따라 밀집 지역에 대한 인원 수는 매우 중요한 정보가 되고 있으며 관련 연구에 대한 관심과 중요도가 증가하고 있는 추세다. 특히 도시를 중심으로 인구가 밀집되면서 항상 과밀 위험이 존재하는 지역에는 안전사고의 위험이 언제나 존재한다. 따라서 인구 밀집도를 파악하고 사고를 예방하는 것은 우리 사회의 매우 중요한 문제가 되었다. 이에 따라 정확한 인구 수를 추정하기 위한 연구가 지속되고 있다.

기존의 연구는 RGB 영상과 함께 깊이 데이터 또는 열화상 데이터 등을 추가로 사용하는 Cross-Modal Crowd Counting을 중심으로 발전해왔다. 이러한 방법은 RGB 영상만으로는 얻을 수 없는 정보를 추가적으로 활용하여 우수한 성능을 보인다. 그러나 일반적으로 깊이 데이터와 열화상 데이터 확보에는 고비용의 특수한 센서가 필요하다는 단점이 있다. 우리 연구는, 성능 향상을 위해 추가적인 센서 데이터를 활용하는 기존 연구의 틀을 깨고 새로운 관점을 제시하고자 한다. 구체적으로, 고비용의 센서 데이터를 사용하지 않음과 동시에 높은 정확도의 Crowd Map을 생성하기 위한 연구를 진행했다.

우리는 최근 이미지 생성 분야에서 우수한 성능을 보여 전세계의 주목을 받고 있는 Diffusion Models(DMs)[6]의 장점에 주목했다. Text-to-Image, Image-to-Image 등의 방식으로 높은 품질의 이미지를 생성할 수 있는 DMs의 특성을 이용해 인구 밀집 정보를 포함하는 유의미한 데이터인 Crowd Map를 생성해 인구 수 추정을 수행하고자 했다.

특히, 우리는 정확한 인구 수 정보를 대중화하는 것에도 집중했기 때문에, 학습시 컴퓨팅 자원의 소모가 많은 DMs의 단점을 보완하고자 Latent Space라는 압축된 저차원의 공간에서 Diffusion 과정을 진행하여 한정된 컴퓨팅 자원으로도 우수한 성능을 보이는 Latent Diffusion Models(LDMs)[1]를 기반 모델로 선정하였다. 모델 학습시에는 Ground Truth 입력 데이터로써 Crowd Map을, LDMs의 Conditioning을 위한 입력 데이터로는 RGB 이미지를 사용하였다. 학습과정에서 Crowd Map과 RGB 이미지 사이의 연관성을 학습한 LDMs가, Sampling 과정에서는 단일 RGB 이미지만을 입력으로 사용하여 고품질의 Crowd Map을 생성할 수 있도록 모델을 구성하였다.

DMs는 주로 이미지 생성 연구의 모델로 사용된다. 또한, Crowd counting에 대한 기존 연구들을 살펴보면, DMs를 활용하여 제안하는 기법들은 찾아볼 수 없을 만큼 우리의 연구는 도전적인 시도다. 기존 연구와 비교한 성능지표만을 본다면 아직 개선해야 할 부분이 많다. 그러나 본 논문이 제안하는 방법은 Crowd Counting 연구의 새로운 방법론을 제시했을 뿐만 아니라, 생성형 모델의 활용 가능성, 앞으로의 연구를 위한 기반 마련, 더 나아가 서로 다른 분야에서 사용된 기술의 장점을 융합하여 현대 사회의 발전에 기여할 수 있다는 사례가 될 것이라 기대한다.

II. 관련 연구

Cross-Modal Crowd Counting Models

Crowd counting의 성능을 향상시키기 위해 제안되는 방법은 다양하지만, 가장 대중적으로 사용되는 방법은 RGB 이미지와 함께 깊이 정보와 열화상 등의 추가적인 정보를 활용하는 것이다. 이를 Cross-Modal Crowd Counting이라고 한다. 이때, 깊이 데이터를 활용하는 기법을 RGB-D Crowd Counting이라고 칭하며 대표적인 모델로는 RDNet[4]이 있다. 열화상 데이터를 활용하는 기법은 RGB-T Crowd Counting이라고 칭하며 대표적인 모델로는 IADM[5]가 있다. 이들은 단일 RGB 영상만을 사용한 Crowd Counting 기법에 비해 우수한 성능을 보인다. 하지만 깊이 또는 열화상 데이터를 얻기 위해서는 특수한 센서를 통해 얻어야 하기 때문에, 단일 RGB 영상을 사용하는 모델보다 상대적으로 비용이 많이 들어 대중화가 어렵다는 단점이 있다.

일반적으로 Crowd Counting Models는 사람의 위치 및 수에 대한 정보를 가지고 있는 Crowd Map을 생성한다. 따라서 Crowd Map을 정확하게 나타내는 것이 가장 중요한 문제라고 할 수 있다.

잠재 공간 (Latent Space)

Latent Space는 고차원 입력 데이터의 특징을 효율적으로 표현하여 매핑하는 저차원의 공간을 의미한다. 데이터의 차원을 줄여 표현을 단순화시킬 수 있기 때문에 컴퓨팅 비용 절감에 도움이 된다. 잠재 공간은 다양한 기계 학습 모델에서 주로 사용되며, 특히 오토인코더(Autoencoder), 변이형 오토인코더(VAE), 생성적 적대 신경망(GAN) 등에서 많이 사용된다.

VQ-Regularization

VQ-Regularization는 Latent Variable의 연속적인 값을 이산적인 값으로 변환하는 과정으로 Latent Variable의 표현이 더 간단하고 구조화된 형태로 인코딩될 수 있도록 하여 안정성과 일반화 능력을 향상시키는데 도움을 준다. 우리 연구는 Crowd Map을 Latent Space로의 매핑을 가능하게 하면서, Latent Space 내에서의 Diffusion 연산을 위해 VQ-Variational AutoEncoder(VQ-VAE)[3]를 인코더와 디코더(Decoder)로써 사용한다.

Diffusion Models (DMs)

주로 정규 분포화된 변수로부터 반복적으로 노이즈를 제거함으로써 데이터 분포 $p(x)$를 학습하기 위해 설계된 확률적 모델이다. 노이즈 제거 과정은 데이터에 T번의 노이즈 추가를 수행하는 마르코프 체인(MarKov Chain)의 반대 과정을 학습하는 것으로 볼 수 있다.

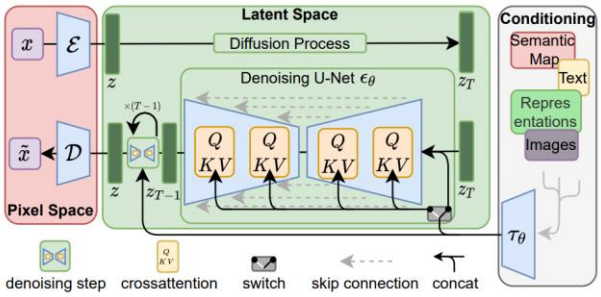

그림 1. 논문의 Base Model인 LDMs의 Architecture

Latent Diffusion Models (LDMs)

LDMs는 Latent Space의 압축된 공간에서 DMs를 사용하는 방법론을 제시한다. [1]에 따르면 Pixel Space에서 DMs를 학습시키기 위해서는 수백개의 GPU가 사용되고, 순차적인 평가 과정으로 인해 추론에 많은 비용이 발생한다고 말한다. 따라서 한정된 컴퓨팅 자원으로 DMs를 학습시키기 위해 Autoencoder를 사용하여 Latent Space안에서 DMs를 적용하는 것을 제안한다. 논문의 연구 결과를 보면, Pixel-Based의 DMs에 비해 상당히 적은 컴퓨팅 자원을 소모함과 동시에, 우수한 성능을 보이는 것을 알 수 있다.

III. 방법

Conditioning Mechanisms

$p(x)$ 와 같은 조건부 확률 분포(Conditional Probability Distribution)는 z의 잡음 버전인 z_t를 사용하여 조건부 노이즈 제거 Autoencoder $\epsilon_\theta(z_t, t, y)$를 구현함으로써 얻을 수 있다. 그 결과, 모델이 Conditioning 입력 y의 영향을 받는 샘플링 결과 z를 얻을 수 있게 된다. Crowd Map $x \in \mathbb{R}^{B\times C_x\times H\times W}$ 은 VQ-VAE 인코더 $\mathcal{E}(x)$에 의해 f배로 Downsampling된 후 VQ-Regularized 된 $z=\mathcal{E}(x)$로 매핑된다. 여기서 $z \in \mathbb{R}^{B\times C_x\times \frac{H}{f}\times \frac{W}{f}}$ 이다. 반면에 RGB 영상 y는 인코더 $\mathcal{T}(y)$에 의해 $\zeta \in \mathbb{R}^{B\times C_y\times \frac{H}{f}\times \frac{W}{f}}$로 매핑된다. 이러한 표현 ζ는 z와 Channel-wise Concatenation되어 Denoising U-Net의 입력이 된다. 즉, Crowd Map x와 RGB 영상 y 쌍을 기반으로 식 1.을 통해 우리의 LDMs가 학습된다. 여기서 ϵ_θ이 최적화된다.

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

현재 단계 z_t를 기반으로 Diffusion 과정 q에서 분포 p와 z_0를 예측할 수 있다. 이는 다음과 같이 정의된다.

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

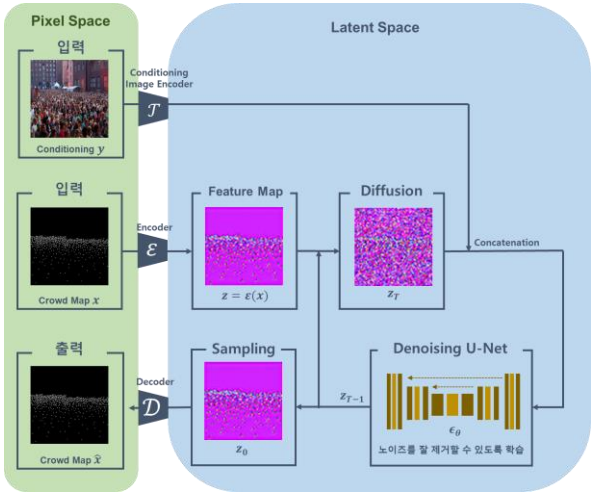

그림 2. 우리 모델의 Architecture를 보여주는 그림. Channel-wise Concatenation 방식을 통한 Conditioning Mechanisms를 보여주고 있다.

Decoder $\mathcal{D}$는 y를 입력으로 했을 때의 z를 Latent Space로부터 재구성한다. 예측된 Crowd Map은 $\hat{x} = \mathcal{D}(z_0) = \mathcal{D}(\mathcal{E}(x))$ 와 같이 나타낼 수 있다.

위치 기반 Crowd Counting

각 사람의 머리 위치에 커널의 가중치 합이 1인 가우시안 커널(Gaussian Kernel)을 적용하여 만들어진 Crowd Map을 입력으로 LDMs를 학습시키고, 학습된 LDMs 샘플링 결과의 이미지 픽셀 값을 전부 합산하면 사람의 수를 계산할 수 있다. 하지만 DMs 기반의 방법은 모델이 노이즈를 완벽하게 제거하지 못하는 경우가 많아 노이즈로부터 자유롭지 못하므로 이미지의 픽셀 값을 전부 합산하는 방법은 효과적이지 못하다.

그 대안으로, Adaptive Thresholding을 통해 이진화된 샘플링 결과에 이미지 픽셀 값이 연속적으로 분포하는 영역을 찾은 뒤, 윤곽선 검출 (Contour Detection) 기법으로 개별 영역의 개수를 세는 방법을 제안한다.

보편적으로 사용되는 Global Thresholding은 이미지 전체 영역의 모든 픽셀 값에 대해 특정 값을 초과하는지 검사하는 방법이다. 하지만 위 방법은 배경에 균일하지 않은 노이즈가 존재하는 경우 결과에 치명적인 영향을 준다. 이러한 문제를 극복하기 위해, 우리는 Local Thresholding, 즉 Adaptive Thresholding을 사용한다. 위 방법을 통해 사람 위치에 해당하는 개별 영역과

그 주변 밝기 값을 통계적으로 분석하여 이진화를 수행하고, 노이즈에 강인한 결과를 얻을 수 있다.

이진화된 샘플링 결과에 Contour Detection을 수행하면 개별 영역에 대한 Perceptual Grouping이 가능하다. [7, 8] 이때, 각 Group의 이미지 평면 상 위치와 개수로부터 위치 기반 Crowd Counting 결과를 얻을 수 있다.

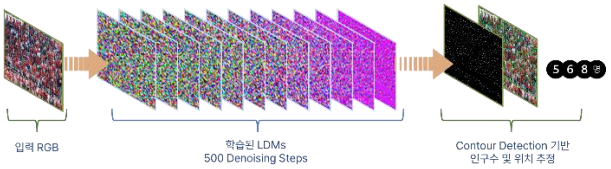

그림 3. 위 그림은 모델의 인구수 및 위치 추정 과정 보여준다. DDIM (Differentiable DIffusion Implicit Models)[9]으로부터 500 번의 노이즈 제거를 수행해 $(\boldsymbol{z_T}, \boldsymbol{z_{T-1}}, \dots, \boldsymbol{z_0})$을 얻은 후 Contour Detection을 통해 인구 수와 위치를 추정한다.

IV. 실험

Benchmark Datasets

ShanghaiTech Part-A: 대표적인 Crowd Counting Dataset으로 Train 이미지 300 장과 Test 이미지 182 장, 총 482 장의 이미지로 구성되어 있다. 우리는 Train 이미지를 9:1 비율로 분배하여 각각 Train, Validation Dataset으로 사용하였다. 따라서, 모델 학습에는 Train, Validation, Test Dataset을 각각 270 장, 30 장, 182 장으로 설정하였다. 데이터 셋에 포함되어 있는 각 이미지에 대한 사람의 위치와 수 정보를 기반으로, 커널의 가중치 합이 1인 Gaussian Kernel을 적용해 Crowd Map 쌍을 생성했다.
ShanghaiTech Part-B: Train 이미지 400 장과 Test 이미지 316 장, 총 716 장의 이미지로 구성되어 있다. 마찬가지로 Train 이미지를 9:1 비율로 분배하여 각각 Train, Validation Dataset으로 사용하였다. 따라서, 모델 학습에는 Train, Validation, Test 이미지를 각각 360 장, 40 장, 316 장로 설정하였다. 또한, ShanghaiTech Part-A에서의 방식과 동일하게 Crowd Map 쌍을 생성했다.

Model Settings

우리가 설정한 모델의 세부사항은 다음과 같다. Learning rate는 1.0×10^{-6}, Mini-Batch Size B를 12, Loss Function은 L_2 로 설정하였다. Diffusion 과정에서의 Sampling Steps t는 1000, Downsampling factor f는 4로 설정했다. 입력 x와 y는 모두 $x, y \in \mathbb{R}^{12\times3\times256\times256}$의 차원으로 설정했다. 그 결과, Latent Space의 z 와 ζ 는 $z, \zeta \in \mathbb{R}^{12\times3\times\frac{256}{4}\times\frac{256}{4}}$의 차원을 갖는다.

표 1. Conditioning Method 결과 비교

Conditioning Method	ShanghaiTech Part-A	
	MAE	RMSE
Cross-Attention	305.8	541.6
Concatenation	**181.1**	**262.2**

그림 4. Conditioning method로써 Concatenation와 Cross-Attention 방식이 적용된 결과 비교를 위한 그림
좌)Concatenation 우)Cross-Attention

Conditioning Method

LDMs는 Cross-Attention Layer를 사용하여 입력 x와 Conditioning y의 연관성을 학습해 효과적으로 이미지를 생성할 수 있다. 이러한 방법은 y가 Text 데이터일 때, 즉 Text-to-Image를 수행할 때 매우 우수한 성능을 보이는 것으로 전해진다. 하지만, 우리가 진행하는 연구는 Conditioning y가 이미지로 주어지는 Image-to-Image를 수행해야 한다. 따라서 우리는 Concatenation 방식과 Cross-Attention 방식 중 어떤 Conditioning Method가 우리의 모델 학습에 더 효과적인지 검증해야 할 필요가 있었다.

표 1.로부터 각 방식을 적용하여 학습시킨 결과를 확인할 수 있다. 이때, Concatenation 방식의 결과가 상대적으로 더 우수했다. 또한, 그림 4.로부터 Cross-Attention을 사용한 모델은 Concatenation에 비해 사람의 위치 및 밀집도에 대한 정보를 정확하게 생성하지 못하는 것을 확인할 수 있다.

해당 결과에 대한 원인은 Cross-Attention을 사용하기 위해 y를 토큰화하여 임베딩하는 과정에서 사람의 위치 정보가 보존되지 않기 때문에, 위치 기반의 Crowd Counting을 수행하는 우리의 모델에 적합하지 않은 방식이라고 분석했다. 따라서 우리는 Concatenation 방식을 Conditioning Method로 선정했다.

Evaluation Metircs

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{C}_i - C_i\right| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\hat{C}_i - C_i\right|^2} \quad (3)$$

표 2. 기존 방법과 우리 방법의 성능 비교 분석 표

Method	ShanghaiTech Part-A		ShanghaiTech Part-B	
	MAE	RMSE	MAE	RMSE
Zhang et al.	181.8	277.7	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**

Mean Absolute Error(MAE)와 Root Mean Squared Error(RMSE)는 Crowd Counting Models를 평가할 때 가장 보편적으로 사용되는 평가 지표다. 우리는 MAE와 RMSE를 사용하여 모델의 성능을 측정했다. MAE는 모든 절대 오차의 평균값이고, 식 2.와 같이 나타낼 수 있다. RMSE는 모든 절대 오차 제곱값의 평균에 제곱근을 적용한 값으로 오차의 표준편차로 해석할 수 있고, 식 3.과 같이 나타낼 수 있다. 식에서 $\hat{C}_i$은 각 이미지에 대해 모델이 예측한 사람의 수이고, C_i는 각 이미지 내 사람 수에 대한 Ground Truth 값이다.

타 연구와의 결과 비교

우리의 모델 결과는 표 2.에서 볼 수 있듯이 ShanghaiTech Part-A 데이터셋에 대해 MAE, RMSE가 각각 181.1, 262.2고, ShanghaiTech Part-B 데이터셋에 대해 MAE, RMSE가 각각 70.9, 98.8이다. 기존 연구들의 성능과 비교해본다면 좋지 않은 결과를 보인다. 그림 5.과 6.의 결과들은 실제 ShanghaiTech Part-A의 Test 결과 사진이다. Ground Truth와 Prediction의 차이가 각각 224, 54, 69, 228, 735로 오차가 크게 발생하는 것을 볼 수 있다. 뿐만 아니라, 사람의 위치에 대한 예측 결과도 완벽하게는 일치하지 않는다. 우리는 결과에 대한 원인을 분석해보았고, 크게 3가지로 결론지을 수 있었다. 그 내용은 아래와 같다.

모델의 한계

VQ-VAE는 우리가 기반 모델로 선정한 LDMs에서 데이터의 차원을 축소 및 확장하는 역할을 한다. 하지만, 다른 차원으로의 매핑은 데이터 정보의 손실로 이어질 수 있다. 따라서 LDMs의 Autoencoder는 성능을 저하시킬 가능성이 있기 때문에, 정보의 손실 없이 데이터를 효과적으로 매핑하기 위한 추가적인 연구가 필요하다.

한정된 컴퓨팅 자원이 원인이 될 수 있다. 생성형 모델에 사용되는 DMs의 가장 큰 장점은 고품질의 이미

지를 생성한다는 점이다. 하지만, 그만큼의 많은 컴퓨팅 자원이 필요하다는 단점이 있다. 우리는 이러한 한계를 극복하고자, Latent Space라는 압축된 공간에서 Diffusion을 진행하여 컴퓨팅 자원을 줄이는 LDMs를 기반 모델로 선정했음에도 불구하고, 여전히 Autoencoder나 GAN과 같은, 다른 생성형 모델에 비해 많은 컴퓨팅 자원을 요구했다. 따라서, 컴퓨팅 자원의 한계로 인한 모델 성능 저하는 불가피 했음을 그 원인으로 분석했다.

DMs의 한계로는 모델 학습을 위한 방대한 데이터셋이 필요하다는 점이 있다[2]. 우리가 진행한 연구 또한 DMs를 기반한 방법이기 때문에, 다양하고 많은 학습 데이터를 요구한다. 하지만 우리가 Benchmark로써 사용한 ShanghaiTech 데이터 셋은 한정된 데이터 양을 통해 학습된 모델을 평가하고 비교하기 위한 데이터셋이다. 따라서 우리가 제시한 방법은 해당 데이터 셋 사용 시 모델이 충분히 학습될 수 없기 때문에 기존 연구에 비해 성능 지표가 좋지 않음을 원인으로 분석했다.

연구 성과

우리의 연구에는 위와 같은 한계점들이 있지만, 분명한 성과도 존재한다. 우리는 LDMs를 기반으로 한 Crowd Counting 수행 가능성을 확인했다. 기존 연구가 제시하는 방법에 비해 우리 모델의 성능 지표는 상대적으로 좋지 않다. 하지만 그림 5.와 6.의 결과를 통해 우리 모델의 활용 가능성을 확인할 수 있었다.

우리의 모델은 Crowd Map을 예측해 생성할 때, 사람의 머리에 대한 위치(초록색 점)를 예측한다. 그림 5.의 결과 사진을 보면 초록색 점이 사진의 상단 부분에 밀집도 높게 찍혀 있는 것을 볼 수 있는데, 이는 사람의 머리가 사진의 위쪽에 밀집되어 있기 때문이다. 반면 사람의 몸이 많은 비중을 차지하고 있는 사진의 하단에는 초록색 점이 거의 찍히지 않았다. 즉, 우리의 모델이 사람의 밀집도를 식별하고, 유의미한 Crowd Map을 생성할 수 있음을 의미한다.

그림 6.의 결과 사진을 보면 초록색 점의 분포로부터 위치 인식 결과가 배경이 아닌, 사람이 있는 위치에 올바르게 나타나는 경향성을 확인할 수 있다. 즉, 우리 모델이 사람만을 명확하게 구별하면서 Crowd Map을 올바르게 예측할 수 있음을 의미한다.

위 결과를 통해 우리는 기존 연구에 비해 성능적인 발전을 이루지는 못했지만, DMs를 Crowd Counting 연구에 적용할 가치가 있고, 이를 발전시킨다면 이미지 생성에서의 우수성이 해당 연구에서의 새로운 전환점을

그림 5. ShanghaiTech Part-A 테스트 셋에 대한 우리 모델의 샘플링 결과

그림 6. ShanghaiTech Part-A 테스트 셋에 대한 우리 모델의 샘플링 결과

향후 연구 계획 및 기대 효과

우리가 분석한 향후 연구 계획 및 기대효과는 다음과 같다. 우선 해결해야 할 문제는 모델이 영상내의 사람 머리 크기 변화에 상관없는 동일한 인식 결과를 갖을 수 있도록 하는 것이다. 그림 6.을 보면 카메라의 구도에 따라 거리가 먼 사람과 가까이 있는 사람의 머리 크기가 다른 것을 볼 수 있다. 하지만 모델이 이에

따른 차이를 인식하지 못하고, 이미지에서 가까이 있는 사람의 머리에 다수의 초록색 점이 생성되는 등의 문제가 발생한다. 이 문제는, 다양한 카메라 구도, 사람 수, 환경 등이 포함된 방대한 데이터 셋을 통해 학습한다면 해결할 수 있을 것이라 예상한다.

두번째로, 사람의 수 오차에 대한 Loss Function을 추가하는 것이다. 우리는 실제로 사람 수 오차에 대한 Loss Function 적용을 시도하였다. 하지만, 사람의 위치를 기반으로 Crowd Map을 생성하는 우리 모델의 특성으로 인해 단순하게 Loss 값을 더하는 형태의 단순한 응용은 학습에 유의미한 도움이 되지 않았다. 따라서 추가적인 연구를 통해 사람 수 예측 오차에 대한 적합한 Loss Function을 설계하여 모델에 적용한다면, Crowd Counting 성능이 향상될 것이라고 기대한다.

V. 결론

Crowd Counting의 정확도를 높이기 위한 다양한 시도는 현대사회에서 가치가 높은 연구 주제이다. 또한 누구나 인구 수 정보를 쉽고 정확하게 취득할 수 있는, 정보의 대중화도 중요한 문제라고 할 수 있다. 정확도를 높이기 위해 추가적인 정보를 활용하는 기존의 기법은 가장 잘 알려져 있고, 동시에 매우 효과적이다. 따라서 현재 많은 연구들은 깊이 또는 열화상 등의 다양한 Modality를 활용하여 우수한 성능을 보이는 Crowd Counting Models를 제안한다.

우리는 정확도와 접근성, 두가지 문제를 모두 해결하고자 DMs를 Crowd Counting 연구에 적용하는, 아직 대중화되지 않은 방법을 시도하였다. 특히 Latent Space에서 Diffusion 과정을 진행하여 적은 컴퓨팅 자원으로 정확한 이미지 생성이 가능한 LDMs를 기반 모델로 채택하여 우리의 목적성을 더했다.

모델의 성능은 ShanghaiTech Part-A에 대한 MAE, RMSE가 각각 181.1과 262.2로 기존 연구에 비해 상대적으로 좋지 않다. 하지만 우리 연구는 DMs가 사람 특징과 인구 밀집도에 대한 정보를 유의미하게 학습한다는 결과를 도출하였고, 이는 생성형 모델이 Crowd Counting 연구에도 효과적으로 적용될 수 있음을 의미한다.

본 논문은 단순하게 성능을 위한 연구를 진행한 것이 아니다. 기존 연구를 개선하는 것에서 벗어나 새로운 방법론을 제시한다. 우리의 연구가 Crowd Counting Models 중 최고 성능을 달성하지는 못했지만, 기존 연구의 틀을 깨고 그 가능성을 확인함으로써 Crowd Counting 연구의 새로운 변곡점이 되었기를 기대한다.

VI. 참고문헌

[1] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Moon, Taehong, et al. "Fine-tuning Diffusion Models with Limited Data." NeurIPS 2022 Workshop on Score-Based Methods.

[3] Van Den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." Advances in neural information processing systems 30 (2017).

[4] Lian, Dongze, et al. "Density map regression guided detection network for rgb-d crowd counting and localization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[5] Liu, Lingbo, et al. "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[6] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.

[7] Gong, Xin-Yi, et al. "An overview of contour detection approaches." International Journal of Automation and Computing 15 (2018): 656-672.

[8] Rezanejad, Morteza, et al. "Contour-guided Image Completion with Perceptual Grouping." arXiv preprint arXiv:2111.11322 (2021).

[9] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." arXiv preprint arXiv:2010.02502 (2020).

손 기 훈

2018.03-2024.02 인하대학교
정보통신공학과 졸업 예정
관심 분야:
Computer Vision, Generative Models
E-mail: thlg60@naver.com

안 민 혁

2018.03~2023.08 인하대학교
정보통신공학과 졸업 예정
관심 분야:
Multi Modal Learning, Computer Vision
E-mail: als7928@daum.net

이 시 원

2017.03~2024.02 인하대학교
정보통신공학과 졸업 예정
관심 분야:
Computer Vision
E-mail: siwon2717@gmail.com