이미지 복원을 위한 구조 및 의미 일관성 딥러닝 모델

이동규, 한동석*

경북대학교

jasmindoe@knu.ac.kr, *dshan@knu.ac.kr

Deep Learning Model Ensuring Structural and Semantic Consistency for Image Restoration

Dong Gyu Lee, Dong Seog han*

Kyunpook National Univ.

요 약

이미지 복원은 이미지에서 훼손되거나 누락된 부분을 자연스러우면서 일관되게 복원하는 것을 목표로 한다. 딥러닝 기법 이전의 방법들은 패치 매칭이나 텍스처 합성을 기반으로 하지만 최근에는 PixelHacker와 같은 확산 모델 기반 기법이 구조 및 의미적 문제를 극복하여 뛰어난 성능을 보여준다. 본 논문에서는 이미지 복원의 구조적 의미적 일관성 문제를 극복하며 동시에 세부적인 요소의 복원 성능을 향상시키기 위한 모델을 제안한다. 제안 모델은 다른 딥러닝 모델 기반의 이미지 복원 방식들과 비교하여 FID(Fréchet inception distance)와 LPIPS(learned perceptual image patch similarity)의 성능이 우수하다.

Ⅰ. 서 론

이미지 복원은 오래된 사진 복원, 특정 객체 제거, 영상 편집 등 다양한 분야에서 사용된다. 초기의 텍스처 합성[1]이나 패치 기반[2] 방법은 단순한 영역에서는 그럴듯한 복원 결과를 보여주지만 복잡하며 작은 영역에서는 의미있는 구조를 복원하는 데 한계가 있었다. 이미지 복원 연구는 딥러닝을 통해 크게 발전하였다. 초기의 방식과 다르게 딥러닝 기반의 방식은 더욱 복잡하고 일관성을 가지는 결과를 보여준다. 콘텍스트 인코더(context encoders)[3]는 인코더-디코더 구조와 적대적 손실함수를 활용해 이미지 내 자연스러운 복원을 시도하였다. 이후 RePaint[4]는 확산 모델을 활용해 점진적 복원을 수행했으며 PnP 인페인팅(plug-and-play inpainting)[5]은 외부 조건을 접목하였다. 최근 PixelHacker[6]는 구조적 일관성과 의미적 일관성을 동시에 고려하는 모델을 제안하여 기존 대비 현저히 향상된 결과를 보여주었다. 그러나 해당 모델의 경우 영역 크기에 대한 복원이나 세부적인 부분에 대한 복원 성능이 저하되는 문제가 있다.

이미지 복원에서는 가려진 영역이 어떠한 반복된 구조 또는 색 등의 특징 패턴을 어느 영역까지 이어지는 구조적 일관성과 복원되는 영역이 주변 환경과 어울릴 수 있는 의미적 일관성을 가지는게 중요하다. 이를 위해 다양한 기법들이 사용되지만 여전히 세부적인 부분이나 특정 영역 크기에 대한 한계점이 존재한다. 본 논문은 딥러닝 기반의 확산 기반 방식을 통해 이미지 복원의 구조 및 의미적 일관성을 유지하며 작은 영역의 크기에 대해서도 상세한 복원이 가능한 모델을 제안한다.

Ⅱ. 이미지 복원 모델 구조

이미지 복원의 구조적 일관성 및 의미적 일관성을 유지하기 위해 원본 이미지에 마스크 정보가 적용된 정보를 입력으로 받는다. 이 입력은 확산 기반 모델에서 구조적 정보와 의미적 정보를 이용하여 복원이 진행된다. 구조적 정보와 의미적 정보는 LCG(latent categories guidance)를 통해 전경 임베딩과 배경 임베딩을 통해 얻는다. 전경 임베딩은 이미지 내 목표로하는 객체 또는 영역을 마스크화한 객체 의미 마스크로 학습하고 배경 임베딩은 특정 객체 또는 영역 이외의 배경 장면을 의미론적 마스크로 주변 문맥과 구조적 정보를 학습한다. 해당 과정에서 모델이 특정 영역 크기나 복잡한 경계영역에서 일관성을 확보하기 위해 C2F(coarse-to-fine) 구조를 사용한다. 원본 이미지는 마스크 이미지와 함께 특정한 크기의 해상도로 변경하기 위해 그림 1의 피라미드 샘플링을 통해 다운 샘플링을 진행한다. 변형된 해상도들은 x^{l_M}로 변형되어 입력으로 사용되고 입력 시 상위 스케일은 하위 스케일의 출력을 업샘플링한다.

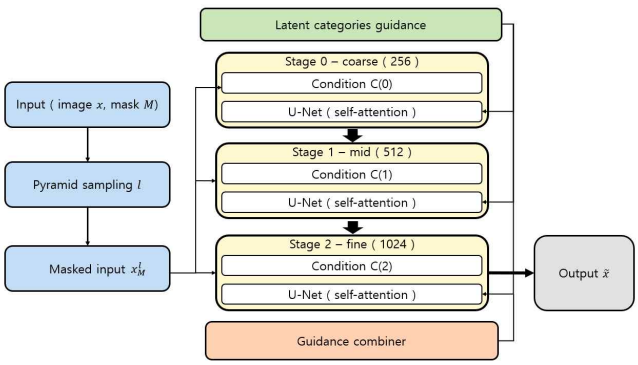

그림 1 제안하는 이미지 복원 모델 구조도

LCG의 전경과 배경의 잠재 임베딩을 미리 정해둔 4가지 타입의 마스크 이미지로 학습하고 그림 1에서 스테이지별 U-Net[8] 형태의 어텐션 경로에 사용하기 위해 특정 입력 형태로 변환한다. C2F의 구조에서 각 스테이지마다 어텐션 구조를 사용한다. 모델의 역확산 시 노이즈 예측은 CFG(class free guidance)[7]를 사용하여 예측을 진행한다.

저해상도에서는 이미지의 전역 구조와 패턴, 색과 같은 부분을 복원하게 학습하고 고해상도에서는 LCG의 잠재 임베딩을 통해 구조적, 의미적 일관성을 확보한다. 동시에 각 단계에서 이미지의 마스크 경계 부분 문제를 위해 경계 보존 손실과 업샘플링 정합 손실을 보정한다. CFG는 저해상도에서는 영향을 크게 주고 단계가 올라갈수록 영향을 줄인다. 이는 초기에 저해상도에서 이미지의 전역적인 구조와 의미를 쉽게 얻도록 유도하고 이후에는 자연스러운 의미론적 형성에 방해가 되지 않게 하기 위함이다.

각 단계에서는 입력이 들어옴과 동시에 LCG의 잠재 임베딩 또한 입력된다. 입력은 이전 단계의 결과값에서 가이던스 컴바이너(guidance combiner)를 통해 일관성을 보정해주고 U-Net 구조의 셀프 어텐션을 통해 복원이 진행된다. 셀프 어텐션은 선형 구조의 모델을 사용하여 기존 어텐션 모듈의 계산량을 감량한 모델을 사용한다. 선형 구조를 보완하기 위해 스위쉬 활성화 함수[10]를 사용하여 선형 함수에 비선형성을 추가해 학습 효율을 증가시켜준다.

손실함수는 각 마스크별 가중 노이즈, 이미지 간 L1 손실, 마스크 경계, 업샘플링 정합 손실로 총 4가지의 손실함수가 사용된다. 손실함수들은 각 스테이지별로 계산되며 가중 노이즈는 가중치를 통해 예측 오차의 학습을 조절한다. L1 손실과 마스크 경계 손실 함수는 복원된 부분의 생성된 결과의 자연스러움와 경계면의 불균형을 방지한다. 업샘플링 정합 손실 함수는 가우시안 블러를 통해 정합하여 일관성을 보정해준다.

Ⅲ. 학습 및 실험결과

모델의 학습을 위한 데이터로는 places2, FFHQ(flickr-faces-hq), celebA-HQ를 사용하고 마스크 이미지는 총 4가지 타입을 사용한다. 마스크 타입은 전경 객체 마스크, 배경 객체 마스크, 배경 무작위 마스크, 배경 무작위 객체 마스크로 사용되고 우선적으로 LCG의 학습을 하게 된다. LCG의 학습은 4가지 마스크 타입을 통해 해당 모듈에서 전경 잠재 임베딩과 배경 잠재 임베딩의 정보를 추출하게 된다. 객체 마스크들의 객체 클래스는 미리 정해둔 클래스를 사용하고 랜덤 마스크는 무작위 부위에 대해 마스크를 생성한다.

피라미드 다운샘플링 단계는 256, 512, 1024로 나누어 사용하게 되고 스테이지 내에서 단계별로 반복되는 횟수인 스텝 T의 값은 60, 40, 30으로 사용한다. 학습 시 초기 스테이지에서 전역적인 정보를 학습하기 위해 스테이지별 가이던스 게인을 크게 주고 다음 스테이지로 갈수록 작아지게 값을 설정한다.

Method	Places2 (large mask)		Places2 (Small mask)		CelebA-HQ	
	FID	LPIPS	FID	LPIPS	FID	LPIPS
CoModGAN	2.97	0.231	1.18	0.167	5.87	0.155
DeepFill	9.88	0.263	3.90	0.164	23.10	0.324
MADF	7.38	0.234	3.10	0.170	7.11	0.245
PixelHacker	2.64	0.177	1.06	0.142	4.89	0.198
Proposed Model	2.62	0.173	0.99	0.101	4.72	0.172

표 1은 딥러닝 방식의 모델들과 제안하는 모델의 성능을 places2와 celebA-HQ의 데이터를 사용하여 FID와 LPIPS를 비교하였다. places2에서는 마스크의 영역 크기[9]에 따라 두가지로 나누어 성능을 하였고 CelebA-HQ는 랜덤한 이미지들을 사용하여 평가하였다. 전체적으로 제안하는 모델이 FID와 LPIPS에서 좋은 성능을 보여준다. Places2 데이터에서 큰 마스크에 대해서는 pixelhacker와 비교해서 큰 차이를 보이지 않는다. 하지만 작은 마스크에 대해서는 특히 좋은 성능을 보여준다. 이는 모델이 구조 및 의미적 일관성을 유지함과 동시에 각 스테이지별로 입력되는 임베딩 정보를 사용하여 낮은 단계에서는 이미지 특징의 전역적인 정보를 학습한다. 그리고 단계가 올라갈수록 학습된 전역정보에 전경 객체 정보 외에도 주변 정보와 의미론적 일관성을 유지할 수 있다. 이때 가이던스 컴바이너(guidnace combiner)가 다음 단계 입력 시 일관성을 보정해주기 때문에 더 좋은 복원 결과를 보여준다. CelebA-HQ는 places2보다 데이터 수가 부족하고 유사한 정보들 구조인 사람 얼굴 데이터이다. 그렇기 때문에 places2 보다는 복원에 있어 모델이 인식해야할 구조에 대해서 어려움이 있어 전체적인 평가 지표가 낮게 나온 것으로 보인다.

Ⅳ. 결론

본 논문에서는 딥러닝 기반의 확산 기반 모델을 사용하여 이미지 복원의 성능을 향상시킨 모델을 제안하였다. 제안 모델은 기존의 확산 기반 모델들이 구조적 일관성과 의미적 일관성 성능에서 좋은 성능을 보여주면서 특정한 크기의 영역 복원에서도 좋은 성능을 보여준다. 복원 단계를 나누어 저해상도에서는 전역 구조와 색과 같은 부분을 복원하고 고해상도로 단계적으로 올라가며 구조 및 의미적 일관성도 확보하게 된다. 동시에 전경과 배경 임베딩의 정보를 사용하고 다음 단계로 넘어갈 때 가이던스의 보정을 통해 일관성을 보정하게 된다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부(산업통상자원부)의 재원으로 한국산업 기술진흥원의 지원을 받아 수행된 연구임(P0024162, 2023년 지역혁신클러스터 육성).

참 고 문 헌

[1] A. A. Efros and T. K. Leung, “Texture synthesis by non-parametric sampling,” in Proceedings of the seventh IEEE international conference on computer vision, vol. 2. IEEE, 1999, pp. 1033 - 1038.

[2] A. Criminisi, P. Pérez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” IEEE Transactions on image processing, vol. 13, no. 9, pp. 1200 - 1212, 2004.

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536 - 2544.

[4] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, “Repaint: Inpainting using denoising diffusion probabilistic models,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11 461 - 11 471.

[5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” arXiv preprint arXiv:2011.13456, 2020.

[6] Z. Xu, K. Duan, X. Shen, Z. Ding, W. Liu, X. Ruan, X. Chen, and X. Wang, “Pixelhacker: Image inpainting with structural and semantic consistency,” arXiv preprint arXiv:2504.20438, 2025.

[7] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” arXiv preprint arXiv:2207.12598, 2022.

[8] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234 - 241.

[9] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, “Mat: Mask-

aware transformer for large hole image inpainting," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 758 - 10 768.

[10] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," arXiv preprint arXiv:1710.05941, 2017.