WiFi CSI 신호 기반 영상 이미지 생성을 위한 AI 모델 학습*

홍지연[0], 김혜정, 남수연, 유재현
성신여자대학교 AI 융합학부
20241283@sungshin.ac.kr, 20241202@sungshin.ac.kr, 20241204@sungshin.ac.kr,
jhyoo@sungshin.ac.kr

AI model training for WiFi CSI-based video image generation

Jiyeon Hong[0], Hyejeong Kim, Suyeon Nam, Jaehyun Yoo
School of AI convergence, Sungshin Women's University
20241283@sungshin.ac.kr, 20241202@sungshin.ac.kr, 20241204@sungshin.ac.kr,
jhyoo@sungshin.ac.kr

요 약

본 논문은 WiFi CSI(Channel State Information) 신호를 활용하여 사람의 위치 정보를 나타내는 이미지를 생성하는 모델을 제안한다. 제안된 모델은 CSI 시계열 데이터를 잠재 공간으로 압축하는 기반 구조를 따르며, 이를 기반으로 공간적 정보를 포함한 이미지를 복원하는 end-to-end 방식으로 구성된다. 실험 결과, 제안한 모델은 사람의 연속적인 위치 특징이 유지된 이미지를 생성함을 확인하였으며, 이를 통해 CSI 신호 로부터 공간적 위치 정보를 효과적으로 추정할 수 있음을 보였다. 이러한 결과는 위치 추정 문제를 기존의 회귀 기반 접근이 아닌 생성 기반 접근으로 해결할 수 있는 가능성을 시사한다.

1. 서 론

최근 컴퓨터 비전 및 기계학습 기술의 발전으로 RGB 카메라, LiDAR, 레이더 센서를 활용한 사람 위치 추정 연구가 활발히 이루어지고 있다. 그러나 RGB 카메라는 사생활 침해 우려가 있으며, LiDAR와 레이더는 장비 비용과 전력 소모 측면에서 한계를 가진다. 이에 따라 기존 무선 통신 인프라를 활용할 수 있는 WiFi CSI(Channel State Information) 기반 비접촉 위치 추정 기술이 주목받고 있다.

WiFi CSI는 송수신기 사이의 무선 채널 변화 정보를 포함하며, 사람의 이동이나 자세 변화에 따라 신호 패턴이 달라진다. 선행 연구에서는 WiFi CSI를 활용해 벽 너머 공간 정보를 영상 형태로 복원할 수 있음을 보였으며[1], 이는 CSI 신호가 사람의 위치 및 공간적 변화를 반영할 수 있음을 시사한다. 그러나 CSI 신호는 시간적 변동성과 환경적 영향을 크게 받기 때문에, 이를 위치나 형상 정보로 안정적으로 변환하는 것은 여전히 어려운 문제이다.

본 연구에서는 WiFi CSI 시계열 데이터를 입력으로 사용하여 사람의 위치 정보가 반영된 인체 메쉬 이미지를 생성하는 모델을 제안한다. 제안 모델은 Transformer와 TCN을 결합하여 CSI의 장기적 시간 의존성과 국소적 시계열 패턴을 함께 학습하도록 구성하였다. 이를 통해 CSI 기반 인체 메쉬 이미지 생성 및 간접 위치 추정 가능성을 확인하고, 환경 변화에 따른 일반화 한계를 완화하는 것을 목표로 한다.

2. 본 론

2.1. CSI 데이터 수집

본 실험에서는 송신기(Tx)와 수신기(Rx)로 각각 ESP32-S3 보드 1대를 사용하여 WiFi CSI 데이터를 수집하였다. 송수신 통신은 ESP-NOW 프로토콜을 기반으로 수행하였으며, ESP-IDF 환경에서 데이터 수집 서버를 구축하였다. 송신 주기는 100 Hz로 설정하였으나, 실제 서버 수신 속도는 네트워크 및 하드웨어 특성으로 인해 약 90 Hz로 측정되었다. 수집된 CSI 데이터 중 본 연구에서는 진폭(amplitude) 정보만을 활용하였다. 실험은 6.5 m × 3.15 m 크기의 실내 공간에서 수행되었으며, CSI 신호와 ESP32-CAM으로 촬영한 사람 이미지 데이터를 동시에 수집하였다. 이미지 데이터는 CSI 데이터보다 낮은 프레임 속도로 획득되므로, CSI Timestamp를 기준으로 가장 가까운 이미지 프레임을 매칭하여 데이터 쌍을 구성하였다. 또한 모델의 일반화 성능을 평가하기 위해 학습 데이터와 테스트 데이터는 서로 독립적인 환경에서 수집하였다.

2.3 학습 이미지 구축

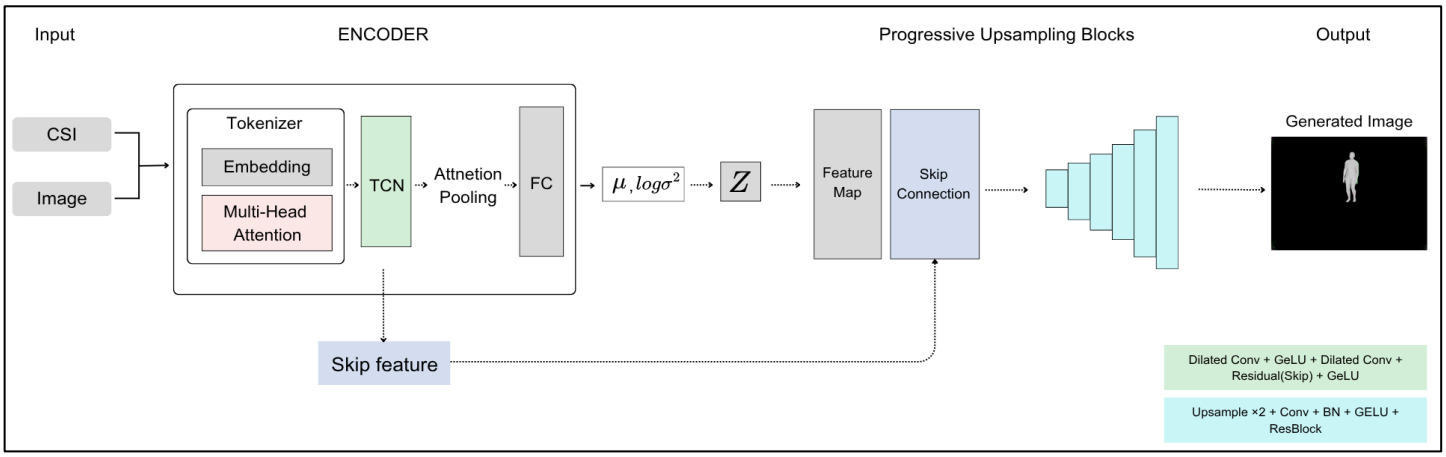
원본 영상을 직접 복원 대상으로 사용할 경우, 배경 정보, 조명 변화, 의복 차이 등과 같은 비본질적 요소가 함께 반영되어 모델의 일반화 성능이 저하될 수 있다. 특히 본 연구의 목적은 RGB 영상 자체를 사실적으로 복원하는 것이 아니라, WiFi CSI 신호로부터 사람의 위치와 전반적인 인체 형상 정보를 추정하는 데 있으므로, 배경이나 색상 정보보다 인체의 구조적 정보만을 학습 대상으로 구성하는 것이 적절하다.

따라서 본 연구에서는 인체의 동작 및 형태 정보만을 효과적으로 반영하기 위해 각 프레임에서 인체 스켈레톤을 추출하고, 이를 기반으로 메쉬 이미지를 생성하였다. 먼저, Ultralytics에서 제공하는 YOLOv8 기반 pose estimation 모델[2]을 이용하여 프레임별 인체 keypoint를 JSON 형태로 추출하였다. 이후 추출된 스켈레톤 정보를 입력으로 받아 3D 인체 메쉬를 직접 추정하는 Key2Mesh 모델[3]을 적용하여 사람의 형태를 메쉬 형태로 재구성하였다.

생성된 메쉬 이미지는 사람의 자세, 위치, 실루엣과 같은 핵심적인 형상 정보를 포함하며, 원본 영상에 포함된 배경 및 조명 변화의 영향을 줄일 수 있다. 이를 통해 모델이 CSI 신호와 인체 위치 및 형상 정보 간의 관계를 보다 집중적으로 학습할 수 있도록 하였다.

* 이 논문은 2026년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음

[그림 1] WiFi CSI 신호로부터 사람의 형태를 메쉬 이미지로 생성하기 위한 Transformer 기반 모델 구조



2.4 모델 구조

기존 WiFi CSI 기반 위치 추정 연구들은 좌표를 직접 예측하는 회귀 기반 접근이 널리 사용되어 왔다. 한편, 최근에는 CSI 신호로부터 이미지나 공간 표현을 생성하려는 시도도 일부 보고되고 있으며, 이러한 접근은 공간 구조를 보다 직관적으로 반영할 수 있는 가능성을 가진다. 본 연구는 이러한 흐름을 바탕으로, WiFi CSI 데이터를 입력으로 받아 사람의 위치 정보를 포함하는 이미지를 생성하는 Transformer 기반 생성 모델을 제안한다.

제안한 모델 구조는 [그림 2]에 나타냈다. 모델은 Encoder 에서 추출된 latent space 를 기반으로 이미지를 점진적으로 생성하는 end-to-end 학습 구조를 따른다.

입력 CSI 데이터는 (B,T,N)형태로, 배치 크기 B, 시간 길이 T, 그리고 서브캐리어 수 N을 의미한다. 입력 데이터는 선형 임베딩을 통해 고차원 feature space 로 투영된 후 Transformer 기반 Tokenizer 에 입력된다. Tokenizer 는 4 개의 attention block 으로 구성되며, 각 block 은 4-head self-attention 과 feed-forward network 로 이루어진다. attention score 계산 시 relative positional bias 를 추가하여 시점 간 상대적 위치 정보를 반영한다. 이를 통해 CSI 시계열의 전역적인 시간 의존성과 시점 간 상관관계를 효과적으로 학습한다.

이후 추출된 feature 는 Temporal Convolutional Network(TCN)에 입력된다. TCN은 서로 다른 dilation rate를 갖는 1 차원 convolution block 으로 구성되며, 각 block 은 convolution, GELU 활성화 함수, 그리고 residual connection 을 포함한다. 이러한 구조는 다양한 시간 스케일에서의 국소적 및 장기적 의존성을 동시에 반영할 수 있다. 이 과정에서 추출된 feature 는 이후 progressive upsampling block 에서 활용하기 위해 skip connection 으로 함께 전달된다. 이후 Fully Connected Layer 를 통해 잠재 공간의 분포 파라미터인 평균 μ 와 로그 분산 $\log\sigma^2$ 로 매핑된다. 잠재 변수는 reparameterization trick 을 통해 다음과 같이 샘플링된다.

$$Z = \mu + \sigma \cdot \epsilon, \epsilon \sim N(0, I)$$

이를 통해 입력 CSI 데이터의 변동성과 불확실성을 반영할 수 있는 확률적 잠재 공간을 구성한다. 이를 통해 입력 CSI 데이터의 변동성과 불확실성을 반영할 수 있는 확률적 잠재 공간을 구성한다.

Progressive upsampling block 은 샘플링된 latent vector z와 skip feature 를 입력으로 받아 이미지를 생성한다. 먼저 latent vector 는 선형 변환을 통해 $(B,512,6,8)$ 형태의 저해상도 feature map으로 변환된다. 동시에 skip feature는 1차원 convolution 과 attention pooling 을 통해 요약된 후 선형 변환을 거쳐 동일한 해상도로 변환되며, 두 feature 는 채널 방향으로 결합된다. 이러한 skip connection 은 encoder 에서 학습된 시간 정보를 이미지 생성 과정에 효과적으로 반영한다.

이후 feature map 은 bilinear upsampling 과 convolution, batch normalization, GELU 활성화 함수, 그리고 residual block 으로 구성된 upsampling block 을 반복적으로 거치며 해상도를 점진적으로 증가시킨다. 최종적으로 bilinear interpolation 을 통해 480×640 해상도로 확장한 후 convolution layer 와 sigmoid activation 을 적용하여 $(B,3,480,640)$ 크기의 RGB 이미지를 출력한다. 제안된 모델은 Transformer 와 TCN 을 결합하여 CSI 시계열의 전역 및 국소적 특징을 동시에 학습하고, 이를 잠재 공간으로 압축한 뒤 skip connection 과 progressive upsampling 구조를 활용하여 고해상도의 이미지를 효과적으로 생성한다.

2.5 손실함수

본 모델의 학습을 위해 재구성 오차를 최소화하고 잠재 공간의 정규화를 유도하는 복합 손실 함수를 설계하였다. 전체 손실 함수는 다음과 같이 정의된다.

$$L_{total} = L_{recon} + \beta \cdot KL + \lambda_{freq} L_{freq} + \lambda_{grad} L_{grad} \quad (1)$$

각 손실 항의 기여도를 조절하기 위해 β , λ_{freq} , λ_{grad} 를 하이퍼파라미터로 설정하였다. 각 손실 함수의 상세 정의와 도입 배경은 다음과 같다.

먼저 L_{recon} 은 재구성된 이미지와 원본 이미지 간의 픽셀 단위 일치성을 보장하는 역할을 한다. 본 연구에서는 이미지의 선명도를 유지하고 경계선이 흐려지는 현상을 방지하기 위해 L1 Loss를 채택하였으며, 이는 식 (2)와 같이 계산된다.

$$L_{recon} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \{\hat{x}_i\}| \quad (2)$$

이때 N은 이미지의 전체 픽셀 수이며 x_i와 $\hat{x}_i$는 각각 i번째 위치의 실제 픽셀 값과 재구성된 픽셀 값을 의미한다. 잠재 변수 z가 표준 정규 분포를 따르도록 유도하여 잠재 공간의 정규성을 확보하기 위해 KL Divergence를 적용한다. KL Divergence는 식 (3)과 같이 정의된다.

$$KL = -\frac{1}{2}\sum_{i=1}^{N}(1 + \log {\sigma_i}^2 - {\mu_i}^2 - {\sigma_i}^2) \quad (3)$$

여기서 d는 잠재 변수의 차원을, μ_i 와 σ_i 인코더를 통해 산출된 잠재 분포의 평균과 표준편차를 의미한다.

단순한 픽셀 오차를 넘어 WiFi 신호의 특성과 이미지의 구조적 정밀도를 반영하기 위해 L_{freq} 과 L_{grad} 을 추가로 도입하였다. L_{freq} 은 식 (4)와 같이 2차원 이산 푸리에 변환(2D DFT) 연산자인 F를 사용하여 이미지의 주파수 도메인 성분을 비교함으로써, 픽셀 영역에서 놓치기 쉬운 미세 패턴 정보를 보존한다.

$$L_{freq} = |F(x) - F(\hat{x})|_1 \quad (4)$$

더불어 L_{grad} 는 이미지 그래디언트 연산자 를 활용하여 공간적 변화량을 일치시킴으로써 실루엣의 윤곽선을 더욱 명확하게 한다. L_{grad}는 식 (5)와 같이 정의된다.

$$L_{grad} = \left|\nabla_x \hat{Y} - \nabla_x Y\right|_1 + \left|\nabla_y \hat{Y} - \nabla_y Y\right|_1 \quad (5)$$

여기서 $\hat{Y}$ 과 Y 는 각각 재구성된 이미지와 실제 이미지를 의미하며, ∇_x 와 ∇_y 는 이미지의 수평 및 수직 방향에 대한 변화량을 나타낸다. 또한 $\|\cdot\|_1$ L1 Norm으로, 두 이미지 간 그래디언트 편차의 평균 절대값 오차를 계산하여 엣지 성분을 보존하는 역할을 한다.

2.6 실험 결과

본 연구에서는 제안 모델의 이미지 생성 성능을 평가하기 위해 SSIM, PSNR, FID를 사용하였다. SSIM은 생성 이미지와 실제 이미지 간의 구조적 유사도를, PSNR은 픽셀 단위 복원 품질을, FID는 생성 이미지와 실제 이미지 집합 간의 분포적 유사도를 평가하는 지표이다. SSIM과 PSNR은 높을수록, FID는 낮을수록 우수한 성능을 의미한다.

성능 비교를 위해 CNN 기반 생성 모델과 MLP 인코더 및 CNN 디코더를 결합한 Base 모델을 대조군으로 설정하였다. 제안 모델은 Transformer와 TCN을 결합하여 CSI 시계열 데이터의 장기적 시간 의존성과 국소적 패턴을 함께 학습하도록 구성하였다.

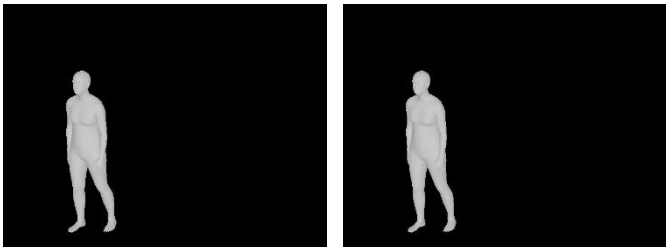
표 1 성능 지표

Method	SSIM	PSNR	FID
Ours	0.9560	21.0049	80.2165
Base	0.9276	21.3290	56.1230
CNN	0.9344	21.9387	120.8693

정량적 평가 결과, 제안 모델은 SSIM에서 0.9560으로 가장 높은 값을 보이며 실제 이미지와의 구조적 유사도 측면에서 우수한 성능을 보였다. 이는 제안 모델이 CSI 시계열 데이터로부터 사람의 전체적인 위치, 자세 방향, 실루엣과 같은 구조적 정보를 비교적 안정적으로 학습했음을 의미한다. 특히 Transformer와 TCN을 결합한 구조가 CSI 신호의 장기적 시간 변화와 국소적 패턴을 함께 반영함으로써, 인체 형상의 전반적인 구조를 유지하는 데 기여한 것으로 해석된다.

반면 PSNR은 21.0049로 비교 모델보다 낮게 나타났으며, FID 또한 Base 모델보다 높게 측정되었다. 이는 제안 모델이 구조적 유사도 측면에서는 강점을 보였지만, 픽셀 단위의 세밀한 복원 품질과 생성 이미지의 분포적 유사성 측면에서는 추가적인 개선이 필요함을 의미한다.

[그림 2] 실제 메쉬 이미지와 제안 모델의 생성 결과



(a) 실제 메쉬 이미지(GT) (b) 제안 모델 생성 결과

[그림 2]는 실제 메쉬 이미지(GT)와 제안 모델의 생성 결과를 비교한 것이다. 생성된 결과는 실제 이미지와 비교했을 때 사람의 중심 위치와 전반적인 실루엣을 비교적 유사하게 복원하였다. 다만 팔, 다리, 경계 부분의 세부 윤곽은 실제 이미지에 비해 다소 흐리게 나타났다. 그럼에도 불구하고 본 실험 결과는 WiFi CSI 시계열 데이터로부터 사람의 위치와 전반적인 형상이 반영된 이미지를 생성할 가능성을 확인했다는 점에서 의의가 있다.

3 결론 및 향후 과제

본 연구에서는 WiFi CSI 시계열 데이터를 활용하여 사람의 형태를 인체 메쉬 이미지로 복원하고, 이를 기반으로 위치 정보를 간접적으로 추정하는 생성 기반 모델을 제안하였다. 제안 모델은 Transformer와 TCN을 결합하여 CSI 신호의 장기적 시간 의존성과 국소적 시계열 패턴을 함께 학습하도록 구성되었다. 실험 결과, 제안 모델은 일부 정량 지표에서 대조 모델보다 낮은 성능을 보였으나, 정성적 결과에서는 사람의 전반적인 윤곽과 위치를 비교적 안정적으로 복원하는 경향을 확인하였다. 이는 WiFi CSI 시계열 데이터로부터 사람의 위치와 형상이 반영된 이미지를 생성할 수 있는 가능성을 보여준다.

다만 본 연구에서는 SSIM, PSNR, FID와 같은 이미지 유사도 중심 지표를 사용하였기 때문에, 실제 위치 추정 성능을 직접적으로 평가하는 데에는 한계가 있다.

향후 연구에서는 centroid error, bbox IoU, silhouette IoU, Dice score와 같은 위치 및 영역 기반 지표를 추가하여 생성 결과의 위치 정확도와 인체 영역 복원 성능을 보다 정량적으로 분석할 예정이다. 또한 복합 손실 함수의 각 항이 성능에 미치는 영향을 확인하기 위해 손실 항별 ablation study를 수행할 계획이다. 나아가 수신기(Rx)를 다중으로 배치하여 다양한 방향에서 CSI 신호를 수집함으로써 공간 정보의 해상도를 높이고, 다양한 실내 환경 및 벽 너머 상황(through-wall scenario)으로 데이터셋을 확장하여 모델의 일반화 성능과 실제 적용 가능성을 검증하고자 한다.

참 고 문 헌

[1] J. Strohmayer, R. Sterzinger, C. Stippel, and M. Kampel, “Through-Wall Imaging Based on WiFi Channel State Information,” arXiv preprint arXiv:2401.17417, 2024.

[2] G. Jocher et al., “YOLOv8 by Ultralytics,” arXiv preprint, 2023.

[3] B. Uguz, O. Suat, B. Karagoz, and E. Akbas, “MoCap-to-Visual Domain Adaptation for Efficient Human Mesh Estimation from 2D Keypoints,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1622-1632, 2024.