변 혜 원 교수 지도
석사학위 청구논문

다중 객체 추적 및 인페인팅 시스템

2024

성신여자대학교 대학원
미래융합기술공학과
이 효 진

다중 객체 추적 및
인페인팅 시스템

변 혜 원 교수 지도

이 논문을 석사 학위 논문으로 제출함

2023년 11월

성신여자대학교 대학원

미래융합기술공학과

이 효 진

인 준 서

이효진의 석사학위 논문으로 인준함

2023년 11월

심사위원장 오 장 민 (서명 또는 인)

심 사 위 원 이 규 중 (서명 또는 인)

심 사 위 원 변 혜 원 (서명 또는 인)

성신여자대학교 대학원

논 문 개 요

영상에서 콘텐츠와 무관한 객체는 시청자가 중심 객체에 집중하지 못하게 하며, 개인정보 유출과 같은 문제를 일으킬 수 있어 제거되어야 한다. 이에 본 논문에서는 경량화된 다중 객체 인페인팅 시스템을 제안한다. 중심 객체 선정 과정, 객체 추적 기술과 이미지 인페인팅(Image Inpainting) 기술을 결합하여 다중 객체를 효과적으로 제거하며, 시스템의 전 과정을 자동화하여 빠른 성능을 보장한다. 제거 대상 객체를 파악하기 위해 다비스-볼드윈 지수(Davies-Bouldin Index)를 도입하여 중심 객체를 선정하는 방법론을 제안한다. 사람의 신체 부위에 대한 탐지 결과를 이용하여 세그멘테이션(Segmentation) 마스크에 준하는 하이브리드(Hybrid) 마스크를 생성하고, YOLOv7을 도입하여 다중 객체 추적 및 마스킹 작업을 자동화한다. 또한, 인페인팅 모델의 합성곱 블록을 이산 웨이블릿 변환(Discrete Wavelet Transform)과 깊이별 분리 합성곱(Depthwise Separable Convolution)으로 구성하여 모델의 복원 능력은 유지하면서 시스템의 추론 속도를 약 9 FPS 정도로 개선하였다. PSNR, SSIM, LPIPS, FPS의 지표를 이용하여 시스템을 평가한 결과, 기존 인페인팅 모델과 비교해 FPS 대비 높은 복원 성능을 보였다.

목 차

그 림 목 차

표 목 차

Ⅰ. 서 론

소셜 미디어의 유행으로 영상 콘텐츠의 접근성이 향상되면서 영상에서의 개인정보 보호 문제에 대한 중요성이 부각 되고 있다. 영상 콘텐츠를 통한 개인정보 유출은 의도하지 않은 객체로 인해 발생하는 경우가 대부분이다. 특히 야외 촬영 등에서 다른 사람의 개인정보를 담고 있는 객체는 촬영 당시에 의식하지 못하는 경우가 많아 촬영자의 주의가 필요하다[1]. 그러나 모든 촬영 조건을 통제하는 것은 불가능하기 때문에 해당 객체를 제거하기 위한 후속 처리 작업이 필요하다.

야외 촬영 영상에서 의도하지 않게 촬영된 객체를 제거하는 문제는 다음과 같은 몇 가지 이슈가 있다. 첫째, 영상에 유지해야 하는 중심 객체와 제거 대상이 되는 불필요한 객체를 식별해야 한다. 둘째, 영상 내에는 여러 개의 불필요한 객체가 존재할 수 있다. 셋째, 다량의 이미지에서 반복적으로 다중 객체를 제거해야 하므로 인페인팅 전 과정에서의 자동화 및 경량화 과정이 필수적이다.

영상에서 객체를 제거하는 기술은 대상 객체를 식별하여 블러링 및 마스킹하는 방법부터 이미지 인페인팅과 같은 고급 기법을 활용하는 것까지 다양하다. 특히 인페인팅을 이용한 객체 제거(Object Removal) 기법은 객체가 제거된 상태로 장면을 복원함으로써 시각적으로 우수한 품질의 이미지를 생성하는 장점이 있다. 그러나, 기존의 인페인팅 연구는 대부분 단일 객체를 대상으로 하고, 주로 객체를 수동으로 마스킹하는 것을 전제로 하며, 객체가 제거된 부분의 영상 복원 품질을 향상하는 데 집중한다. 이러한 방식은 처

1) Faklaris, C., Cafaro, F., Blevins, A., O'Haver, M. A., and Singhal, N. (2020). A snapshot of bystander attitudes about mobile live-streaming video in public settings. In Informatics, 7(2), 10.

리 시간이 오래 걸리는 단점이 있다.

본 논문에서는 경량화된 다중 객체 인페인팅 시스템을 제안한다. 영상에서 중심 객체와 제거 대상이 되는 여러 객체를 식별한 후, 해당 객체의 마스킹 및 인페인팅 과정을 자동화하여 약 9 FPS 정도의 추론 속도를 보장한다. 이를 위해 다비스-볼드윈 지수[2]를 도입하여 중심 객체를 선정하고, YOLOv7을 도입하여 영상에 있는 다중 객체에 대한 마스킹 작업을 자동화한다. 특히, 개인정보 보호가 필요한 사람 객체에 대한 성능 향상을 위해 경계 상자(Bounding Box) 마스크와 세그멘테이션 마스크를 결합한 하이브리드 마스크를 제안한다. 또한, 이산 웨이블릿 변환을 도입하고 인페인팅 모델의 합성곱 블록을 개선하여 복잡한 패턴이 있는 영상에 대한 복원 성능을 향상시키고 전체 시스템을 최적화한다. 본 논문의 기여점은 다음과 같이 요약할 수 있다.

- 영상에서 다중 객체를 실시간으로 추적하고 인페인팅 하는 새로운 시스템을 제안한다.
- 다비스-볼드윈 지수를 도입해 영상의 중심 객체를 선정한다.
- 신체 부위의 경계 상자를 마스킹하여 실시간으로 세그멘테이션 마스크에 준하는 하이브리드 마스크를 생성한다.
- YOLOv7을 도입하여 영상의 불필요한 객체에 대한 마스킹을 자동화한다.
- 전체 시스템의 빠른 처리를 위해 이산 웨이블릿 변환과 깊이별 분리 합성곱[3]으로 구성된 새로운 합성곱 블록을 제안한다.

---

2) Davies, D. L., and Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

3) Pramod, R. T., Katti, H., and Arun, S. P. (2018). Human peripheral blur is optimal for object recognition. arXiv preprint arXiv:1807.08476.

Ⅱ. 이론적 배경

1. 객체 추적

객체 추적(Object Tracking) 모델은 객체 탐지(Object Detection) 모델의 분리 여부에 따라 2가지 방식으로 분류된다. 객체 탐지 모델에서 탐지된 객체 정보를 독립된 다른 모델로 전달하여 객체를 추적하는 방식과 객체 탐지 과정과 추적 과정을 단일 모델로 구성하는 방식이 그것이다.

SORT(Simple Online and Real-time Tracking)[4] 및 DeepSORT[5]는 객체 탐지 모델과 추적 모델이 독립된 모델이다. 이에 뛰어난 탐지 성능을 보여주는 기존의 객체 탐지 모델을 활용하여 객체의 정보를 얻고, 이렇게 얻은 정보를 기반으로 객체 추적을 진행한다. SORT는 객체 탐지 모델에서 객체의 위치 정보만을 이용하며, 칼만 필터(Kalman Filter)와 헝가리안 알고리즘(Hungarian Algorithm)을 활용해 객체 매칭(Object Matching)을 진행한다. DeepSORT는 SORT의 객체 추적 성능을 개선하고자 객체의 위치 정보뿐 아니라 객체 탐지 모델의 중간 특성 맵(Feature Map)을 활용하여 객체를 추적한다. 이를 통해 다중 객체 추적 정확도(Multi Object Tracking Accuracy)를 높였다. 객체 탐지 과정과 추적 과정이 분리된 방식은 객체 추적 모델이 지양하는 ID 스위칭(Identification Switching) 비율을 줄일 수 있으나 추적 과정의 알고리즘으로 인해 추론 속도가 느리다는 단점이 있다.

JDE(Joint Detection and Embedding)[6] 및 FairMOT[7]는 객체 탐지 과정

4) Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In 2016 IEEE international conference on image processing (ICIP), 3464-3468.
5) Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP), 3645-3649.

과 추적 과정을 단일 모델로 구성하여 객체의 경계 상자와 객체 추적을 위한 임베딩 벡터(Embedding Vector)를 해당 모델의 결과물로 구성하였다. JDE는 기존 객체 추적 모델의 추론 속도를 개선하고자 단일 모델 구성을 최초로 제시하였다. 해당 모델은 SORT와 비교해 다중 객체 탐지 정확도 및 속도 측면에서 우수한 성능을 보였으나, 앵커(Anchor)와 같이 객체 탐지 시에만 유용한 정보를 임베딩 벡터에도 피드백하여 ID 스위칭 발생 확률이 높았다. 이에 FairMOT는 중심 모델을 앵커-프리 객체 탐지 모델(Anchor-Free Object Detection)로 구성하여 객체 추적 성능을 높인 모델을 제안하였다. 해당 모델은 기존 객체 추적 모델과 비교해 우수한 다중 객체 추적 성능과 실시간성을 보여주었다.

2. 객체 제거

인페인팅을 이용한 객체 제거에는 여러 생성 모델을 이용할 수 있다. 그러나 이미지를 자연스럽게 생성하는 것이 목표이므로 내부적으로 생성한 이미지의 자연스러움을 평가하는 생성적 적대 신경망(Generative Adversarial Network)의 활용이 주로 연구되었다.

Uittenbogaard et al.[8]은 다양한 각도의 이미지를 활용하여 객체를 제거하는 모델을 제안하였다. 해당 모델은 같은 시점의 다양한 배경 정보를 이용하여 결과 이미지를 자연스럽게 생성하지만, 다양한 각도의 여러 이미지가

6) Wang, Z., Zheng, L., Liu, Y., Li, Y., and Wang, S. (2020). Towards real-time multi-object tracking. In European Conference on Computer Vision, 107-122.
7) Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 129, 3069-3087.
8) Uittenbogaard, R., Sebastian, C., Vijverberg, J., Boom, B., and Gavrila, D. M. (2019). Privacy protection in street-view panoramas using depth and multi-view imagery. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10581-10590.

필요하다는 점과 단일 객체를 처리한다는 점을 한계로 가진다. Alkobi, Noa et al.[9]은 여러 개의 생성자(Generative)와 판별자(Discriminator)를 피라미드 형식으로 구성하여 이미지 복원 품질을 높이고자 하였다. 그러나 복원이 마스킹 영역 바깥에서부터 독립적인 단계로 진행되어 폐색된 객체나 복잡한 배경에 대해선 적절한 복원이 이루어지지 않았다. Shetty, R. R. et al.[10]은 객체를 확실히 제거하고자 복원한 이미지를 객체 탐지 모델에 전달하여 특성을 추출하고 이를 손실 함수에 반영하였다. 이는 영상에서 지우고자 하는 객체의 클래스가 유일한 경우 효과적이었으나, 지우고자 하는 객체와 같은 클래스의 다른 객체가 영상에 존재하는 경우, 객체 제거가 제대로 이루어지지 않은 양 피드백하였다. 따라서 해당 모델은 영상에 동일 클래스의 다중 객체가 존재하는 경우를 고려하지 못했다. Darapaneni, Narayana, et al.[11]은 YOLOv3를 이용해 객체 검출을 수행하고 맥락 주의 구조(Contextual Attention)[12]를 탑재한 SRGAN(Super Resolution GAN)을 이용해 이미지 인페인팅을 시도하였다. 해당 모델은 작은 객체에 대해서 높은 복원력을 보여주었지만, 이미지 복원 시 전체 이미지의 색감 변화를 일으킨다는 단점이 있다.

Deepfillv2[13]은 Free-form 마스크에 대한 이미지 복원 품질을 높이고자

9) Alkobi, N., Shaham, T. R., and Michaeli, T. (2023). Internal Diverse Image Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 648-658.

10) Shetty, R. R., Fritz, M., and Schiele, B. (2018). Adversarial scene editing: Automatic object removal from weak supervision. Advances in Neural Information Processing Systems, 31.

11) Darapaneni, N., Kherde, V., Rao, K., Nikam, D., Katdare, S., Shukla, A., ... and Paduri, A. R. (2022). Contextual Attention Mechanism, SRGAN Based Inpainting System for Eliminating Interruptions from Images. arXiv preprint arXiv:2204.02591.

12) Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, 5505-5514.

13) Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2019). Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF international conference on computer vision, 4471-4480.

그룹별 합성곱(Gated Convolution) 연산을 활용하였다. 이를 통해 마스킹 영역의 모양이 다양해져도 이미지를 자연스럽게 복원하였다. AOT-GAN[14]은 수용 영역(Receptive Field)을 넓힌 합성곱 블록과 판별자의 마스킹 영역 예측을 제안하여 다양한 마스크에 대한 이미지 복원 능력을 높이고자 하였다. 그러나 두 모델은 마스크의 크기가 커지면 배경의 패턴을 제대로 파악하지 못해 복원한 이미지의 품질이 낮아지는 경우가 발생하였다. 이에 MAT[15]과 LaMa[16]는 기존 연산을 배경 정보를 잘 활용할 수 있는 연산으로 교체하여 크기가 큰 마스크에 대한 복원 성능을 높이고자 하였다. MAT은 기존 합성곱 블록을 트랜스포머(Transformer)로 교체하였고, LaMa는 기존 합성곱 연산을 푸리에 합성곱(Fourier Convolution) 연산으로 교체하였다. 두 모델 모두 마스킹 영역을 제외한 배경 객체를 확실히 유지하며 고해상도의 이미지로 복원하지만, 연산 과정에서 이용할 수 있는 배경 정보가 부족해지면 제거하고자 하는 객체를 남기는 경우가 발생하였다. IA[17]는 세그멘테이션 모델과 기존 인페인팅 모델을 결합하여 제거하고자 하는 객체를 배경과 분리하여 인페인팅 과정을 진행하였다. 해당 모델은 결합한 세그멘테이션 모델로 인해 객체 단위 마스크를 생성해낼 수 있으며 이미지에서 객체만을 지우는 것에 특화되었다. 그러나 초기 위치 지정이 필요하여 단일 객체를 대상으로 한다는 점과 긴 추론 시간을 한계로 갖는다.

---

14) Zeng, Y., Fu, J., Chao, H., and Guo, B. (2022). Aggregated contextual transformations for high-resolution image inpainting. IEEE Transactions on Visualization and Computer Graphics.

15) Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., and Jia, J. (2022). Mat: Mask-aware transformer for large hole image inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10758-10768.

16) Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., ... and Lempitsky, V. (2022). Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2149-2159.

17) Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., and Chen, Z. (2023). Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790.

최근에는 노이즈 생성 과정을 역연산이 가능한 수식으로 정의하여 이미지를 생성하는 Diffusion이 이미지 생성 분야에서 주목받고 있다. 이에 Diffusion을 이용한 인페인팅 또한 활발히 연구되고 있다.

RePaint[18]는 DDPM (Denoising Diffusion Probabilistic Model)을 활용하여 극단적으로 크기가 큰 마스크나 활용할 배경 정보가 부족한 마스크에 대한 인페인팅 성능을 향상시키고자 하였다. Uni-paint[19]는 사전 학습된 Stable Diffusion을 활용하여 지우고자 하는 객체를 텍스트로 제공하는 멀티 모달(Multi Modal) 인페인팅을 가능하게 하였다. 그러나 Diffusion이 생성적 적대 신경망보다 더 다양하고 고품질의 이미지를 생성하는 강점은 분명하지만, 모델의 크기가 충분히 커야 한다는 단점 또한 분명하다. 따라서 Diffusion은 모델이 거대해짐에 따라 모델의 추론 시간이 증가하고 모델의 이식성이 감소하므로 실용적인 측면에서 제약이 존재한다.

3. 모델 경량화

데이터가 축적된 여러 분야에서 딥러닝 모델이 개발, 이용됨에 따라 모델의 성능뿐 아니라 모델의 크기 및 추론 시간이 중요한 지표로 활용되고 있다. 이에 따라 기존 모델에서 보조적인 연산을 제외하거나 중심 연산을 교체하는 등의 방법을 통해 모델을 경량화하는 연구가 진행되었다.

Xception[3]은 Inception 모델을 경량화하고자 합성곱 블록의 순차적인 연산 과정을 개선한 깊이별 분리 합성곱 연산을 제안하였다. 깊이별 분리 합성곱 연산은 기존 3차원 연산을 두 방향의 차원으로 먼저 계산한 후, 나머지 한

18) Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11461-11471.

19) Yang, S., Chen, X., and Liao, J. (2023). Uni-paint: A Unified Framework for Multimodal Image Inpainting with Pretrained Diffusion Model. In Proceedings of the 31st ACM International Conference on Multimedia, 3190-3199.

차원을 독립적으로 계산한다. 이를 통해 필터 연산의 영향력은 유지하되 연산에 이용되는 파라미터의 수를 획기적으로 줄여 Inception 모델을 경량화하였다. NSB-GAN[20]은 BigGAN에 웨이블릿 변환(Wavelet Transform)을 적용하여 적은 컴퓨팅 자원에서 고해상도 이미지를 생성하고자 하였다. 해당 연구에서는 웨이블릿 변환이 이미지의 저주파와 고주파를 분할 하는 데 효과적임을 보였지만, 컴퓨팅 자원에 따라 이미지의 복원 정도가 제한된다는 점을 한계로 꼽았다.

MobileStyleGAN[21]은 깊이별 분리 합성곱 연산과 웨이블릿 변환을 결합한 합성곱 블록을 이용하여 고품질의 이미지를 생성하되 모델의 크기와 계산 복잡도를 줄인 StyleGAN을 제시하였다. 이와 유사하게 Mobile-SRGAN[22]과 WMR-DepthwiseNet[23] 또한 깊이별 분리 합성곱과 이산 웨이블릿 변환(Discrete Wavelet Transform), 역 이산 웨이블릿 변환(Inverse Discrete Wavelet Transform) 함수로 구성한 합성곱 블록을 이용하여 모델의 추론 시간을 효과적으로 줄일 수 있음을 실험으로 입증하였다.

---

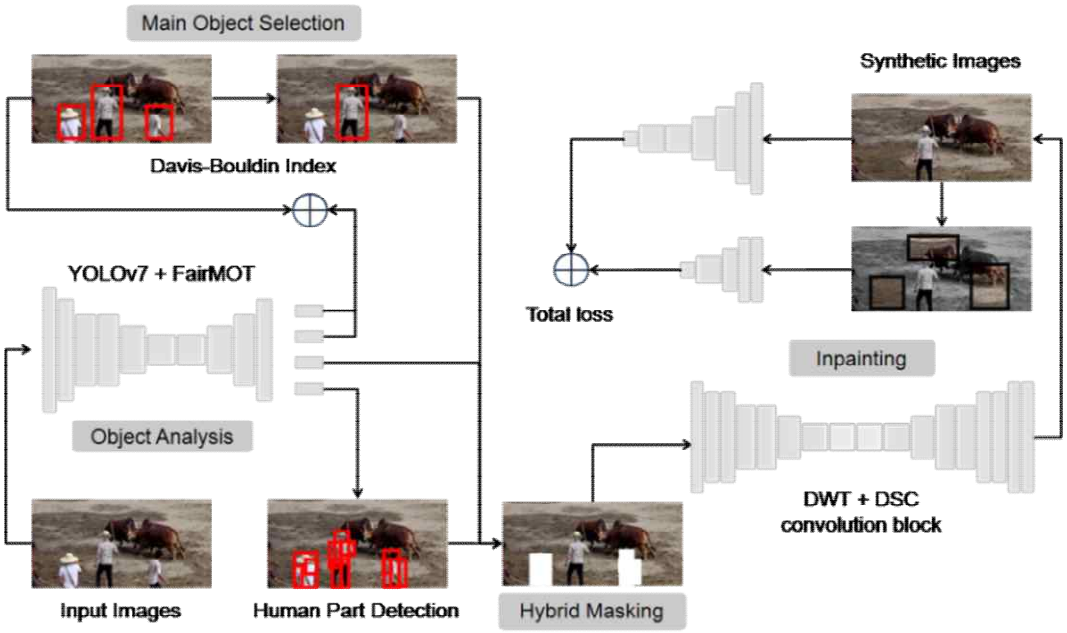
20) Han, S., Srivastava, A., Hurwitz, C., Sattigeri, P., and Cox, D. D. (2020). not-so-BigGAN: Generating High-Fidelity Images on Small Compute with Wavelet-based Super-Resolution. arXiv preprint arXiv:2009.04433.

21) Belousov, S. (2021). Mobilestylegan: A lightweight convolutional neural network for high-fidelity image synthesis. arXiv preprint arXiv:2104.04767.

22) Vasileiou, C., Smith, J., Thiagarajan, S., Nigh, M., Makris, Y., and Torlak, M. (2022, October). Efficient CNN-based super resolution algorithms for mmWave mobile radar imaging. In 2022 IEEE International Conference on Image Processing (ICIP), 3803-3807.

23) Monday, H. N., Li, J., Nneji, G. U., Hossin, M. A., Nahar, S., Jackson, J., and Chikwendu, I. A. (2022). WMR-DepthwiseNet: A Wavelet Multi-Resolution Depthwise Separable Convolutional Neural Network for COVID-19 Diagnosis. Diagnostics, 12(3), 765.

Ⅲ. 다중 객체 제거 시스템



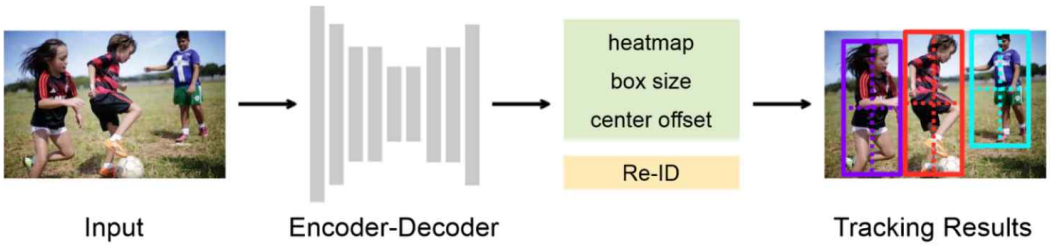
[그림 3-1] 제안하는 객체 제거 시스템

제안하는 다중 객체 제거 시스템은 [그림 3-1]처럼 객체 분석, 중심 객체 선정, 이미지 복원 단계로 구성하였다. 객체 분석 단계는 영상 내 객체를 분석하는 단계로, 전체 시스템의 추론 시간과 중심 객체 추적을 위해 기존과 달리 사람과 신체 부위를 탐지하여 추적한다. 이어지는 중심 객체 선정 단계는 영상에서 불필요한 객체를 제거하는 단계로 초기 프레임에서 중심 객체를 선정하는 과정과 불필요한 객체를 마스킹하는 과정을 포함하고 있다. 이를 위해 중심 객체를 정의하고, 다비스-볼드윈 지수를 도입하여 중심 객체를 특정한다. 또한, 탐지된 신체 부위의 경계 상자를 이용해 하이브리드

마스크를 생성한다. 마지막으로 이미지 복원 단계는 마스킹 영역을 복원하는 단계로, 맥락 주의 구조(Contextual Attention)의 구성을 변경하고 개량한 합성곱 블록을 인페인팅 모델에 적용해 마스킹 영역을 복원한다.

1. 객체 분석

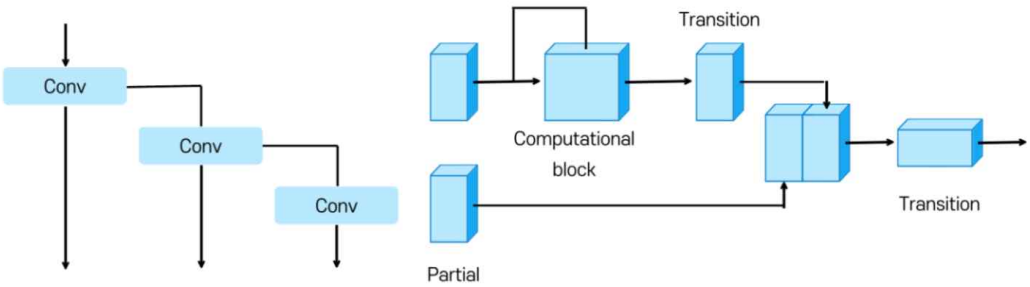
객체 분석 단계는 영상에 존재하는 모든 객체를 식별하는 단계이다. 이는 이어지는 중심 객체 선정 단계에서 중심 객체와 불필요한 객체를 선별하는 데 이용할 객체 정보를 얻기 위함이다. 해당 단계에서는 다중 객체 추적 모델을 이용하여 객체를 제거할 필요가 있는 영상으로 이미지뿐 아니라 비디오 또한 고려하였다. 또한, 다중 객체 추적 모델 중, FPS 대비 높은 다중 객체 추적 정확도를 보이는 FairMOT[24]를 이용하여 전체 시스템의 추론 시간을 줄이고자 하였다. 더불어 FairMOT의 특성(Feature) 추출을 위한 백본 네트워크(Backbone Network)를 YOLOv7으로 교체하여 사람과 신체 부위 탐지 속도를 높이고자 하였다.



[그림 3-2] 객체 추적 모델

24) FairMOT: On the Fairness of Detection and Re-Identification in Multi-Object Tracking. Available: https://github.com/ifzhang/FairMOT

FairMOT는 앵커-프리 객체 탐지 모델을 활용한 객체 추적 모델로, [그림 3-2]의 과정을 따른다. 먼저, 입력 이미지를 인코더-디코더(Encoder-Decoder), 즉 앵커-프리 객체 탐지 모델에 전달하여 이미지의 특성을 추출한다. 이때 앵커-프리 객체 탐지 모델은 객체 추적 성능에 직접적인 영향을 미치므로 사전 학습된 모델을 이용하는 것이 일반적이다. 기존 구조에서는 객체 탐지 성능을 고려하여 백본 네트워크(Backbone Network)로 ResNet34 모델에 깊은 층 집계(Deep Layer Aggregation)를 적용해 이용하였다. 그러나 본 논문에서는 YOLO 모델의 실시간 대비 객체 탐지 성능 향상 및 앵커-프리 객체 탐지 모델로의 구조적 진화를 고려하여 YOLOv7[25]을 백본 네트워크로 이용하였다.



[그림 3-3] 연결 방법에 따른 연산 흐름
(왼) 연결 기반 구조, (오른) 깊이와 너비를 고려한 복합 스케일링 구조[23]

YOLOv7은 YOLOv5에서 모델의 추론 시간은 유지하되 탐지 성능을 향상시킨 모델이다. 이는 데이터 세트의 환경 다양성을 고려한 증강 기법과 활성화 함수 개선 등이 적용된 결과이다. 모델 개선에 사용된 기법이

25) Wang, C. Y., Bochkovskiy, A., and Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7464-7475.

YOLOv4와 유사하여 YOLOv7과 YOLOv4는 구조적인 유사성을 갖는 듯 보이나, 이 두 모델에는 확연한 차이점이 존재한다. YOLOv4는 [그림 3-3]의 왼쪽과 같이 학습 시 그래디언트(Gradient) 갱신 흐름을 유지하고자 연산의 깊이를 늘리는 방식으로 합성곱 연산을 추가했다. 이와 달리 YOLOv7은 합성곱 연산을 연산 깊이 추가와 너비 추가로 나누어 병렬적으로 구성하였다. 이러한 변화를 통해 YOLOv7은 학습 효율이 높으며, 다른 YOLO(You Only Look Once) 모델보다 FPS 대비 객체 탐지 성능이 높다는 이점을 가진다.

백본 네트워크를 이용한 특성 추출이 완료되면 추출된 특성은 객체 탐지 분기(Object Detection branch)와 정체성 임베딩 분기(Identity Embedding branch)로 나누어 처리한다. 객체 탐지 분기에서는 회귀 문제(Regression task)를 수행하여 객체의 경계 상자를 예측한다. 제안한 시스템에서 객체 탐지 분기의 결과물은 사람 객체와 신체 부위이다. 정체성 임베딩 분기에서는 입력 특성을 객체 간 유사성 계산을 위한 임베딩 벡터로 변환한다. 변환된 벡터(X, Y)는 객체별로 [수식 3-1]의 코사인 유사도(Cosine Similarity)를 계산하는 데 이용하며, 이 값을 통해 객체 간 유사성을 확인하여 ID를 할당하고 추적한다.

$$Cosine\ \ Similarity = \frac{X \cdot Y}{\|X\|_2 \cdot \|Y\|_2}$$

[수식 3-1] 코사인 유사도

2. 중심 객체 선정

중심 객체 선정 단계는 중심 객체를 선정하여 불필요한 객체를 파악하고 이에 대한 마스킹을 진행하는 단계이다. 이를 위해 중심 객체를 정의하고

중심 객체 선정 기준을 다비스-볼드윈 지수에 적용하여 중심 객체를 하나로 특정하였다. 이후, 탐지된 신체 부위의 경계 상자 마스킹을 통해 불필요한 객체에 대한 마스킹을 진행하였다. 이를 통해 중심 객체 선정 과정에서 발생하는 연산 지연을 최소화하고 세그멘테이션 마스크와 유사한 하이브리드 마스크를 생성하였다.

본 논문에서는 영상의 중심 콘텐츠와 타인의 개인정보 객체를 분리하고자 중심 콘텐츠를 콘텐츠의 흐름을 이끄는 사람 객체로 정의하며 중심 객체라 지칭한다. 더불어 중심 객체로 선정되지 않은 나머지 사람 객체는 콘텐츠에 필요하지 않은 사람 객체이므로 불필요한 객체로 지칭한다.

영상에서 중심 객체는 영상을 통틀어 가장 비중 있는 객체이며 객체의 크기, 위치 등으로 파악한다. 이에 따라 가장 큰 객체, 중앙에 존재하는 객체, 촬영기기에 가장 가까운 객체, 고정된 위치에 있는 객체, 영상에서 사라지지 않는 객체, 가장 빈번하게 나타나는 객체로 중심 객체를 설명할 수 있다. 이 중, 중심 객체가 영상에서 가장 큰 객체이거나 중앙에 존재하는 객체, 촬영기기와 가장 가까운 객체일 경우에는 한 장의 프레임만으로 중심 객체를 파악할 수 있다. 이와 달리 중심 객체가 고정된 위치에 있는 객체이거나 영상에서 사라지지 않는 객체, 가장 빈번하게 나타나는 객체일 경우에는 전체 동영상을 분석하여 중심 객체를 식별해야 한다. 이 경우, 중심 객체를 파악하는 것은 전체 동영상을 해석해야 하므로 추가적인 연산과 시간이 필요하다.

본 논문에서 중심 객체는 영상의 초기 프레임부터 추적해야 하는 객체로, 이를 충족시키기 위해 첫 프레임에서 중심 객체가 결정되어야 한다. 이에 첫 프레임을 기준으로 크기가 크고 중앙에 가까운 사람 객체를 중심 객체로 정의하였다. 촬영기기와의 거리를 비교하는 방법은 영상 깊이 분석(Depth Estimation)이 필요한 방법으로, 중심 객체 선정 과정에서 발생하는 연산 지

연을 최소화하고자 중심 객체 선정 기준에서 제외하였다.

정의한 중심 객체 선정 기준을 토대로 [수식 3-2]의 다비스-볼드윈 지수(DB)를 계산하여 가장 낮은 값을 갖는 객체를 중심 객체로 선정하였다. 다비스-볼드윈 지수는 클러스터의 중심(A) 간 거리(M) 대비 클러스터 내 데이터(X)와의 거리(S)를 계산하여 클러스터 내부 거리와 다른 클러스터와의 거리를 비교하는 지수이다. 이때 X_i와 A_i는 각각 i번째 클러스터의 내부 데이터와 클러스터 중심을 의미한다.

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^q \right\}^{\frac{1}{q}}, \quad M_{ij} = \left\{ \sum_{k=1}^{N} |a_{ki} - a_{kj}|^p \right\}^{\frac{1}{p}},$$
$$DB = \frac{1}{N} \sum_{i=1}^{N} \left[\max_{i \neq j} \left\{ \frac{S_i + S_j}{M_{ij}} \right\} \right]$$

[수식 3-2] 다비스-볼드윈 지수

본 논문에서는 클러스터 간 거리 비교를 위해 영상 중앙과 객체의 경계 상자에 클러스터링을 적용하였다. 객체의 중앙과 객체의 경계 상자 꼭짓점은 각기 다른 클러스터의 중심으로 가정하고, 영상 중앙을 객체 중앙 클러스터의 내부 데이터로 가정하여 다비스-볼드윈 지수를 계산하였다. 해당 가정을 통해 객체 중앙과 경계 상자의 각 꼭짓점까지의 거리가 동일해져 클러스터 간 거리(M)는 일정한 값을 가지게 된다. 또한, 외부 클러스터의 내부 거리 또한 존재하지 않기 때문에 클러스터 내부 거리(S)도 일정한 값을 가진다. 이때 다비스-볼드윈 지수는 객체의 크기 대비 영상 중앙까지의 거리가 되어, 해당 지수를 통해 중심 객체를 선정하면 객체의 크기와 영상 중앙까지의 거리에 비슷한 가중치를 부여할 수 있다. 따라서 다비스-볼드윈 지수가 낮을수록 객체 중심과 중앙까지의 거리가 짧고 객체의 크기가 크다는

것을 의미한다.

$$I(Z=z)=\frac{P(Z=z)}{P(Z\neq z)}=\left(\frac{y_z}{x_z}\right)\left(\frac{y_o}{x_o}\right)^{-1}$$

[수식 3-3] 중심 객체 가능성

중심 객체 가능성(I_z)은 중심 객체(z)일 확률과 중심 객체로 선정되지 않았을 때(o)의 확률을 비교한 것으로, 객체 크기(x)와 영상 중앙까지의 거리(y)를 고려할 때, [수식 3-3]으로 표현될 수 있다. 이로 인해 중심 객체는 객체의 크기는 커야 하며 영상 중앙까지의 거리는 짧아야 한다.

$$I(Z=z)=\frac{x_o y_z}{x_z y_o}=1 \quad (a) \qquad y_o=n y_z,\ x_z=n x_o \quad (b)$$

$$W_z H_z = n\,W_o H_o \quad (c)$$

$$\sqrt{\frac{W_z^2+H_z^2}{4}}=\sqrt{\frac{(n\,W_o)^2+(nH_o)^2}{4}}=\sqrt{\frac{(n^2\,W_o^2+n^2H_o^2-H_o^2)+H_o^2}{4}} \quad (d)$$

$$\frac{(d)}{(c)}=\frac{\sqrt{n^2\,W_o^2+(n^2-1)H_o^2}}{n\,W_o} \quad (e)$$

$$\lim_{n\to\infty}\frac{(d)}{(c)}=\sqrt{2} \quad (f)$$

[수식 3-4] 객체 크기 측정에 따른 중심 객체 가능성 비교

또한, [수식 3-4]는 영상 중앙까지의 거리가 늘어남에 따라 중심 객체 가능성을 동일한 값으로 유지하고자 할 때(a, b) 필요한 객체 크기를 면적과 비교한 식이다. 객체의 경계 상자 면적을 이용했을 때(c)는 영상 중앙까지 거리가 늘어난 배수만큼 객체 면적의 너비 또는 높이가 증가하면 되었다. 이와 달리 객체의 경계 상자 꼭짓점까지의 거리를 이용했을 때(d)는 영상

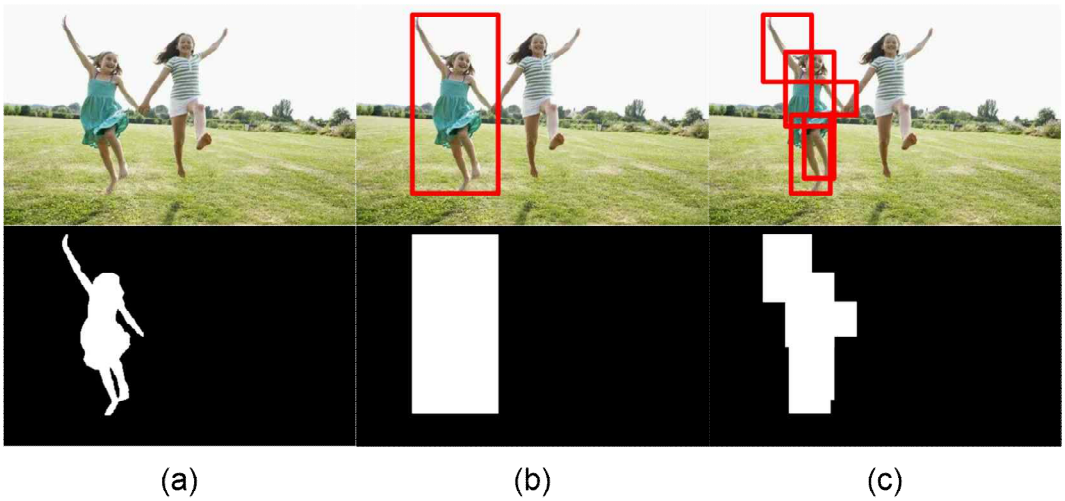
중앙까지의 거리가 늘어난 배수보다 더 큰 값을 가져야 한다(f). 이는 객체 면적 값을 그대로 이용했을 때보다 객체의 경계 상자 꼭짓점까지의 거리를 이용했을 때 중심 객체 선정에 대한 객체 면적의 영향력이 감소했음을 의미한다. 따라서 객체의 경계 상자 꼭짓점까지의 거리를 이용하여 중심 객체를 선정하면 객체의 크기와 영상 중앙까지의 거리에 비슷한 가중치를 부여할 수 있다.

다비스-볼드윈 지수를 활용한 중심 객체 선정 결과는 [그림 3-4]에서 확인할 수 있다. 이렇게 선정한 중심 객체는 ID를 기록하여 다음 프레임의 마스킹 대상에서 제외하였다.

[그림 3-4] 다비스-볼드윈 지수로 선정된 중심 객체 이미지

마스킹 과정은 중심 객체로 선정되지 않은 불필요한 객체의 신체 부위 경계 상자에 따라 마스킹을 진행한다. [그림 3-5]는 제안한 하이브리드 마스크를 경계 상자 마스크, 세그멘테이션 마스크와 함께 나타낸 그림이다. 하이브리드 마스크는 [그림 3-5]의 3열처럼 탐지된 신체 부위 경계 상자로 마스킹

한 이미지를 의미한다. 경계 상자 마스크는 탐지된 사람 객체의 자세에 따라 너비와 크기가 크게 달라질 수 있다. 특히 [그림 3-5]처럼 팔을 뻗고 있는 자세에서 마스크의 너비와 크기가 달라지며, 이 경우 인페인팅 모델에 배경 정보를 충분히 전달하지 못한다. 이와 달리 사람의 신체 부위를 탐지하여 마스킹한 하이브리드 마스크는 자세에 따른 마스크 크기 변동이 적으며, 경계 상자 마스크와 비교해 세그멘테이션 마스크와의 차이가 작다. 따라서 이러한 마스킹 방법을 이용하면 세그멘테이션 과정을 거치지 않고 세그멘테이션 마스크에 준하는 마스크를 생성할 수 있다.



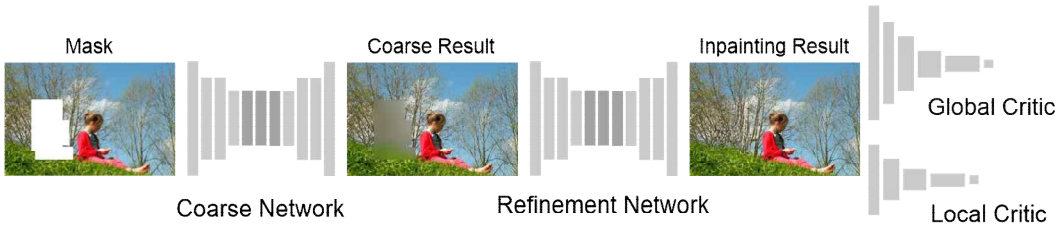
[그림 3-5] 마스크 종류 비교
(a) 세그멘테이션 마스크, (b) 경계 상자 마스크, (c) 하이브리드 마스크

3. 인페인팅

인페인팅 단계는 이미지 인페인팅을 통해 이미지를 복원하는 단계이다. 앞선 중심 객체 선정 단계에서 마스킹된 이미지를 인페인팅을 통해 복원하

여 이미지의 시각적인 품질을 높이는 것을 목표로 한다. 이를 위해 맥락 주의 구조의 생성적 적대 신경망을 활용했으며, 이미지의 패턴 분석 및 상세한 복원을 위해 일차 네트워크(Coarse Network)의 비중을 늘렸다. 또한, 일차 네트워크에서 합성곱 연산 시 하르(Haar) 함수를 이용한 이산 웨이블릿 변환을 적용하여 이미지의 구조적 정보를 깊이 있게 확인하였다. 또한, 판별자의 중간 특성을 비교하여 이미지 복원 품질을 높이고자 하였다.

맥락 주의 구조는 일차 네트워크를 통해 전반적인 복원을 완료하는 1차 복원, 세부 조정 네트워크(Refinement Network)를 통해 이미지 품질을 높이는 2차 복원으로 진행된다. 이렇게 생성된 이미지는 전체 이미지와 복원된 영역만을 고려하는 두 개의 판별자에 의해 평가되어 전체 모델에 피드백된다. 해당 구조는 [그림 3-6]을 통해 확인할 수 있다.

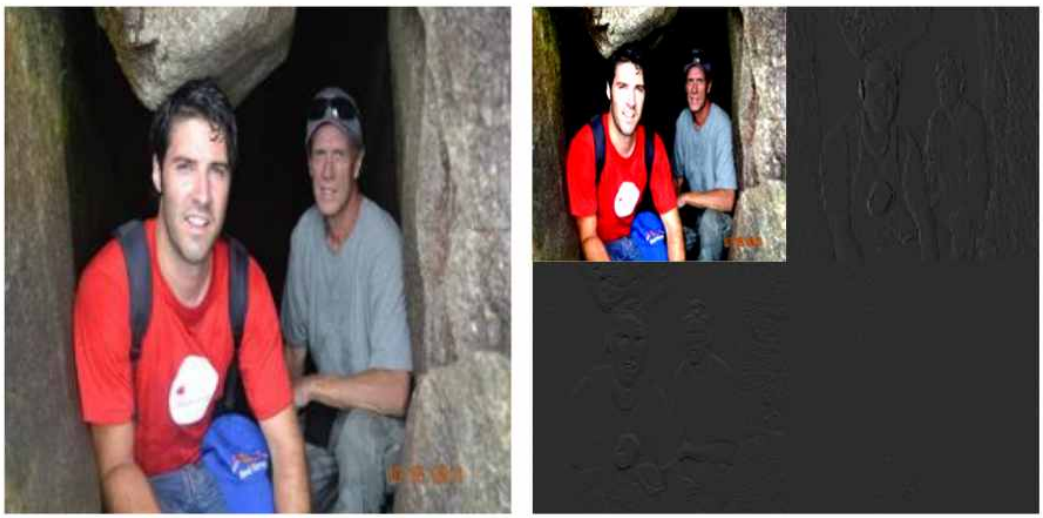


[그림 3-6] 맥락 주의 구조[12)]

그러나 이 구조는 세부 조정 네트워크를 일차 네트워크와 같은 크기로 설정하여 불필요한 연산을 반복하고, 합성곱 블록이 영상의 전반적인 패턴에만 주목한다는 한계점이 있다. 이에 본 논문에서 제안한 인페인팅 모델은 세부 조정 네트워크를 줄이고 일차 네트워크의 중간 합성곱 블록의 수를 늘렸으며, 웨이블릿 변환의 일종인 이산 웨이블릿 변환, 역 이산 웨이블릿 변환(Inverse Discrete Wavelet Transform)을 합성곱 연산 전후에 이용하여

이미지 복원에 초점을 맞췄다.

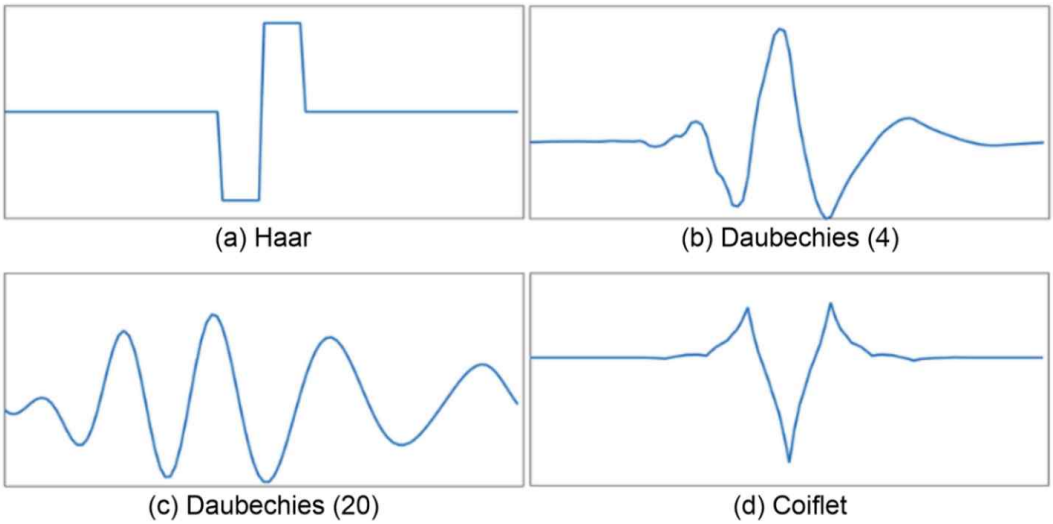
웨이블릿 변환은 입력 이미지를 주파수 대역으로 변환하여 이미지를 다양한 주파수 대역으로 투영한다. 이를 통해 이미지의 픽셀 간 패턴을 분석할 때보다 더 많은 구조적 정보를 얻을 수 있어 이미지 복원에 용이하다. 또한, 웨이블릿 변환을 합성곱 연산과 이용하면 편향(Bias)과 같은 고정적인 연산을 학습 파라미터와 분리하여 효율적으로 학습 파라미터를 갱신할 수 있다. 이는 모델의 추론 시간을 줄이는 효과 또한 가져온다.



[그림 3-7] 이산 웨이블릿 변환 결과 예시
(왼) 입력 이미지, (오른) 이산 웨이블릿 변환 결과

이산 웨이블릿 변환은 변환 결과를 이산형 분포로 가정하는 웨이블릿 변환으로 [그림 3-7]처럼 입력 이미지를 동일한 크기의 주파수 채널 4개로 분리한다. 데이터를 이산형 분포로 투영하기 때문에 다른 웨이블릿 변환과 비교해 정보 손실이 크다는 단점이 있으나, 역 이산 웨이블릿 변환과 함께 이

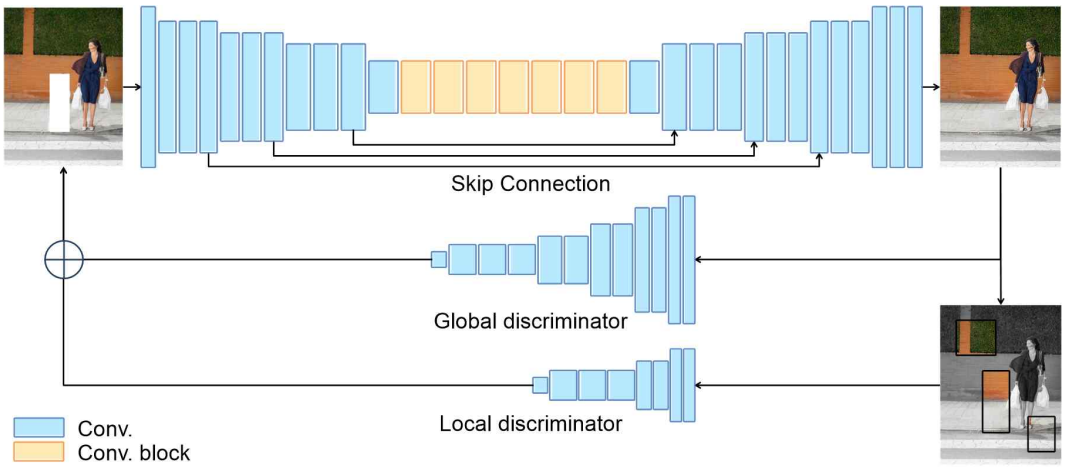
용하면 잃어버리는 정보의 양을 줄일 수 있다. 또한, 이산 웨이블릿 변환은 변환이 간단하여 연산 시간이 짧다는 이점이 존재한다.



[그림 3-8] 웨이블릿 변환 필터 함수

웨이블릿 변환에 이용하는 웨이블릿 필터 함수는 추출하고자 하는 형태에 따라 달라진다. [그림 3-8]은 다양한 웨이블릿 필터 함수 중, 하르 함수, 파라미터를 달리 한 도비시(Daubechies) 함수 2종, 코이플렛(Coiflet) 함수를 시각화한 것이다. 하르 함수는 가장 단순한 형태의 웨이블릿 필터 함수로 빠른 연산이 가능하지만, 다른 함수와 비교해 손실되는 정보량이 많다. 코이플렛 함수는 독특한 형태의 주파수 감지를 위해 사용되는 웨이블릿 필터 함수이다. 도비시 함수는 웨이블릿 필터 함수를 일반화한 것으로 파라미터를 통해 형태를 조절할 수 있으며, 하르, 코이플렛 함수를 비롯한 대부분의 웨이블릿 필터 함수의 형태를 추정할 수 있다는 것이 특징이다. 웨이블릿 필터 함수를 통한 주파수 감지는 복잡한 형태일수록 정교한 주파수 분해가 가

능하며 이 때문에 도비시 함수를 이용하는 것이 일반적이다. 그러나 이산 웨이블릿 변환, 역 이산 웨이블릿 변환에 하르 함수를 이용하면 웨이블릿 변환을 곱셈 연산 없이 효율적으로 구현 가능하여, 제안한 인페인팅 모델에선 하르 함수를 기저 함수로 하는 이산 웨이블릿 변환 함수를 이용하였다.



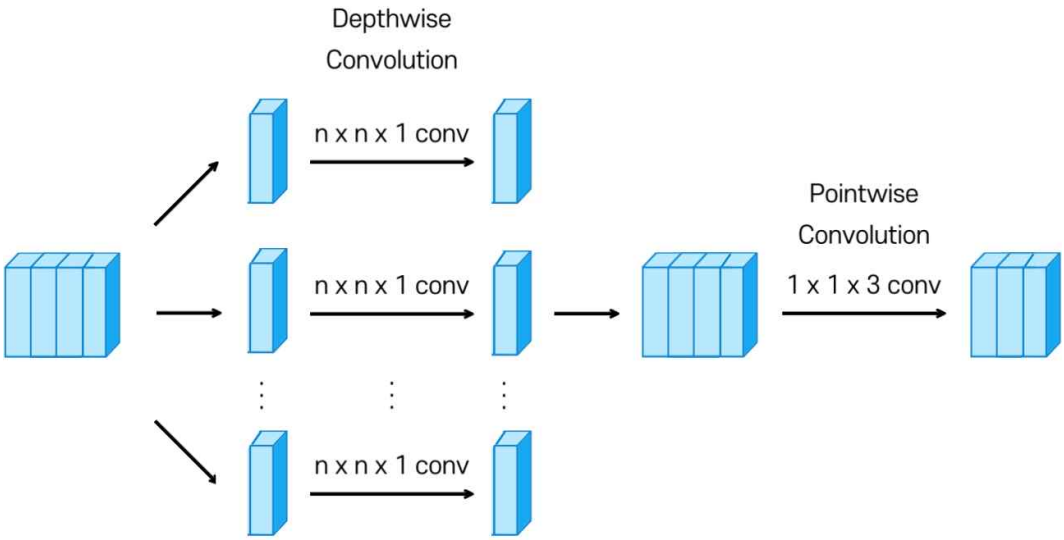
[그림 3-9] 이미지 인페인팅 모델

[그림 3-9]는 맥락 주의 구조를 변형해 구성한 인페인팅 모델의 그림이다. 생성자(Generator)에서 다운 샘플링(Downsampling)을 위해 스트라이드(Stride)를 조절한 합성곱 연산을 통해 특성 맵의 크기를 축소한다. 중심에 존재하는 합성곱 블록은 마스킹 영역을 복원하기 위해 외부 패턴을 참고한다. 이후, 스킵 커넥션(Skip Connection)과 업샘플링(Upsampling)을 통해 이미지의 전반적인 특징을 유지하며 자연스러운 이미지를 생성한다. 판별자(Discriminator)는 맥락 주의 구조처럼 이미지의 전반적인 복원 품질과 마스킹 영역의 복원 품질을 확인하고자 2개의 판별자로 구성하였다. 2종의 판별자는 각각 전역적(Global), 지역적(Local) 판별자로 일정 크기의 패치를 결과

로 도출하여 이미지의 사실성을 특성 단위로 확인한다. 또한, 학습 과정에서 그래디언트 피드백 흐름을 통일시키고자 두 판별자의 패치를 결합하여 입력으로 주어진 전체 이미지의 진위를 판별한다. 이를 통해 전체 이미지의 생성 품질과 마스크 영역의 생성 품질을 생성자에 피드백한다.

4. 모델 경량화

기존 인페인팅 모델은 이미지 복원 성능에 집중하여 모델의 추론 시간이 길다. 본 논문에서는 이를 개선하고자 기존 합성곱 연산을 깊이별 분리 합성곱 연산으로 교체하고, 합성곱 블록을 인페인팅에 특화된 형태로 개량하여 모델의 추론 시간을 줄이고자 하였다.



[그림 3-10] 깊이별 분리 합성곱[3)]

깊이별 분리 합성곱 연산은 합성곱 연산에 이용되는 파라미터 수를 줄여

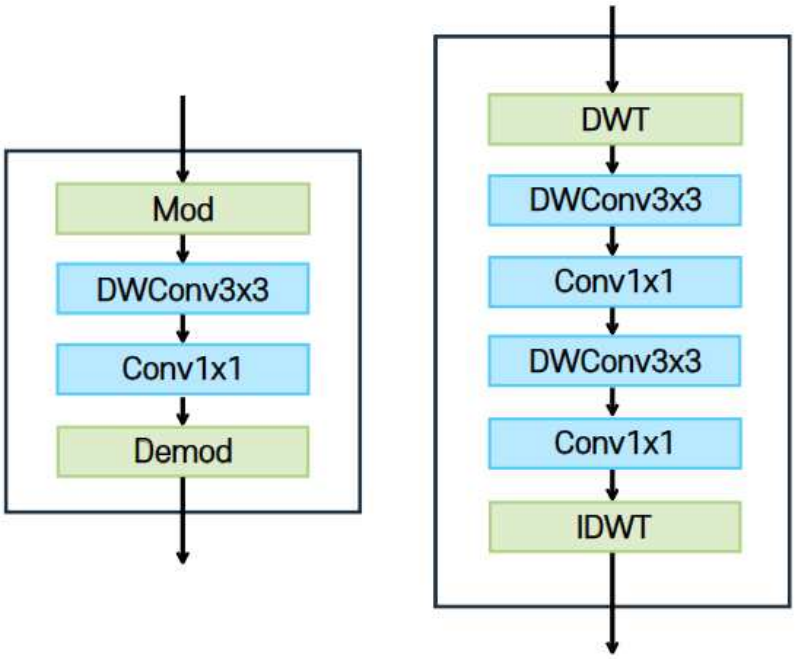
연산 효율성이 높도록 개량된 연산이다. 이로 인해 깊이별 분리 합성곱 연산을 이용하면 전체 모델의 복잡성과 메모리 사용량을 감소시킬 뿐 아니라 과적합 또한 방지할 수 있다. 깊이별 분리 합성곱은 기존 합성곱 연산이 하나의 합성곱 연산으로 끝나는 데 반해 [그림 3-10]에서 볼 수 있듯 깊이별 합성곱(Depthwise Convolution) 연산과 점별 합성곱(Pointwise Convolution) 연산의 두 단계로 진행한다.

깊이별 합성곱 연산은 입력 특성 맵의 각 채널에 합성곱 연산을 진행한다. 입력 채널마다 합성곱 연산을 수행하기 때문에 채널별로 독립적인 필터가 사용되며, 이로 인해 연산 시 입력 채널 수와 동일한 필터 수가 필요하다. 이 연산은 그룹의 수가 입력 특성 맵의 채널 크기와 같은 그룹별 합성곱(Grouped Convolution) 연산으로 구성할 수 있으므로 그룹별 합성곱 연산의 이점인 연산량 감소와 그룹 간 독립성을 기대할 수 있다. 또한, 입력 특성 맵의 공간적인 정보를 추출하고 매핑(Mapping)하여 출력 특성 맵의 크기를 조절할 수 있다.

점별 합성곱 연산은 필터 크기가 1인 합성곱 연산으로, 입력 특성 맵에 각 채널에서 동일한 위치에 있는 값들의 선형 결합으로 연산이 진행한다. 채널 간 연산에 집중함으로써 깊이별 합성곱 연산에서 찾아내지 못한 새로운 특징을 찾아낼 수 있다는 이점이 있다. 또한, 계산에 이용되는 필터 수를 조정하여 출력 특성 맵의 채널 수를 조절할 수 있다.

합성곱 블록은 MobileStyleGAN[20]의 블록과 유사하나, 웨이블릿 변환을 통한 특성 추출의 효과를 극대화하기 위해 중간 합성곱 연산을 2번으로 구성하였다. [그림 3-11]의 왼쪽 부분은 MobileStyleGAN에서 모델 경량화를 위해 사용된 합성곱 블록의 형태로 깊이별 분리 합성곱 연산 앞뒤로 정규화(Normalization)와 스케일링(Scaling) 과정을 거치는 것이 특징이다. 이는 입력 특성 맵의 노이즈와 변동성을 줄인다는 이점이 있다. 그러나 이러한 정

규화 과정은 통계량 계산이 선행되어야 하므로 입력 이미지의 크기에 따라 연산 시간이 기하급수적으로 증가할 위험이 있다. 이에 정규화 대신 웨이블릿 변환을 합성곱 연산 이전에 진행하여 이미지를 다양한 주파수 대역으로 재구성하는 방식을 이용하였다. 또한, 잦은 이산 웨이블릿 변환, 역 이산 웨이블릿 변환 및 깊이별 분리 합성곱 연산으로 잃어버린 연산량을 일부 유지하고자 깊이별 분리 합성곱 연산을 연달아 진행하도록 구성하였다.



[그림 3-11] 합성곱 블록 비교
(왼) DSC 연산을 이용한 합성곱 블록[20], (오른) 제안한 합성곱 블록

따라서 합성곱 블록의 연산을 정리하면 다음과 같다. 먼저 웨이블릿 변환 함수를 이용해 연산의 입력 데이터를 주파수 특성 차원으로 투영한 후, 깊이별 분리 합성곱 연산을 진행한다. 이를 통해 얻은 특성 맵은 웨이블릿 역

변환 함수를 통해 기존의 차원으로 투영되어 재구성한다. 이러한 분해 및 재구성 과정을 반복하면서 저주파 대역은 입력 이미지의 일반적인 특징을 추출하고 고주파 대역은 세부적인 정보를 추출한다. 이렇게 추출된 저주파 및 고주파 대역의 정보를 결합하여 이미지를 자연스럽게 복원한다.

5. 손실 함수

제안한 인페인팅 모델은 맥락 주의 구조를 활용하여 이미지의 복원 성능을 높이고자 하였고, 이를 위해 전반적인 품질과 마스킹 영역의 품질을 파악하여 생성자에 피드백하도록 하였다. 이를 위해 손실 함수는 생성적 적대 신경망의 손실 함수 구성을 기준으로 이미지의 전반적인 품질 및 마스킹 영역의 품질을 확인하는 손실 함수로 구성되었다.

지각 손실 함수(Perceptual Loss)는 VGG16과 같이 높은 분류 성능을 가진 사전 학습된 거대 분류기의 중간 특성 맵을 활용하여 원본 이미지와 생성한 이미지의 특성 차이를 계산하는 방식이다. 많은 수의 클래스를 분류하기 때문에 중간 특성 맵에서 이미지의 일반적인 특성을 추출할 수 있다. 이 때문에 인페인팅에서는 복원한 이미지가 마스킹 영역을 제외한 이미지 전반의 특성을 유지하는지 평가하는 데 이용한다. 그러나 지각 손실은 복원 영역의 변화에 둔감하여 일정 손실 값을 유지하기 때문에 복원 성능을 정확하게 측정하는 데 적절하지 않다. 이에 전역적(Global), 지역적(Local) 판별자의 패치를 이용하여 이미지의 전반적인 품질 및 마스킹 영역의 품질을 확인하는 손실 함수를 구성하였다.

$$L_p = E_{x \sim P_X}\{D_{Global}(x)\} - E_{z \sim P_Z}\{D_{Global}(G(z))\} + E_{x \sim P_X}\{D_{Local}(x)\} - E_{z \sim P_Z}\{D_{Local}(G(z))\}$$

[수식 3-5] 개량한 지각 손실 함수

[수식 3-5]는 전역적 판별자와 지역적 판별자의 패치를 통한 손실 함수 계산이다. 두 판별자는 원본 이미지와 생성한 이미지에 대하여 일정 크기의 패치를 결과로 도출하여 이미지의 사실성을 특성 단위로 계산한다. 이러한 구성은 복원한 이미지의 품질을 정확히 피드백할 뿐 아니라 생성자와 판별자 간의 학습 불안정 해소를 돕는다. 해당 손실 함수는 복원한 이미지가 실제 이미지와 가까울수록 작은 값을 갖는다.

$$L_{adv} = E_x\left[\log\{D_{Total}(D_{Global}(x) + D_{Local}(x))\}\right]$$

[수식 3-6] 전체 판별자로 개량한 적대적 손실

일반적인 맥락 주의 구조는 판별자의 손실 함수를 전역적, 지역적으로 각각 피드백하는 것에 그친다. 그러나 제안한 모델은 학습 과정에서 판별자의 그래디언트 피드백 흐름을 통일시키고자 전체(Total) 판별자를 도입하였다. 따라서 [수식 3-6]처럼 전체 판별자는 전역적, 지역적 판별자의 패치를 결합하여 전체 이미지의 진위를 확인한다. 이렇듯 판별자를 여러 개로 구성하면서 생성자와의 학습 속도를 맞추기 위해 각 판별자는 각 판별자 학습 시 다른 판별자를 고정하여 각각 피드백하였다.

$$L = \alpha L_p + \beta L_{adv}$$

[수식 3-7] 전체 손실 함수

[수식 3-7]은 전체 손실 함수로 판별자에 대한 손실 함수만으로 구성하여 모델의 학습 안정성을 높이고자 하였다. 이를 통해 모델 학습이 안정적으로 수렴하여 기울기 소실(Vanishing Gradient)이나 모드 붕괴(Mode Collapse)와 같은 문제를 방지할 수 있다. 또한, 학습 데이터 세트의 과적합(Over-fitting)을 예방하고 생성자와 판별자 간의 균형 학습을 유도하고자 전체 손실 함수를 간단하게 구성하였다.

VI. 실험 및 결과

1. 데이터 세트

[그림 4-1] BGVP 데이터 세트[26]

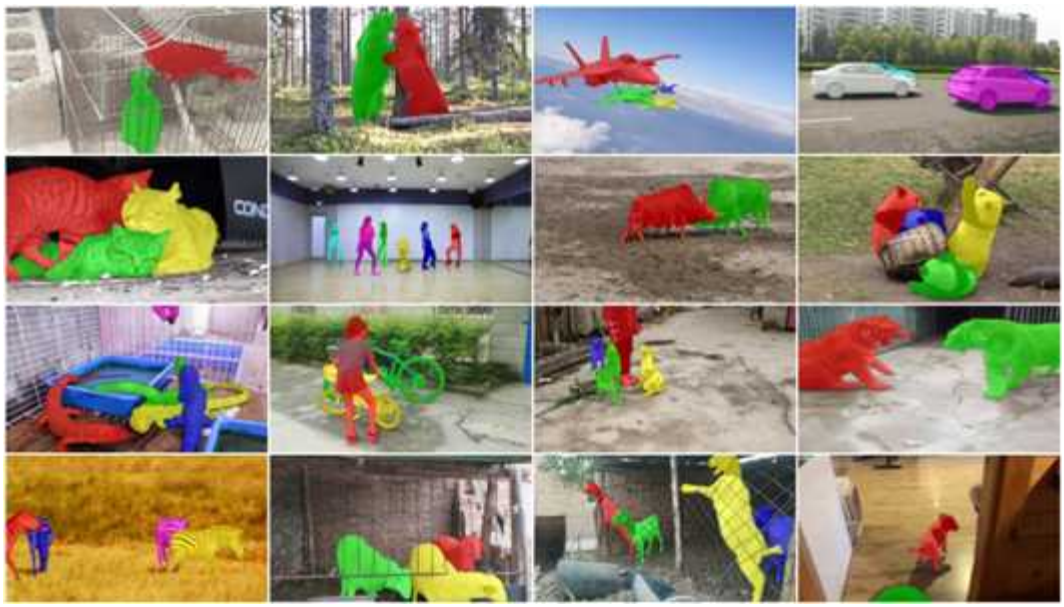
본 논문에서 중심 객체와 지워야 할 객체의 클래스는 편의상 사람 클래스로 제한하였다. 이에 BGVP(Background Vulnerable Pedestrian Dataset)[26] 데이터와 WiderPerson[27] 데이터와 같이 다량의 사람이 존재하는 이미지 데

26) Sharma, D., Hade, T., and Tian, Q. (2022). Comparison Of Deep Object Detectors On A New Vulnerable Pedestrian Dataset. arXiv preprint arXiv:2212.06218.

27) Zhang, S., Xie, Y., Wan, J., Xia, H., Li, S. Z., and Guo, G. (2019). Widerperson: A diverse dataset for dense pedestrian detection in the wild. IEEE Transactions on Multimedia, 22(2), 380-393.

이터를 학습 데이터로 선정하였다.

또한, OVIS(Occluded Video Instance Segmentation)[28] 데이터를 테스트 데이터로 이용하였다. 해당 데이터 세트는 역동적인 움직임을 가진 객체와 폐색 객체가 주를 이루는 것이 특징이다. [그림 4-2]에서 볼 수 있듯이 25개의 다양한 카테고리로 영상 내 객체의 클래스를 분류하고 있으나, 본 논문에서는 사람 클래스의 이미지를 대상으로 하였다.



[그림 4-2] OVIS 데이터 세트[28]

실험에는 Colab이 이용되었으며 이에 대한 컴퓨터의 사양은 [표 4-1]과 같다. 또한, 실험에 이용한 데이터는 학습 데이터(Train data), 검증에 이용한 데이터(Validation data), 지표 계산 등의 실험에 이용한 데이터(Test

28) Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., ... and Bai, S. (2022). Occluded video instance segmentation: A benchmark. International Journal of Computer Vision, 130(8), 2022-2039.

data)로 나누어지며 8:1:1의 비율로 나누어 이용하였다. 모델 학습에 이용한 파라미터는 배치(Batch Size)가 16, 평균 에포크(Epoch)가 500이다. 또한, 학습은 SGD와 Adam 중, 더 작은 손실 값을 보인 Adam을 통해 최적화되었다.

[표 4-1] 시스템 환경

Computing Environment	Workstation
Processor	Intel(R) Xeon(R) CPU @ 2.20GHz
Memory	83.5GB
Operating System	Ubuntu 18.04.3 LTS
Graphics Card	NVIDIA A100-SXM4-40GB

2. 평가 지표

인페인팅 실험에서 복원한 이미지의 시각적 품질은 다음의 3가지 평가 지표를 통해 평가하였다.

$$PSNR = 10 \times \log_{10}\left(\frac{H \times W \times MAX_I^2}{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}[R(i,j) - G(i,j)]}\right)$$

[수식 4-1] PSNR

[수식 4-1]은 PSNR(Peak Signal-to-Noise Ratio)로 생성한 이미지의 노이즈 비율을 계산하는 지표이다. 원본 이미지(R)와 생성한 이미지(G) 사이의 픽셀 간 차이의 평균을 최대 차이(MAX_I)와 비교하여 영상을 복원하는 과정에서 생긴 선명도 손실을 집중적으로 측정한다.

$$SSIM = \frac{(2\mu_R\mu_G + C_1)(2\sigma_{RG} + C_2)}{(\mu_R^2 + \mu_G^2 + C_1)(\sigma_R^2 + \sigma_G^2 + C_2)}$$

[수식 4-2] SSIM

[수식 4-2]는 SSIM(Structural Similarity Index Measure)으로, 데이터의 구조적인 유사성을 평가하는 지표이다. PSNR과 마찬가지로 원본 이미지(R)와 생성한 이미지(G) 사이의 차이에 집중하는데 각 집단의 분포를 고려하고자 각 픽셀의 평균(μ), 표준편차(σ), 공분산(σ_{RG})과 같은 통계량을 이용한다.

$$LPIPS = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left| w^l \odot \left(\hat{y}^l_{R,hw} - \hat{y}^l_{G,hw} \right) \right|_2^2$$

[수식 4-3] LPIPS

LPIPS(Learned Perceptual Image Patch Similarity)는 AlexNet, VGG, SqueezeNet 모델의 중간 특성 맵을 활용한다. [수식 4-3]처럼 각 모델에서 원본 이미지(R)와 생성한 이미지(G)의 차이를 가중치(w^l)에 대한 평균의 합으로 계산한다. LPIPS는 좋은 성능을 내는 거대 분류기의 중간층(l)을 통해 이미지의 특성을 추출하고 이를 기반으로 이미지 간의 차이를 계산한다는 점에서 FID(Fréchet Inception Distance)와 유사하다. 그러나 FID는 Inception 모델을, LPIPS는 AlexNet, VGG, SqueezeNet 모델을 이용한다는 차이가 있다. 이 차이로 인해 LPIPS는 인간이 구조적으로 이미지를 비교하는 정도를 수치로 나타내는 평가 지표로 활용된다.

3. 객체 제거

제안한 시스템을 통한 객체를 제거하기에 앞서 중심 다비스-볼드윈 지수를 통한 객체 선정 실험과 경계 상자 마스크 및 하이브리드 마스크에 대한 이미지 복원 비교 실험, 전체 판별자 손실 함수의 적절성 실험을 진행하였다.

[그림 4-3] 중심 객체 선정 방식에 따른 결과 비교
(a) 원본 이미지, (b) 중앙 객체 선택, (c) 가장 큰 객체 선택, (d) 다비스-볼드윈 지수에 의한 선택

[그림 4-3]은 중심 객체 선정 결과를 선정 기준별로 정리한 것이다. 1열과 2열의 이미지에서 중앙 객체는 영상의 중심 객체로 볼 소지가 다분하나, 가장 큰 객체 선택 방식에선 탐지된 객체의 크기가 작아 중앙 객체를 중심 객체로 선정하지 않았다. 3열과 4열은 중앙에 가깝지만 크기가 큰 객체를 중심 객체로 보았을 때이며, 이 경우 중앙 객체 선택 방식과 가장 큰 객체 선

택 방식에서는 각각 다른 객체를 중심 객체로 선정하였다. 특히 가장 큰 객체를 중심 객체로 선택하는 방식은 3열처럼 영상에 잠시 모습을 드러낸 객체를 중심 객체로 선택하여 실제 중심 객체와 큰 차이를 보였다. 이와 비교해 다비스-볼드윈 지수에 의한 선택 방식은 중앙 객체 선택 방식과 가장 큰 객체 선택 방식을 적절히 융합하여 영상의 중심 객체를 적절히 선택하는 양상을 보였다.

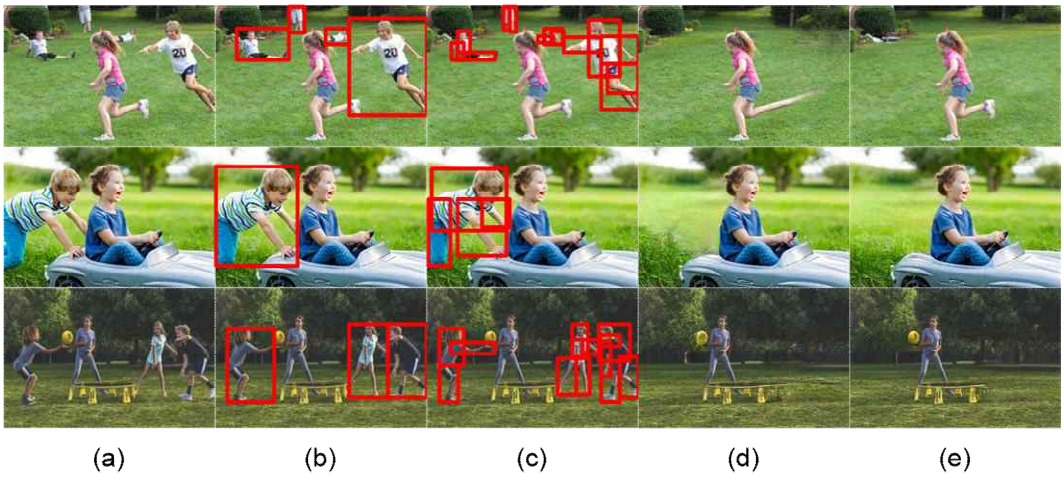
[표 4-2] 선정 방식에 따른 중심 객체 예측 성능

Main object selection	Nearest to the center	Bigest area	Davies-Bouldin Index
accuracy ↑	0.4	0.56	0.87

[표 4-2]는 제안한 중심 객체 선정 방식에 따른 중심 객체 예측 성능을 정리한 표이다. 영상에 대한 중심 객체 데이터는 학습 데이터에서 임의로 선별하는 과정과 동료 평가를 통해 구성하였다. 이를 통해 기존의 중앙에 가까운 객체와 면적이 큰 객체는 각각 이용하였을 때 정확도가 0.4, 0.56으로 영상의 중심 객체를 예측하기 어렵다. 그러나 다비스-볼드윈 지수를 통해 객체의 크기와 영상 중앙까지의 거리를 고루 고려했을 때는 정확도가 0.87로 측정되었다. 이는 제안한 중심 객체 선정 방식이 기존의 방식보다 효과적으로 중심 객체를 예측함을 의미한다.

[그림 4-4]는 마스크의 종류에 따른 인페인팅 결과를 정리한 것이다. 1행의 경우, (b)를 보면 중심 객체와 오른쪽에 존재하는 사람 객체의 전체 경계 상자가 겹쳐 (d)에서 중심 객체의 신발이 사라지고 다리가 늘어나게 복원된 것을 볼 수 있다. 이와 달리 (c)를 보면 하이브리드 마스크의 마스킹 영역이 겹치지 않아 (e)에서 중심 객체가 제대로 보존된 것을 볼 수 있다. 이와 비슷하게 2행은 하이브리드 마스크를 이용했을 때, 배경 풀의 정보를 더 받아 자연스럽게 이미지를 복원하였다. 더불어 3행 또한, 하이브리드 마스크를 이

용했을 때 배경 잔디와 나무의 정보를 제공 받아 하단에 있는 노란 객체와 노란 공이 원래의 형태를 유지하며 복원되었다.



[그림 4-4] 마스크 종류에 따른 인페인팅 결과 비교
(a) 원본 이미지, (b) 경계 상자, (c) 제안한 하이브리드 마스크, (d) 경계 상자 마스크를 통해 복원한 이미지, (e) 제안한 하이브리드 마스크를 통해 복원한 이미지

[표 4-3] 마스크 종류에 따른 인페인팅 성능

Mask type	PSNR ↑	SSIM ↑	LPIPS ↓
bounding box	29.93	0.81	0.083
hybrid	32.013	0.89	0.068

[표 4-3]은 마스크 종류에 따른 인페인팅 결과를 평가 지표로 비교한 것이다. 경계 상자 마스크를 이용했을 때는 PSNR, SSIM, LPIPS가 각각 29.93, 0.81, 0.083였으나, 제안한 하이브리드 마스크를 이용했을 때는 32.013, 0.89, 0.068으로 PSNR과 SSIM의 수치는 상승했으며, LPIPS의 수치는 감소했다. 이는 제안한 하이브리드 마스크가 전체 사람 객체에 대한 경계 상자 마스크보다 주변 정보를 활용하기 쉬워 이미지가 더 자연스럽게 복원되었기

때문이다.

[표 4-4] 전체 판별자의 손실 함수 구성에 따른 성능

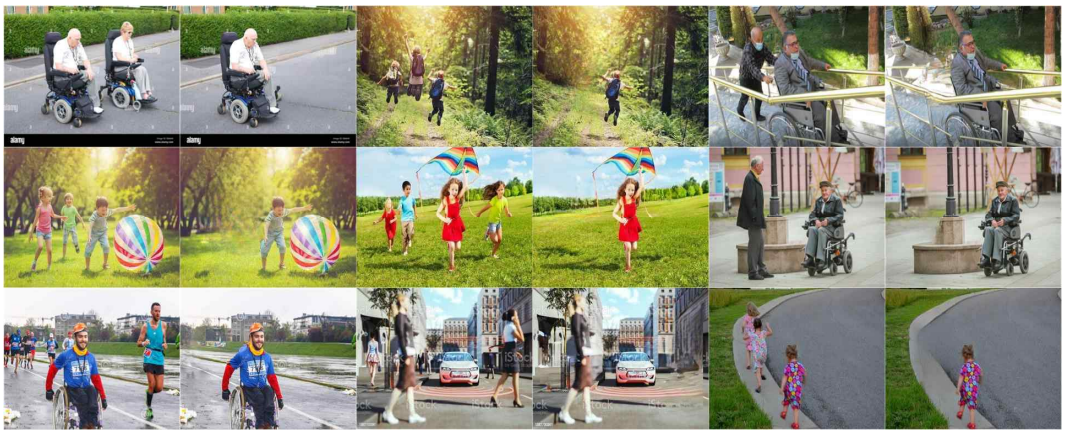
Discriminator loss	PSNR ↑	SSIM ↑	LPIPS ↓
Global, Local	31.9	0.87	0.079
Global, Local, Total	32.013	0.892	0.067

[표 4-4]는 제안한 시스템에서 구성한 전체 판별자의 손실 함수 유무에 따라 복원한 이미지의 품질 차이를 정리한 표이다. 기존 전역적, 지역적 품질을 확인하는 판별자만으로 손실 함수를 구성할 때는 PSNR, SSIM, LPIPS가 각각 31.9, 0.87, 0.079로 측정되었다. 판별자 간 학습 흐름을 제어하는 전체 판별자가 추가되면서 이미지 생성 품질 지표는 각각 32.013, 0.892, 0.067로 이전과 비교해 소폭 상승하고, 감소했음을 알 수 있다. 이는 전체 판별자의 추가가 인페인팅 모델의 학습 불균형을 일부 해소하여 결과 이미지의 품질이 더 향상되었음을 의미한다.

[그림 4-5]에서 [그림 4-7]은 데이터에 따른 객체 제거 시스템의 결과물로, [그림 4-5]는 학습 데이터 세트, [그림 4-6]은 Pedestrian[29], Penn-Fudan Database[30] 데이터 세트, [그림 4-7]은 Places365[31] 데이터 세트에 대한 결과물이다. 또한, 홀수 열은 입력 이미지를, 짝수 열은 복원한 이미지를 나타낸다. [그림 4-5]를 주목하면, 입력 이미지의 전체적인 색감과 구성을 유지하고 있으며, 배경의 패턴을 연장하여 이미지를 자연스럽게 생성함을 알 수 있다. 더불어 단일 객체뿐 아니라 다중 객체도 제거함을 시각

29) Wang, L., Shi, J., Song, G., and Shen, I. F. (2007). Object detection combining recognition and segmentation. In Asian conference on computer vision, 189-199.
30) Karthika, N. J., and Chandran, S. (2020). Addressing the false positives in pedestrian detection. In Electronic Systems and Intelligent Computing: Proceedings of ESIC 2020, 1083-1092.
31) López-Cifuentes, A., Escudero-Vinolo, M., Bescós, J., and García-Martín, Á. (2020). Semantic-aware scene recognition. Pattern Recognition, 102, 107256.

적으로 확인할 수 있다. 그러나 1행 6열의 그림과 3행 6열의 그림처럼 영상에서 나타나지 않는 독특한 패턴이 객체에 가려졌을 때 배경 패턴으로 이를 파악할 수 없어 다소 우그러진 형태로 복원되었다.



[그림 4-5] 제안한 시스템의 인페인팅 결과

[그림 4-6] Pedestrian, Penn-Fudan Database 데이터 세트 인페인팅 결과

[그림 4-7] Places365 데이터 세트 인페인팅 결과

4. 모델 경량화

본 논문에서는 모델의 중간 합성곱 블록을 개량하여 인페인팅 모델을 경량화하고자 하였다. 이와 더불어 합성곱 블록의 개수를 조절하여 생성 이미지 품질 대비 전체 모델의 추론 시간을 줄이는 실험을 진행하였다. 이때 이전과

같은 이미지 복원 정도를 유지하고자 PSNR을 주 평가 지표로 설정하여 FPS를 관측하는 모델 경량화 실험을 진행하였다. 해당 경량화 실험은 PSNR이 큰 차이로 관측되는 지점에서 중단하였다.

[표 4-5] 중심 합성곱 블록 개수에 따른 성능

Num. of conv. block	PSNR ↑	Params ↓	FPS ↑
9 conv.	33.094	925,156	5.7
6 conv.	31.98.	906,057	7.3
5 conv.	32.6	899,024	8.2
4 conv.	32.013	893,991	8.9
3.5 conv.	27.21	-	-

[표 4-5]는 제안한 합성곱 블록의 수에 따른 이미지 생성 품질과 FPS를 정리한 표이다. 제안한 합성곱 블록의 수가 줄어듦에 따라 중간 합성곱 블록에서 학습해야 하는 파라미터 수가 점차 줄어들고 있으며, 전체 시스템에 대한 추론 시간이 줄어 FPS가 증가하는 것을 볼 수 있다. 합성곱 블록이 초기 개수의 반 정도일 때 다른 모델보다 높은 PSNR 수치를 보였다. 이는 불필요한 연산이 줄면서 파라미터의 연산 효율이 높아졌기 때문이다. 합성곱 블록이 3.5개일 때 PSNR 수치가 급감함을 알 수 있다. 이는 합성곱 블록이 3.5개일 때 생성한 이미지와 블록의 개수가 4개 이상인 모델이 생성한 이미지가 이미지를 복원하는 측면에서 큰 성능 차이를 보이는 것으로 해석할 수 있다.

5. 모델 비교

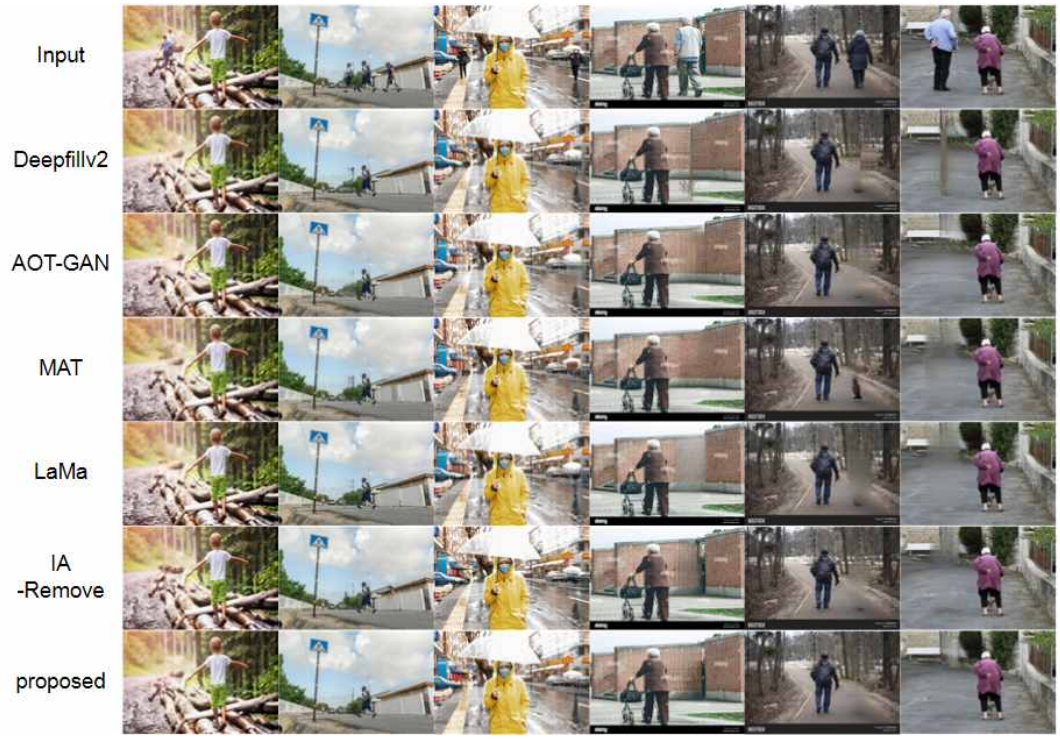
본 논문에서 제안한 시스템은 입력 이미지에서 불필요한 사람 객체를 탐지하고 마스킹한다. 그러나 기존의 인페인팅 모델은 이미 마스킹 된 이미지를 입력으로 받기 때문에 이를 이용하기 위해선 마스킹 된 이미지가 필요하

다. 이에 제안한 시스템의 객체 분석 단계와 중심 개체 선정 단계를 거쳐 마스킹한 이미지를 저장하여 다른 모델과 비교하는 데 활용하였다. [그림 4-8]은 모델 비교를 위해 생성한 마스크 그림이다.

[그림 4-8] 마스크 생성 결과

[그림 4-9]는 지우고자 하는 객체가 작은 크기의 다중 객체일 경우와 단조로운 배경에서 비교적 크기가 큰 단일 객체일 경우의 모델에 따른 결과 이미지이다. 1열처럼 지우고자 하는 객체가 단일 객체일 경우에는 참고해야 하는 배경 정보가 적으므로 비교한 모든 모델이 자연스러운 이미지를 생성하였다. 2열과 3열 또한, 지우고자 하는 객체가 작아 다중 객체임에도 비교적 자연스러운 이미지가 생성되었다 그러나 4열에서 6열처럼 지우고자 하는 객체의 크기가 커지자 Deepfillv2[13]과 MAT[15]가 이미지를 제대로 복원하지 못하는 경우가 생겼다. Deepfillv2의 경우, 제거하고자 하는 객체가 커질수록 배경 정보를 제대로 활용하지 못해 6열처럼 눈에 띄게 우그러지는 경향을

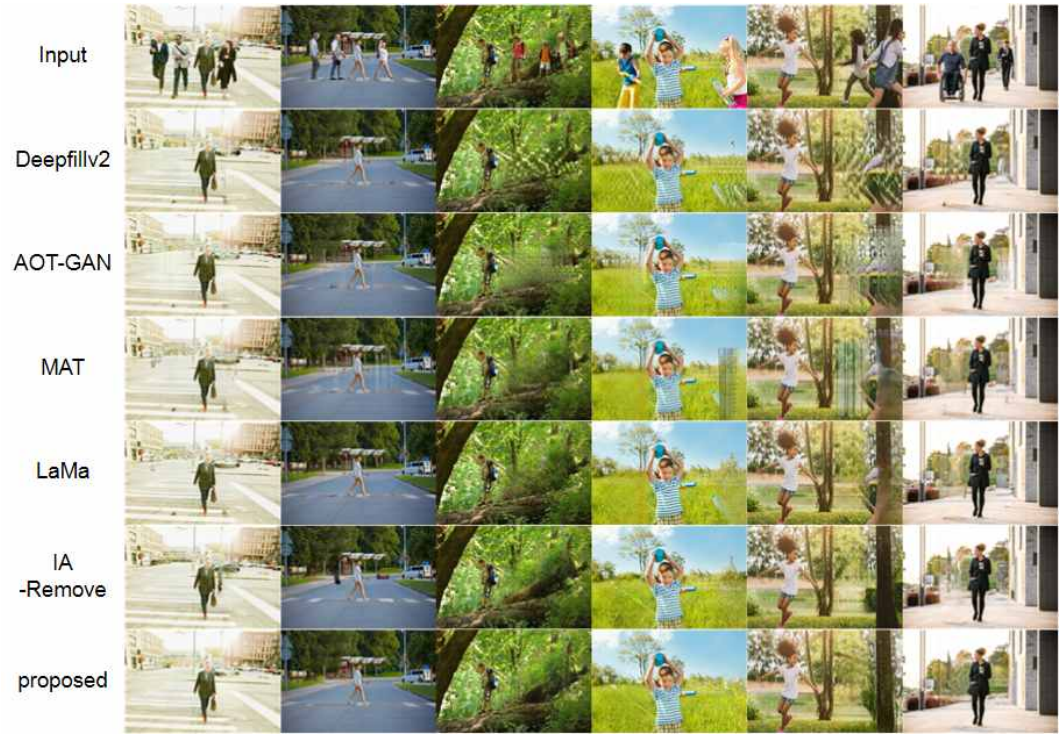
보였다. MAT의 경우, 4열에서 6열의 그림을 보면 마스킹 영역은 섬세하게 복원했지만, 특히 6열의 그림에서 주변 배경과 동화되지 못하는 경우가 있었다. 또한, LaMa의 경우, 4열의 이미지에서 주변 패턴을 제대로 확인하지 못해 번진 이미지로 복원된 것을 볼 수 있다.



[그림 4-9] 단순한 환경에 대한 인페인팅 예시

[그림 4-10]은 사람이 아닌 다른 객체가 마스킹 영역에 걸쳐 있거나 배경에 다양한 패턴이 존재하는 복잡한 배경에 대한 인페인팅 결과 이미지이다. 1행은 원본 이미지이며, 2행은 Deepfillv2, 3행은 AOT-GAN, 4행은 MAT, 5행은 LaMa, 6행은 IA 모델이며, 마지막으로 7행은 제안한 시스템의 결과 이미지이다.1열과 2열, 3열은 지우고자 하는 객체가 연이어 존재하여 배경

정보가 제한되는 경우다. 1열과 2열은 이미지 전반에 배경 색이 존재하는 이미지의 결과로 대부분 눈에 띄는 우그러짐 없이 이미지를 복원하였다. 그에 반해 3열은 배경이 되는 나무에 대한 정보가 부족해 많은 모델에서 번짐(Blurring) 현상이 나타났다.



[그림 4-10] 복잡한 환경에 대한 인페인팅 예시

[그림 4-10]의 4열부터 6열은 지우고자 하는 객체 배경에 이미지에서 잘 나타나지 않은 패턴이 존재하는 경우이다. Deepfillv2과 AOT-GAN[14]의 경우, 마스킹 영역의 주변과 내부의 패턴을 반복하여 마스킹 영역을 복원하면서 결과 이미지에 의도하지 않은 패턴이 생겼다. MAT와 LaMa[16]는 객체 크기에 대한 영향은 받지 않았으며, 마스킹 영역 주변의 전반적인 패턴을

이용해 이미지를 복원하였다. 그러나 복원된 이미지의 시각적인 품질을 높이고자 3열, 4열처럼 결과 이미지에 지우고자 하는 객체 일부를 남기는 경우가 존재했다.

IA[17]는 모델 내부에서 세그멘테이션을 일부 진행함에 따라 지우고자 하는 객체 뒤에 존재하는 사물의 형태를 대부분 남겨 원본 이미지와의 환경을 대부분 보존했다. 제안한 시스템은 마스킹 영역의 패턴을 일부 연장하여 이미지를 복원하는 형태로, 이미지를 자연스럽게 복원했다. 또한, 마스킹 영역 주변에서 확실한 색 차이를 보여주는 경우, 과감히 색을 유지하고 끝맺음으로써 번짐 현상을 최소화한 결과를 보여주었다.

[그림 4-11] 인페인팅 실패 사례 비교

[그림 4-11]은 제안한 시스템의 이미지 복원 실패 사례를 다른 모델의 결과 이미지와 비교한 그림이다. 제안한 시스템은 이미지를 자연스럽게 복원하기 위해서 배경의 패턴을 연장하는 경향을 보였다. 이 때문에 지우고자 하는 객체의 경계 상자가 이미지 구석에 존재하는 경우, 배경 정보를 충분히 이용하지 못해, 주변 객체가 상하 또는 좌우로 늘려 복원되었다. 더불어 5열이나 6열처럼 주위 배경에 일관성이 없는 경우, 지우고자 하는 영역에 존재하는 작은 객체를 큰 객체처럼 복원하는 등 원본의 형태와 다르게 이미지를 복원하기도 하였다.

[표 4-6] 인페인팅 성능 비교

	PSNR ↑	SSIM ↑	LPIPS ↓	FPS ↑
Deepfillv2[13]	**32.886**	0.559	0.09	**12.16**
AOT-GAN[14]	32.45	0.786	0.11	6
MAT[15]	31.16	0.835	0.068	6.31
LaMa[16]	31.908	0.83	0.074	6.35
IA-Remove[17]	31.51	**0.891**	**0.067**	2.4
proposed	32.013	0.819	**0.067**	8.9

[표 4-6]은 기존 모델과 제안한 시스템의 인페인팅 성능을 정리한 것이다. SSIM의 경우, IA 모델이 가장 높게 측정되었지만, 제안한 시스템의 SSIM 값과 많은 차이를 보이지 않는다. LPIPS는 IA와 제안한 시스템에서 가장 낮은 값을 보였다. PSNR의 경우, 모델별 인페인팅 결과 이미지에서 번짐 현상을 가장 많이 보인 Deepfillv2의 측정값이 가장 높게 관측되었다. 이는 PSNR이 픽셀 간 차이를 계산하면서 인간의 지각적 판단과 정확히 비례하지 않기 때문이다[32]. 그러나 제안한 시스템의 PSNR이 Deepfillv2의 PSNR과 큰 차이를 보이지 않으므로, 결과적으로 제안한 시스템이 상당히 자연스

32) Suresh, K., and Sakthi, U. (2018). Robust multi-thresholding in noisy grayscale images using Otsu's function and harmony search optimization algorithm. In Advances in Electronics, Communication and Computing: ETAEERE-2016, 491-499.

럽게 이미지를 복원하고 있음을 알 수 있다. 나아가 종합적인 이미지 복원 성능을 FPS와 비교할 때 제안한 시스템은 FPS 대비 가장 높은 복원 성능을 보였다. [표 4-7]은 비교 실험에 이용한 모델별로 불필요한 다중 객체 제거 시스템에 대한 기능을 정리한 것이다.

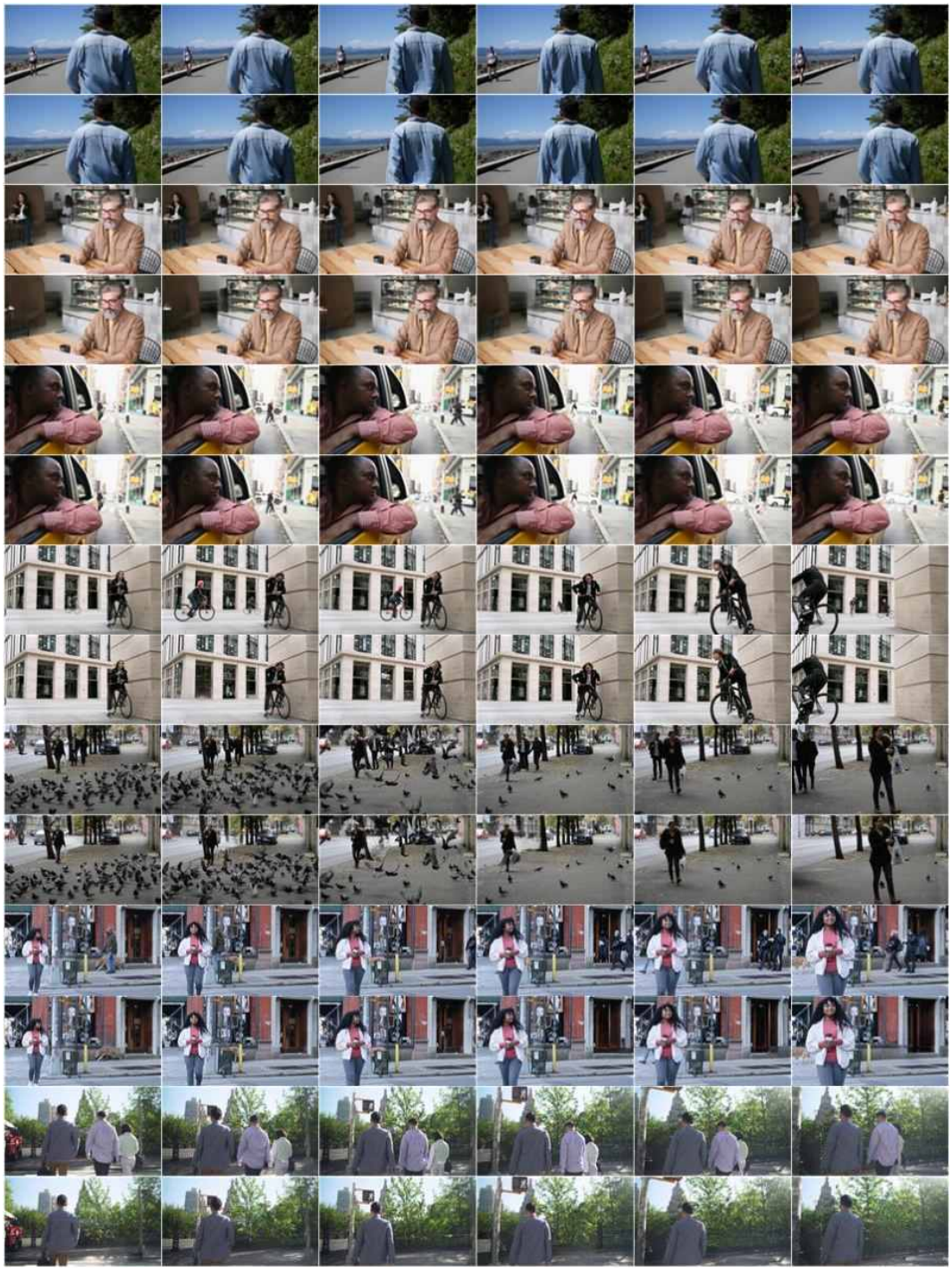
[표 4-7] 모델별 기능 비교

	Deepfillv2	AOT-GAN	MAT	LaMa	IA-Remove	proposed
1) main object select	✗	✗	✗	✗	✗	✓
2) multi object	✓	✓	✓	✓	✗	✓
3) object masking	✗	✗	✗	✗	✓	✓
4) video input	✗	✗	✗	✗	✓	✓

6. 비디오 입력

본 논문에서 제안한 시스템은 이미지 처리 성능에 기대어 이미지뿐 아니라 비디오로 또한 입력 데이터로 고려될 수 있다. [그림 4-12]는 제안한 시스템이 비디오 입력을 처리한 결과물로써, 홀수 행이 입력 프레임이며 짝수 행이 불필요한 객체가 제거된 프레임이다. 첫 번째 프레임에서 선정된 중심 객체는 객체 추적 모델의 객체 매칭을 통해 비디오 전체에서 중심 객체로 처리되었다. 이에 따라 지워야 하는 객체가 상하좌우로 움직임에도 결과 프레임에서 중심 객체가 유지되었다. 프레임별로 입력과 출력을 비교하면, 입력 프레임에서 불필요한 객체만 제거되었음을 확인할 수 있으며, 마스킹 영역이 주변 배경과 어우러져 큰 번짐 현상없이 복원되었음을 알 수 있다. 그러나 객체를 하이브리드 마스크로 마스킹하였기 때문에 동영상으로 비교했을 때, 일부 프레임에서 배경이 우그러지는 현상이 포착되거나 배경 객체가

왜곡되는 경우가 발생했다. 또한, 배경에 존재하는 다른 객체가 가려졌을 때 해당 객체가 특정 프레임에선 사라졌다가 다음 프레임에서 다시 관찰되는 현상 또한 발생하였다.



[그림 4-12] 비디오 인페인팅 결과

V. 결론

본 논문에서는 불필요한 다중 객체를 효율적으로 제거하고자 경량화된 다중 객체 인페인팅 시스템을 제안하였다. 이를 위해 중심 객체 선정 과정에서 다비스-볼드윈 지수를 도입하였으며, YOLOv7과 객체 추적 모델을 통해 영상의 객체를 식별하였고, 인페인팅 모델의 성능은 유지하되 FPS를 높이고자 하이브리드 마스크와 새로운 합성곱 블록을 제안하였다.

중심 객체 선정 과정에서 도입한 다비스-볼드윈 지수는 객체의 크기와 영상 중앙까지의 거리를 같은 단위에서 비교한다는 데 의의가 있으며, 대부분의 중심 객체에서 낮은 값을 보여 중심 객체 선정 지표로 적합하였다. 그러나 다비스-볼드윈 지수는 비교하는 객체의 크기가 상하 또는 좌우로 상당히 길어지면 외부 클러스터와의 거리가 멀어진 것으로 해석되어 중심 객체를 다르게 선정할 위험이 있다. 또한, 영상의 특수성에 따라 중심 객체의 정의가 달라질 수 있으나, 이를 반영할 수 없다는 한계가 있다. 따라서 이러한 특수 상황까지 고려하여 중심 객체를 선정하고자 한다면 행동을 기반으로 사람 객체를 탐지하는 Cheong, Bin, et al.[33]의 연구처럼 추가적인 중심 객체 선정 모델이 필요하다.

자동 마스킹과 객체 추적 과정은 탐지된 객체 정보에 따라 진행되어 YOLOv7 성능에 의존한다는 구조적인 문제가 있다. 이는 YOLOv7보다 객체 탐지 성능이 뛰어난 다른 객체 탐지 모델을 이용하면 해결할 수 있는 부분이나, 객체 탐지 모델의 연산량이 증가함에 따라 전체 시스템의 추론 시간 또한 증가할 수 있어 연구가 필요한 부분이다.

---

33) Cheng, B., Chen, P., Zhang, X., Fang, K., Qin, X., and Liu, W. (2023). Personalized Privacy Protection-Preserving Collaborative Filtering Algorithm for Recommendation Systems. Applied Sciences, 13(7), 4600.

하이브리드 마스크는 YOLO의 신체 부위 탐지 성능에 기대어 세그멘테이션 마스크와 유사하게 만든 마스크이다. 마스크 생성 과정에 필요한 시간을 고려했을 때 하이브리드 마스크는 세그멘테이션 마스크의 대안이 될 수 있으나, 여전히 인페인팅에서 세그멘테이션 마스크가 하이브리드 마스크보다 높은 성능을 보이는 것은 자명하다. 이에 향후 연구로 자세 탐지(Pose Estimation) 등의 기법을 통해 통해 하이브리드 마스크를 보완하는 것을 고려할 수 있다.

참 고 문 헌

1) Faklaris, C., Cafaro, F., Blevins, A., O'Haver, M. A., and Singhal, N. (2020). A snapshot of bystander attitudes about mobile live-streaming video in public settings. In Informatics, 7(2), 10.

2) Davies, D. L., and Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

3) Pramod, R. T., Katti, H., and Arun, S. P. (2018). Human peripheral blur is optimal for object recognition. arXiv preprint arXiv:1807.08476.

4) Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In 2016 IEEE international conference on image processing (ICIP), 3464-3468.

5) Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP), 3645-3649.

6) Wang, Z., Zheng, L., Liu, Y., Li, Y., and Wang, S. (2020). Towards real-time multi-object tracking. In European Conference on Computer Vision, 107-122.

7) Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 129, 3069-3087.

8) Uittenbogaard, R., Sebastian, C., Vijverberg, J., Boom, B., and Gavrila,

D. M. (2019). Privacy protection in street-view panoramas using depth and multi-view imagery. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10581-10590.

9) Alkobi, N., Shaham, T. R., and Michaeli, T. (2023). Internal Diverse Image Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 648-658.

10) Shetty, R. R., Fritz, M., and Schiele, B. (2018). Adversarial scene editing: Automatic object removal from weak supervision. Advances in Neural Information Processing Systems, 31.

11) Darapaneni, N., Kherde, V., Rao, K., Nikam, D., Katdare, S., Shukla, A., ... and Paduri, A. R. (2022). Contextual Attention Mechanism, SRGAN Based Inpainting System for Eliminating Interruptions from Images. arXiv preprint arXiv:2204.02591.

12) Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, 5505-5514.

13) Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2019). Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF international conference on computer vision, 4471-4480.

14) Zeng, Y., Fu, J., Chao, H., and Guo, B. (2022). Aggregated contextual transformations for high-resolution image inpainting. IEEE Transactions on Visualization and Computer Graphics.

15) Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., and Jia, J. (2022). Mat: Mask-aware transformer for large hole image inpainting. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10758-10768.

16) Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., ... and Lempitsky, V. (2022). Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2149-2159.

17) Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., and Chen, Z. (2023). Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790.

18) Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11461-11471.

19) Yang, S., Chen, X., and Liao, J. (2023). Uni-paint: A Unified Framework for Multimodal Image Inpainting with Pretrained Diffusion Model. In Proceedings of the 31st ACM International Conference on Multimedia, 3190-3199.

20) Han, S., Srivastava, A., Hurwitz, C., Sattigeri, P., and Cox, D. D. (2020). not-so-BigGAN: Generating High-Fidelity Images on Small Compute with Wavelet-based Super-Resolution. arXiv preprint arXiv:2009.04433.

21) Belousov, S. (2021). Mobilestylegan: A lightweight convolutional neural network for high-fidelity image synthesis. arXiv preprint arXiv:2104.04767.

22) Vasileiou, C., Smith, J., Thiagarajan, S., Nigh, M., Makris, Y., and Torlak, M. (2022, October). Efficient CNN-based super resolution algorithms for mmWave mobile radar imaging. In 2022 IEEE International Conference on Image Processing (ICIP), 3803-3807.

23) Monday, H. N., Li, J., Nneji, G. U., Hossin, M. A., Nahar, S., Jackson, J., and Chikwendu, I. A. (2022). WMR-DepthwiseNet: A Wavelet Multi-Resolution Depthwise Separable Convolutional Neural Network for COVID-19 Diagnosis. Diagnostics, 12(3), 765.

24) FairMOT: On the Fairness of Detection and Re-Identification in Multi-Object Tracking. Available: https://github.com/ifzhang/FairMOT

25) Wang, C. Y., Bochkovskiy, A., and Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7464-7475.

26) Sharma, D., Hade, T., and Tian, Q. (2022). Comparison Of Deep Object Detectors On A New Vulnerable Pedestrian Dataset. arXiv preprint arXiv:2212.06218.

27) Zhang, S., Xie, Y., Wan, J., Xia, H., Li, S. Z., and Guo, G. (2019). Widerperson: A diverse dataset for dense pedestrian detection in the wild. IEEE Transactions on Multimedia, 22(2), 380-393.

28) Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., ... and Bai, S. (2022). Occluded video instance segmentation: A benchmark. International Journal of Computer Vision, 130(8), 2022-2039.

29) Wang, L., Shi, J., Song, G., and Shen, I. F. (2007). Object detection combining recognition and segmentation. In Asian conference on

computer vision, 189-199.

30) Karthika, N. J., and Chandran, S. (2020). Addressing the false positives in pedestrian detection. In Electronic Systems and Intelligent Computing: Proceedings of ESIC 2020, 1083-1092.

31) López-Cifuentes, A., Escudero-Vinolo, M., Bescós, J., and García-Martín, Á. (2020). Semantic-aware scene recognition. Pattern Recognition, 102, 107256.

32) Suresh, K., and Sakthi, U. (2018). Robust multi-thresholding in noisy grayscale images using Otsu's function and harmony search optimization algorithm. In Advances in Electronics, Communication and Computing: ETAEERE-2016, 491-499.

33) Cheng, B., Chen, P., Zhang, X., Fang, K., Qin, X., and Liu, W. (2023). Personalized Privacy Protection-Preserving Collaborative Filtering Algorithm for Recommendation Systems. Applied Sciences, 13(7), 4600.

ABSTRACT

Multi object tracking and inpainting system

Lee Hyo Jin
Department of Future Convergence
Technology Engineering
Graduate School of
Sungshin University

In the image, objects unrelated to the content can distract viewers from focusing on the main object, and may pose problems such as privacy breaches, necessitating their removal. In this paper, we propose a lightweight multiple-object inpainting system. By combining the main object selection process, object tracking technology, and image inpainting techniques, we effectively remove multiple objects while automating the entire process for swift performance. To identify the objects to be removed, we introduce the Davies-Bouldin Index and propose a methodology for selecting main objects. We generate a hybrid mask corresponding to the segmentation mask using detection results for human body parts and automate multiple object tracking and masking operations with the introduction of YOLOv7. Furthermore, we structure the convolution blocks of the inpainting model using Discrete Wavelet Transform and Depthwise Separable Convolution, maintaining the model's restoration capabilities while improving the system's inference speed to approximately 9 FPS. Evaluation of the system using metrics such as PSNR, SSIM, LPIPS, and FPS shows a higher restoration performance compared to existing inpainting models in terms of FPS.