2023년도 한국통신학회 동계종합학술발표회

객체 분할 정보를 활용한 선화 생성 모델의 성능 개선

최재웅, 하정민, 이혜진, *이재구
국민대학교 일반대학원 컴퓨터공학과

* jaekoo@kookmin.ac.kr

Improving Line Drawing Generation with Instance Segmentation Information

JaeWoong Choi, JungMin Ha, Hyejin Lee, *Jaekoo Lee
College of Computer Science, Kookmin University

요 약

선화(line drawing)는 사진을 특정 스타일의 선묘 기법 형태로 변환시키는 과업으로, 사진을 선묘 기법 형태로 스타일 전이(style transfer)를 하는 모델이다. 본 논문은 사진에 대한 선화 쌍이 없더라도 선묘 기법으로 그려진 결과를 추출한다. 기존 방법[1]은 이를 위해 의미론적(semantic)인 정보인 CLIP(contrastive language-image pretraining), 그리고 기하학적(geometric) 정보인 깊이(depth) 추정을 사용하였으나, 깊이 정보에 대한 실측 정보가 부족하다는 단점이 있었다. 따라서 본 논문은 실측 정보(ground-truth)가 존재하는 객체 분할을 사용하여 기하학적 정보를 추가하는 방법을 제안하였다. 이러한 분할 정보를 추가함으로써, 깊이 추정에 대한 실측 정보가 상대적으로 적다는 단점을 보완하였으며, 최종적으로 배경 및 기하학적 정보를 더 잘 고려하는 좋은 성능의 이미지를 생성하였다.

Ⅰ. 서 론

현재 스타일 전이(style transfer)를 통해 사진을 변환하는 모델이 컨텐츠로써 많이 활용되고 있다. 본 논문은 사진에 대해 선묘 기법 형태인 선화(line drawing)를 만드는 과업으로써, 사진의 선묘 기법으로 그려진 쌍이 없더라도 선화를 만드는 모델을 제안한다.

선묘 기법 쌍이 있는 사진을 기반으로 학습한 선화 생성 모델[8]의 경우, 부족한 데이터의 양으로 제한이 있었다. 기존 방법[1]은 이러한 문제를 해결하기 위해 사진에 대한 선화의 CLIP(contrastive language-image pretraining)[2] 특징을(feature) 해당 사진의 CLIP 특징과 일치시키는 의미론적 정보와, 기하학적(geometric) 정보를 위한 깊이(depth)[4] 정보만으로 쌍이 없는 사진에 대해서도 선화를 추출했다.

그러나 본 논문은 객체 분할(instance segmentation) 정보를 추가함으로써 성능을 더욱 향상시켰다. 그 결과 선묘 기법 쌍이 없는 데이터 집합에 대해 기존 방법보다 더 나은 선화를 확인 할 수 있었으며, 기존 방법보다 CLIP 점수와 사용자 선호도 조사(user study)에서 좋은 성능을 확인할 수 있었다.

Ⅱ. 본론

본 논문의 최종 과업은 사진이 주어졌을 때, 선화를 생성하는 것이다. 기존 방법[1]은 기하학적인 정보를 위한 깊이 지도(depth map)를 실측 정보(ground truth)로 가지고 있는 데이터 집합(dataset)이 많이 존재하지 않아, 깊이 지도를 수도 레이블링(pseudo-Labeling)으로 실측 정보를 대체하였다. 그렇기에 깊이 지도를 만드는 모델에 따라 성능에 차이가 생긴다는 문제가 생겼다. 이러한 문제를 해결하기 위해, 본 논문은 실측자료가 상대적으로 많이 존재하는 객체 분할 정보를 추가해줌으로써 성능의 일관성을 보장해주었으며, 배경을 더욱 고려하는 등의 더 높은 성능의 결과를 얻을 수 있었다.

사용된 손실함수는 총 5 가지이다. 그림 1 과 같이 첫째는 적대적(adversarial) 손실함수[5]로, 스타일 도메인을 따르는 사진을 생성한다. D_B 는 판별자(Discriminator)로서 생성된 선화의 스타일(Style) 데이터 세트와 스타일 정보가 유사하도록 학습된다.

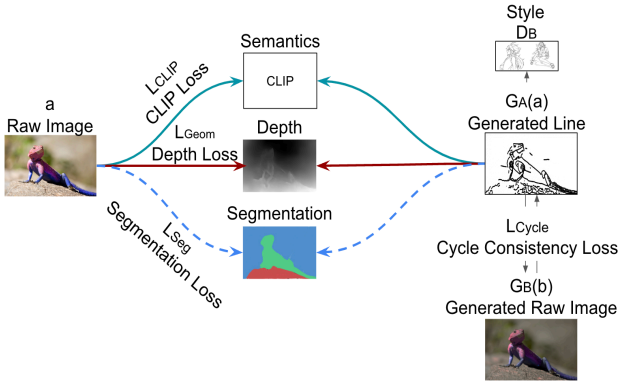

그림 1. 본 논문 제안 방법의 모델 구조

$$L_{GAN} = E_{a\sim A}[D_A(a)^2] + E_{b\sim B}[(1-D_A(G_B(b)))^2] + E_{b\sim B}[D_B(b)^2] + E_{a\sim A}[(1-D_B(G_A(a)))^2] \quad (1)$$

둘째, 깊이 손실함수[4]이다. 사진과 선화의 깊이가 유사하도록 학습한다. I 는 특징 추출기이며, F(a)는 수도 레이블이다.

$$L_{Geom} = \| G_{Geom}(I(G_A(a))) - F(a) \| \quad (2)$$

셋째, CLIP 손실함수[2] 이다. 사진과 선화 간 CLIP 임베딩(embedding) 거리(distance)를 줄인다.

$$L_{CLIP} = \| CLIP(G_A(a)) - CLIP(a) \| \quad (3)$$

넷째, Cycle 손실함수[9]는, 원본(a)을 통해 선화(b)를 생성하며, 선화(b)를 통해 원본(a)을 다시 복원한다.

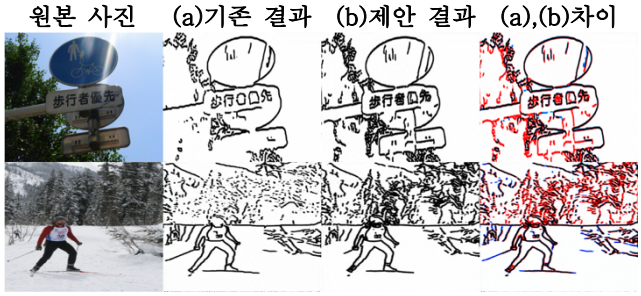

그림 2. 제안 모델의 결과와 기존 모델의 결과 차이의 시각화

그림 3. 훈련 시점에 따른 결과

$$L_{Cycle} = \| G_B(G_A(a)) - a \| + \| G_A(G_B(b)) - b \| \quad (4)$$

마지막으로 그림 1 에서 점선으로 표시된 새로운 손실함수인 분할 손실함수 L_{Seg} 를 제안하였다. 이는 선화의 분할 결과와, 원본 사진의 분할 결과가 유사하도록 학습하는 손실함수이다. 원본 사진의 정보를 잘 담고 있는 선화를 생성하였다면, 원본 사진의 분할 결과와 선화의 분할 결과 또한 유사할 것이라 가정한다.

$$L_{Seg} = \| G_{Seg}(G_A(a)) - gt(a) \| \quad (5)$$

G_{Seg} 는 U-Net[10]을 사용하여 선화의 분할된 결과를 냈으며, gt(a)는 입력값인 원본 사진의 분할 결과이다. a 는 원본 사진이며 $G_A(a)$는 선화이며, 최종적으로, 선화의 분할과 실측값의 분할의 차이를 크로스-엔트로피(cross-entropy)를 통해 줄였다. 이를 통해 선화에 대한 분할 정보를 추가하였다.

Ⅲ. 실험

본 논문은 자연에 존재하는 다양한 사진을 담고 있는 COCO-Stuff[7] 데이터 집합으로 학습되었으며, 평가를 위해 객체별로 나누어진 높은 성능의 사진을 담고 있는 MIT-Adobe FiveK[6] 데이터 집합을 사용하였다.

그림 2 는 원본 사진에 대한 본 논문의 모델과 기존 모델을 학습시킨 결과를 출력한 모델과 기존 모델과 비교하여 수정된 선을 시각화하였다(빨간선: 추가, 파란선: 제거). 그림 3 은 선화의 생성 과정을 훈련 시점(epoch)별로 출력하였으며, 이를 통하여 생성 과정 중에도 더 나은 결과를 생성하는 것을 확인할 수 있었다.

표 1 의 실험 결과를 보면 CLIP 점수지표를 사용하여 정량적으로 성능을 평가하였다. CLIP 점수 지표[1]는 CLIP(contrastive language-image pre-training) 모델에 선화 사진을 넣어 해당 사진이 어떤 문맥(context)을 담고 있는 사진인지 분류(classification) 결과를 확률적으로 보여주는 평가지표이다. 본 모델이 기존 모델보다 CLIP 점수가 2.47%만큼 높았다.

또한, 그림 4 는 사용자 선호도(user study) 지표를 조사한 방법이며, 해당 지표는 본 모델과 기존 모델의 선화를 15 명을 대상으로 각 100 장씩 평가하였다. 이 지표 또한 본 모델이 기반 모델보다 33.6%만큼 높았다.

그림 4. 사용자 선호도 지표 평가 방법

표 1. 두 모델에 대한 CLIP 점수와 사용자 선호도

	CLIP 점수(↑)	User Study(↑)
기존 모델	0.4147	0.332
제안 모델	**0.4394**	**0.668**

Ⅳ. 결론

본 논문에서는 쌍으로 이루어지지 않은 사진에 대한 라인 드로잉의 성능을 향상하기 위해 객체 분할이 추가된 손실함수를 새롭게 제안했다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2022-00167194,미션 크리티컬 시스템을 위한 신뢰 가능한 인공지능).

참 고 문 헌

[1] Chan, Caroline, Frédo Durand, and Phillip Isola. "Learning to generate line drawings that convey geometry and semantics." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

[3] Kim, Taeksoo, et al. "Learning to discover cross-domain relations with generative adversarial networks." *International conference on machine learning*. PMLR, 2017.

[4] Miangoleh, S. Mahdi H., et al. "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021

[5] Mao, Xudong, et al. "Least squares generative adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

[6] Bychkovsky Vladimir, et al. "Durand. Learning photographic global tonal adjustment with a database of input/output image pairs." In CVPR 2011, pages 97 - 104. IEEE, 2011.

[7] Lin Tsung-Yi, et al. "Microsoft coco: Common objects in context." In European conference on computer vision, pages 740 - 755. Springer, 2014.

[8] Yi Ran, et al. "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10743- 10752, 2019.

[9] Zhu Jun-Yan, et al. "Unpaired image-to-image translation using cycle consistent adversarial networks." In Proceedings of the IEEE international conference on computer vision 2017.

[10] Ronneberger Olaf, et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation" MICCAI 2015.