BigGAN 인버전을 활용한 흑백 영상 색상화

김건웅 [O,1], 강경국 [2], 김성태 [1], 이화윤 [1], 김세훈 [3], 김종현 [3], 조성현 [1,2]
포항공과대학교 {인공지능대학원 [1], 컴퓨터공학과 [2]}, 삼성전자 [3]
{k2woong92, kkang831, seongtae0205, hwayoon2}@postech.ac.kr,
{sh0264.kim, jh015.kim}@samsung.com, s.cho@postech.ac.kr

요 약

최근 딥러닝의 발전과 함께 다양한 흑백 영상 색상화 연구가 소개되었고 적대적 생성 신경망 인버전을 활용한 흑백 영상 색상화 기법이 좋은 결과를 보이고 있다. 이들은 흑백 영상을 적대적 생성신경망의 잠재공간에 사상해서 사실적인 색상이 포함된 영상을 생성한다. 하지만 적대적 생성 신경망 인버전 기법을 활용한 흑백 영상 색상화는 적대적 생성 신경망의 표현력 부족으로 복잡한 영상의 색상화에 실패하는 경향을 보인다. 따라서 본 연구에서는 적대적 생성 신경망의 중간 특징 공간을 인버전 공간으로 이용하여 다양한 범주의 흑백 영상을 높은 품질로 색상화하는 방법을 제안한다.

1. 서론

흑백 영상 색상화는 주어진 흑백 영상에 대해 적절한 색상 정보를 추정하여 자연스럽게 채색된 영상을 생성하는 작업이다. 최근 딥러닝의 급격한 발전에 따라 인공 신경망 (Artificial Neural Nets) 기반의 흑백 영상 색상화 연구가 다양하게 소개되었다[4,5,6,7,15,16,17,18]. 인공 신경망 기반의 초기 연구에서는 학습 데이터 셋에 존재하는 여러 흑백 영상과 컬러 영상 쌍 (pair)에 대해 흑백 영상의 채색 결과를 컬러 영상과 비교하는 손실 함수를 사용하여 인공 신경망 모델을 학습한다[6]. 하지만 이와 같이 픽셀 값 차이 기반의 손실 함수로 학습된 모델은 데이터 셋의 평균 색상을 생성하게 되며 그 결과로 흐릿한 색상의 영상을 생성하는 문제를 갖는다.

이러한 문제를 해결하기 위해서 적대적 생성망 (Generative Adversarial Nets)[7]의 사실적인 영상 생성 능력을 흑백 영상 색상화 문제의 사전 지식 (prior)으로 이용하려는 방법들이 제안되었다[4,5,7,18]. GANs 는 잠재 벡터 z로부터 사실적인 이미지를 생성하려는 생성자 (Generator)와 생성자로부터 생성된 이미지를 실제 이미지와 구별하는 판별자 (Discriminator)가 서로 경쟁하는 형태로 학습되는 구조로 생성자가 매우 사실적인 이미지를 생성하도록 학습된다. GANs 를 흑백 영상 색상화 문제의 사전 지식으로 사용하기 위해서 주로 GAN 인버전 (GAN inversion) 기법이 이용된다. GAN 인버전 기법은 주어진 입력 이미지에 대해 해당 이미지를 생성할 수 있는 GAN 의 잠재 벡터를 찾는 기법으로, 흑백 영상 색상화에서는 주어진 흑백 이미지를 동일한 구조 (structure)를 갖는 이미지를 생성하는 GANs 의 잠재 벡터 z로 인코딩 (encoding)하고, 이를 GANs 의 생성자를 이용해서 디코딩 (decoding)하여 사실적인 색상의 이미지를 생성한다.

그림 1. BigGAN 인버전을 통한 영상 재구축 결과. 영상의 내용물을 재구축하는 데에 실패

[4,5,11]에서는 단일 클래스의 영상들로 학습된 StyleGAN[1,2]을 사용하여 높은 품질의 영상 복원에 성공하였지만, 단일 클래스의 영상만을 복원할 수 있다는 한계를 가졌다. [18]은 ImageNet 1K[9] 데이터 셋으로 학습된 BigGAN[3]을 이용하여 일반적인 흑백 영상에 대한 색상화 방법을 제안하였다. 이들은 BigGAN 인버전을 수행하여 BigGAN 생성자의 특징 맵 (feature map)을 오토 인코더 (auto encoder) 구조의 가이던스 (guidance)로 사용하여 생생한 색상을 복원한다. 하지만 다양한 객체가 포함된 복잡한 영상에 대해서는 완벽히 잘못 추정하는 문제가 존재한다. 이러한 성능 저하의 핵심 원인으로는 BigGAN 인버전의 어려움이 있다. 그림 1 과 같이 BigGAN 인버전은 잠재공간 Z 의 표현력 부족으로 원본 영상의 구조를 복원하기

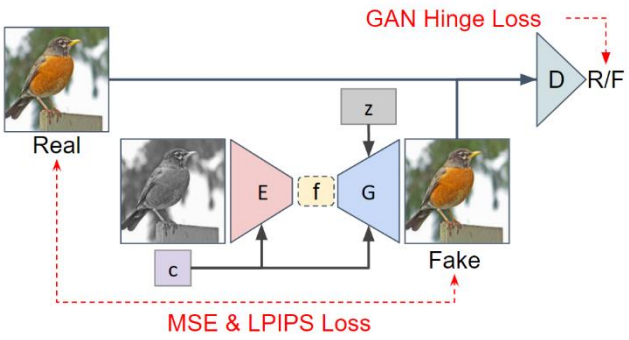

그림 2. 본 논문의 전체 프레임워크. c 는 영상의 클래스 정보, F 는 중간 특징 잠재공간, E 는 인코더, G 와 D 는 생성자와 판별자

어렵다. 이러한 경우 영상 복원에 사용될 인버전 결과가 복원 과정에 잘못된 정보를 전달하게 되며 정확한 영상 복원을 어렵게 만든다.

최근 위와 같은 잠재공간의 표현력 부족 문제를 해결하기 위해서, StyleGAN[1,2]의 중간 특징 맵 (intermediate feature map)을 인버전 잠재공간으로 이용하는 연구가 제안되었다[10]. GAN 의 중간 특징 맵은 기존 GAN 의 잠재벡터 z 보다 훨씬 큰 표현력 (expressive power)를 지니며 이는 GAN 인버전에서 다룰 수 있는 이미지의 범위를 크게 증가시킨다.

본 논문에서는 주어진 흑백 영상을 BigGAN[3]의 중간 특징 맵 인버전을 통한 흑백 영상 채색화 프레임워크를 제안한다. 구체적으로 입력 흑백 영상을 중간 특징 맵으로 인코딩하기 위한 인코더를 설계하였고 BigGAN 생성자와 함께 결합하여 (jointly) 학습한다.

다음으로 다양한 실험을 통해 본 논문의 결과가 기존의 잠재공간 Z로 사상했을 때 영상의 구조가 망가지는 한계점을 극복하고, 일반 영상에 대한 흑백 영상 색상화 성능을 크게 증가시킴을 보인다.

2. 색상화 모델 설계

2.1 BigGAN 인버전을 활용한 색상화

본 연구의 전체 프레임워크를 그림 2 에 나타냈다. 일반적인 흑백 영상에 색상화를 수행하기 위해서 ImageNet 1000K 로 학습된 BigGAN[3]을 인버전 대상 모델로 사용한다. 주어진 흑백 영상과 해당 클래스 라벨이 인코더 (encoder, E)를 통해 BigGAN 의 중간 특징 맵으로 인코딩되며 BigGAN 의 생성자 (G)를 통해 채색된 RGB 영상을 생성한다. 이후 RGB 영상을 Lab color model 로 변환하고 L(Lightness) 정보를 입력으로 사용된 흑백 영상의 L 정보로 대체한다. 이 과정을 통해서 원본 영상의 구조를 잘 유지함과 동시에 풍부한 색상을 갖는 영

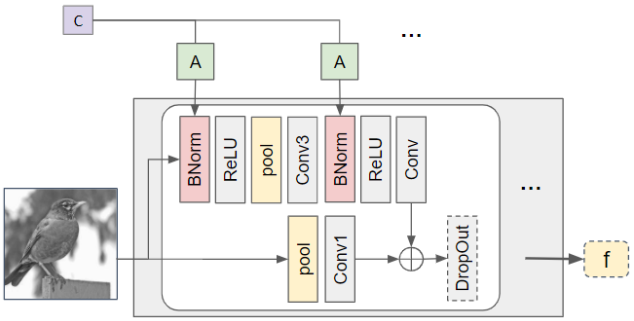

그림 3. 본 논문에서 제시하는 ResNet 기반의 인코더 신경망. c 는 영상의 클래스 정보, A 는 MLP, f 는 중간 특징 인버전 결과

상을 얻을 수 있게 된다.

클래스 라벨 (c)은 인코더 및 생성자의 중간 계층 (layer)에 투입되어 의미론적 내용이 고려된 색상을 생성하도록 만든다. 또한 기존 BigGAN[3] 생성자는 영상 생성에 사용된 잠재변수 z 의 일부를 클래스 라벨 (c)와 붙여서 (concatenate) 생성자의 중간 레이어 입력으로 사용하는데, 본 연구에서는 해당 잠재변수 z 를 표준정규분포에서 임의로 샘플링 (random sampling)된 값으로 사용한다. 이는 BigGAN 생성자가 인코더가 추정한 중간 특징 맵을 입력으로 사용함에 따라 잠재변수 z의 중요도가 상대적으로 낮아지기 때문이다. 또한 중간 특징 맵을 사용하는 상황에서 잠재변수 z가 결과에 미치는 영향이 미비함을 실험적으로 확인하였다.

최종 생성 영상이 학습 데이터 셋의 분포와 같아지도록 판별자 (D)를 도입한다. 판별자는 BigGAN[3]의 생성자와 함께 학습된 모델을 사용하며 색상화 영상과 원본 영상을 각각 가짜(fake)와 진짜(real)로 판별하도록 학습된다. 또한 클래스 정보를 입력으로 받아 참 거짓을 판별하게 되며 이는 각각의 클래스가 가진 특징을 더 잘 생성하도록 유도한다. 이와 같이 미리 학습된 판별자를 사용한 적대적 학습은 인코더와 생성자가 판별자의 사전지식을 색상 생성에 활용할 수 있도록 만든다. 전체적인 학습 손실 함수는 2.3 절에 자세히 설명되어 있다.

2.2 인코더 구조 설계

BigGAN[3] 인버전에 사용된 인코더 구조를 그림 3 에 나타냈다. 인코더의 입력으로 사용되는 흑백 영상은 색상정보를 제외한 고품질의 내용 구조 및 의미론적 정보가 포함한다. 이 정보를 최대한 활용하기 위해서 인코더 모델을 Residual Block 으로 구성하였다. 또한, 인코더에 클래스 정보를 반영하기 위해서 MLP 가 클래스 정보를 입력으로 받아 Batch Normalization[12]의 scale 과 bias parameter 를 추정한다. 인버전 과정에 필요한 down-sampling 은 average pooling 이 사용되었고 학습데이터에 과적합 되는 것을 방지하기 위해서

Residual Block 의 마지막에 Dropout[8]을 추가했다.

중간 특징 맵의 경우 해상도가 커지면 색상의 생생함이 감소하고 반면, 해상도가 작아지면 영상의 내용물 구조를 잘못 추정해서 색상 bleeding 이 발생하는 경향성을 보였다. 따라서 실험을 통해 중간 특징 공간의 해상도를 16 x 16 으로 정했다.

2.3 인코더 학습

학습에 사용된 데이터 셋은 ImageNet 1K[9] training dataset 이고 1000 개의 class 에 대한 1,281,167 개의 영상으로 구성된다. 해당 데이터 셋은 색상정보가 풍부하지 못한 다수의 영상을 포함하고 있기 때문에 [11]에서 제안된 colorfullness 점수를 기준으로 “not colorful”에 해당되는 126,939 개의 영상을 training 데이터에서 제외시켰다. 영상은 인코더의 입력으로 사용되기 전에 흑백 영상으로 변환되는 과정을 거친다.

$$L_E = \lambda_1 MSE(\hat{x}, x) + \lambda_2 LPIPS(\hat{x}, x) \quad (1)$$

$$L_G = -\lambda_3 E[D(\hat{x}, c, y)] + L_E \quad (2)$$

$$L_D = -\lambda_4 (E\left[min\left(0, -1 + D(x, c, y)\right)\right] + E[min(0, -1 - D(\hat{x}, c, y))]) \quad (3)$$

수식 1, 2 는 인코더와 생성자를, 수식 3 은 판별자를 학습시키기 위해 사용된 손실함수를 나타낸다. 수식 1 은 색상화 영상의 전반적인 내용물 구조와 색상 톤을 유지하기 위해서 적용된 RGB color model 에 대한 MSE 손실함수와 LPIPS[13] 손실함수를 나타낸다. 수식 2 와 3 은 풍부하고 사실감있는 색상을 생성하기 위해서 적용된 GAN Hinge[14] 손실함수를 나타낸다. 수식에서 c는 클래스 조건, y 는 real/fake 라벨, $\hat{x}$은 추정된 색상화 결과를 의미하며 학습시 L_G, L_D 가 적대적으로 학습된다. λ_1, λ_2, λ_3, λ_4는 손실함수가 적용되는 비중을 나타내며 그 값은 각각 1, 0.2, 0.03, 0.5 가 적용되었다. BigGAN[3] 생성자와 판별자 모두 ImageNet[9]으로 학습된 모델 가중치 (weight)에서 fine-tuning 된다.

본 논문에서는 최적화를 위한 Adam 옵티마이저가 사용되며 초기 러닝 레이트 (learning rate)는 인코더 및 생성자 0.00003, 판별자 0.0001 로 시작하고 매 epoch 마다 0.9 배씩 감소한다. 학습에 사용된 배치 크기는 64 이며 NIVIDA RTX 3090 그래픽 카드 4 대가 사용되었다. 학습은 총 12 epoch 을 진행했고 약 3 일의 학습시간이 소요되었다.

3. 실험 결과

그림 4. 다양한 영상에 대한 색상화 결과 비교.

본 논문에서 제안하는 프레임워크를 검증하기 위해 ImageNet 1K[9]의 validation set 을 test set 으로 하여 기존 방법과의 비교를 진행하였다.

3.1 기존 방법과의 정성적 비교

그림 4 에서는 서로 다른 3 개의 흑백 영상에 대해서 본 연구와 기존 방법들의 결과를 비교하였다. 기존 방법들은 색상 정보가 풍부하게 만들어 내지 못하여 흐릿한 영상을 생성하거나, 입력 영상의 의미론적 정보를 완벽하게 이해하지 못하여 색상 bleeding 현상 또는 부자연스러운 색상의 결과를 보인다. 반면, 본 논문의 결과는 높은 정확도로 풍부하고 사실적인 채색 결과를 보여준다. 이는 미리 학습된 BigGAN[3]이 다양한 클래스에 대해 사실적인 영상을 생성할 수 있으며, BigGAN 의 중간 특징 공간을 이용하여 다양한 이미지를 표현할 수 있기에

그림 5. 내용 구성이 복잡한 영상의 색상화 결과.

가능하다.

그림 5 는 여러 객체가 복잡하게 포함되어 있는 영상에 대한 색상화 결과 비교이다. 일반적으로 영상의 복잡도가 올라갈수록 영상내의 의미론적 정보를 이해하기 어렵기 때문에 색상화 난이도 역시 증가한다. 이와 같은 이유로 기존 방법의 경우 사물간의 경계를 제대로 파악하지 못하여 색상 bleeding 현상이 발생하거나 채도가 낮은 결과를 보인다. 반면, 본 연구의 결과는 영상 내의 작은 구성요소들의 경계를 세밀하게 구분하여 생생하고 정확한 색상 생성 결과를 보여준다. 이는 적대적 생성 신경망 인버전에 사용된 중간 특징공간이 공간 해상도를 가지고 있어서 영상 내용물의 구조를 더 정확하게 추정할 수 있기에 가능하다.

3.2 기존 방법과의 정량적 비교

Method	FID[19]	Colorfullness[11]
CIC[6]	11.32	33.04
Chroma[15]	8.21	26.27
InstColor[16]	7.89	25.51
DeOldify[17]	3.49	23.79
ToVivid[18]	4.08	35.13
✓Ours	1.30	35.91

표 1. 색상화 결과에 대한 정량적 비교. FID 는 영상의 사실성을 나타냄, Colorfullnee 는 영상이 갖는 색상의 풍부함을 나타냄

표 1 은 색상화 결과에 대한 FID(Frechet Inception Score) [19] 및 Colofullness[11] 점수에 대한 정량 분석 결과를 보여준다. FID 는 색상화 결과와 원본 영상 사이에 분포 유사도를 측정하는 지표로 낮을 수록 색상화 결과가 실제 영상과 유사한 품질과 사실성을 갖는다. Colorfullnes[11]는 생성된 색상의 생생함 정도를 나타내며 수치가 클 수록 영상이 더 생생한 색상을 포함한다. 정성적 결과와 마찬가지로 본 연구의 결과는 기존 방법 대비 가장 낮은 FID 수치와 가장 높은 Colorfullness 값을 달성하였다.

4. 결론

본 논문에서는 BigGAN[3]의 중간 특징 공간 인버전 기법을 활용한 일반영상 색상화 기술을 제안한다. 또한, 실험을 통해 본 논문에서 제안된 프레임워크가 기존 방법 대비 풍부하고 생생한 색상을 높은 품질로 복원하고 다양한 범주에 대해서 일관성 있고 견고하게 동작함을 보였다.

감사의 글

이 논문은 삼성전자 무선사업부의 지원을 받아 수행된 연구임.

참고문헌

[1] Tero Karas, et al. "A style-based generator architecture for generative adversarial networks", In CVPR 2019.
[2] Tero Karas, et al. "Analyzing and improving the image quality of stylegan", In CVPR 2020.
[3] Andrew Brock, et al. "Large scale gan training for high fidelity natural image synthesis", In ICLR 2019.
[4] Tao Yang, et al. "GAN prior embedded network for blind face restoration in the wild", In CVPR 2021.
[5] Xintao Wang, et al. "Towards real-world blind face restoration with generative facial prior", In CVPR 2021.
[6] Richard Zhang, et al. "Colorful image colorization" In ECCV, 2016.
[7] Ian J. Goodfellow, et al. "Generative Adversarial Networks", In NeurIPS 2014.

[8] Nitish Srivastava, et al. "Dropout: a simple way to prevent neural networks from overfitting", In JMLR 2014.
[9] Olga Russakovsky, et al. "ImageNet large scale visual recognition challenge", In IJCV 2015.
[10] Kyoungkook Kang, et. al "GAN Inversion for Out-of-Range Images with Geometric Transformations", In ICCV 2021.
[11] David Hasler, et al. "Measuring colorfulness in natural images", In SPIE 2003.
[12] Sergey Ioffe, et. al "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", In PMLR 2015.
[13] Richard Zhang, et. al "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric", In CVPR 2018.
[14] Jae Hyun Lim, et al. "Geometric GAN", In arXiv preprint arXiv:1705.02894, 2017.
[15] Patricia Vitoria, et al. "Chromagan: Adversarial picture colorization with semantic class distribution", In WACV 2020.
[16] Jheng-Wei Su, et al. "Instance-aware image colorization", In CVPR 2020.
[17] Jason Antic. "Deoldify" In 2019. https://github.com/jantic/DeOldify.
[18] Yanze Wu, et al. "Towards Vivid and Diverse Image Colorization with Generative Color Prior", In ICCV 2021.
[19] Martin Heusel, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium", In NeurIPS 2017.