ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 48, No. 12, pp. 1329-1334, 2021. 12
https://doi.org/10.5626/JOK.2021.48.12.1329

RDID-GAN: 비식별화 이미지 데이터 복원을 통한 효과적인 학습데이터 생성

(RDID-GAN: Reconstructing a De-identified Image Dataset to Generate Effective Learning Data)

오 원 석 † 배 강 민 †† 배 유 석 †††
(Wonseok Oh) (Kangmin Bae) (Yuseok Bae)

요 약 최근 여러 사회 문제들을 예방 및 신속하게 대처하기 위해 CCTV가 설치되고 있고 인공지능을 활용해 이를 효과적으로 처리하는 방안이 연구되고 있다. 하지만, CCTV에서 수집한 데이터는 개인정보 침해의 우려가 있어 비식별화 작업 없이는 자유롭게 사회문제 해결을 위한 모델을 연구하는데 사용할 수 없다. 따라서, 본 논문에서는 RDID-GAN을 제안하여 비식별화된 사람의 얼굴을 임의로 복원하여 개인정보 침해의 우려를 줄이고 네트워크 학습에도 부정적인 영향을 주지 않는 효과적인 데이터셋 제작 방안을 제안한다. RDID-GAN은 attention module을 활용해 비식별화된 부분에 집중하여 합당한 결과를 생성할 수 있도록 하였다. 우리는 실험을 통해 해당 모델과 기존의 제안된 image-to-image 변환 모델을 정성적 및 정량적으로 비교하였다.

키워드: 적대적 생성망, 비식별화 데이터셋, 사람 인식기, 시각 감시, image-to-image 변환

Abstract Recently, CCTVs have been installed to prevent or handle various social problems, and there are many efforts to develop visual surveillance systems based on deep neural networks. However, the datasets collected from CCTVs are inappropriate to train models due to privacy issues. Therefore, in this paper, we proposed RDID-GAN, an effective dataset de-identification method that can remove privacy issues and negative effects raised by modifying the dataset using a de-identification procedure. RDID-GAN focuses on a de-identified region to produce competitive results by adopting the attention module. Through the experiments, we compared RDID-GAN and the conventional image-to-image translation models qualitatively and quantitatively.

Keywords: GAN, de-identified dataset, person detector, visual surveillance, image-to-image translation

· This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis and No.2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network)

† 비 회 원 : 고려대학교 화공생명공학과 학생
5kong@korea.ac.kr
†† 정 회 원 : 한국전자통신연구원 시각지능연구실 연구원(ETRI)
kmbae@etri.re.kr
(Corresponding author임)
††† 정 회 원 : 한국전자통신연구원 시각지능연구실 책임연구원
baeys@etri.re.kr

논문접수 : 2021년 9월 16일
(Received 16 September 2021)
논문수정 : 2021년 11월 9일
(Revised 9 November 2021)
심사완료 : 2021년 11월 10일
(Accepted 10 November 2021)

정보과학회논문지 제48권 제12호(2021. 12)

1. Introduction

최근 Deep Neural Network(DNN)를 이용해 다양한 CCTV 환경에서 발생하는 문제를 해결하고자 하는 노력이 있어 왔다. 또한, 컴퓨팅 파워의 발전으로 DNN의 계층이 증가하자 이를 학습하기 위해 다량의 데이터를 필요로 하기 시작했다. 하지만, CCTV로부터 획득한 데이터는 개인 정보 침해의 우려가 심각하게 존재한다. 따라서 CCTV 기반의 데이터셋에서는 개인 정보를 보호하기 위해 사람의 얼굴 부분을 모자이크, 즉 비식별화하여 영상을 제공한다. 사람의 얼굴 부분을 모자이크하는 것은 개인 정보를 보호하기 위한 확실한 방법이지만 DNN을 학습할 경우 실제 데이터와의 도메인 차이 때문에 성능 하락이 발생할 수 있다. 따라서 얼굴 부분이 비식별화된 영상을 임의의 얼굴로 복구하여 개인 정보 침해의 소지를 제거하고 네트워크의 검출 성능을 높이고자 하였다.

CCTV를 활용하여 제작된 데이터셋은 bounding box를 이용해 얼굴이나 신체의 특정 부분을 모자이크 처리하는 방식을 통해 비식별화를 진행한다. 특히 지자체에서 제공하는 데이터셋의 경우 행동 인식 등을 위한 비디오 데이터셋이 제공되지만 모든 프레임에 걸쳐 얼굴이 비식별화되어 있다. 하지만 기존 DNN의 경우 COCO[1], ImageNet[2], CelebA[3]와 같이 개인 정보 침해의 우려가 없는 데이터셋을 이용해 학습하기 때문에 사람의 얼굴이 가려진 비식별화 영상에서 사람의 위치를 잘 검출하지 못하는 경향이 있다. 특히 비식별화 강도의 세기가 강해질 수록 검출률이 더욱 낮아지는 경향이 있다. 이는 비식별화된 CCTV 영상에서 사람의 행동 및 상태를 인식하는 연구를 진행하는데 있어 어려움을 주고 있다.

본 논문에서는 사람 얼굴 비식별화의 강도를 조절해가며 이것이 네트워크의 인식 성능에 미치는 영향을 측정하였다. 그리고 비식별화된 데이터셋에서 모자이크 된 부분에 집중하여 복구하는 네트워크를 제안한다. 비식별화 데이터셋의 경우 기존의 COCO[1] 데이터셋을 활용하여 제작하였다. 데이터셋에서 제공된 사람의 annotation 파일을 이용하여 비식별화를 진행하였으며 Keypoint RCNN[4]을 이용해 사람 검출 성능을 측정하였다. 그리고 제작된 데이터셋은 비식별화를 복원할 수 있는 복구 네트워크를 학습할 때도 활용하였다. 복구 네트워크는 attention module을 이용하여 모자이크가 발생한 부분에 집중하여 비식별화된 부분을 복원할 수 있도록 학습하였다. 제작한 데이터셋에서 복구 네트워크의 복원 결과를 정성적, 정량적으로 기존의 다른 알고리즘들과 비교하였다. 정량적 평가로는 Keypoint RCNN[4]을 이용해 사람 검출 성능을 AP로 측정하는 방식을 통해 복구된 이미지의 품질을 측정하고 원본 이미지와 다른 알고리즘의 복원 결과를 비교 분석하였다.

2. Related Works

GAN[5]에서는 adversarial loss를 제안하여 두개의 네트워크가 서로 경쟁하게 하여 학습을 유도한다. 그 결과 생성 네트워크는 데이터셋의 분포를 학습하여 학습 데이터와 유사한 결과를 도출하는데 도움을 준다. 이러한 이유로 GAN[5] 기반 모델은 이미지 생성 및 변환 분야에서 활용되고 있다. 특히 SRGAN[6]과 같은 super resolution 네트워크를 학습하는 과정이나 DeepFill[7]과 같은 image inpainting 네트워크 학습에 많이 활용되고 있으며 image-to-image translation을 위한 네트워크 학습에도 활용되고 있다.

Image-to-image translation 네트워크로는 대표적으로 Pix2Pix[8]와 CycleGAN[9]이 있다. 먼저 Pix2Pix[8]는 입력 x에 대해 label y가 쌍으로 존재하는 데이터셋이 존재할 때 이를 학습하기 위한 네트워크이다. 이때 generator가 생성하는 결과와 y를 비교하는 L1 loss 이외에 adversarial loss를 추가로 활용하여 좋은 성능을 얻었다. 하지만 데이터셋이 쌍으로 존재해야 하기 때문에 데이터를 수집할 때 많은 노력을 필요로 한다.

CycleGAN[9]은 cycle-consistency loss를 활용하여 데이터셋이 쌍으로 존재하지 않아도 학습이 가능하도록 학습 방법을 변경하였다. 두 개의 generator를 동시에 학습하여 최종적으로 서로가 역함수 관계가 되도록 학습하고 입력 이미지의 형태를 유지하도록 하기 위해 identity loss를 추가하였다. 게다가 데이터셋을 제작할 때 쌍으로 존재하지 않아도 되기 때문에 필요한 노력이 적지만 원하는 부분에 attention을 주어 이미지를 생성하도록 유도하는 데 어려움이 있다.

Attention 방법으로는 대표적으로 channel attention과 spatial attention이 있다. Channel attention[10]만 이용하기보다 두 가지를 모두 활용하는 모듈인 BAM[11]과 CBAM[12]이 더 좋은 활용성을 보여주고 있다. BAM[11] channel attention과 spatial attention을 다른 branch에서 나누어 계산한 뒤 feature에 attention을 주는 방식을 활용한다. CBAM[12] 같은 경우에는 channel attention module과 spatial attention module을 직렬로 연결하여 attention 강도를 높였다. 그 결과 BAM[11]보다 높은 attention 성능을 보인다.

3. RDID-GAN

비식별화된 데이터에서 사람을 잘 인식하기 위해 얼굴 부분이 모자이크가 되어 있는 사진을 복원하여 인식

률을 높인다. 또한 비식별화된 부분에 집중하여 데이터를 복원할 수 있도록 attention module이 적용된 RDID-GAN을 제안한다. RDID-GAN은 attention module을 적용하여 사람 몸통의 형상에 집중하여 비식별화된 얼굴 부분을 복원하고 최종적으로 사람 인식 성능을 높일 수 있다.

3.1 Dataset

비식별화된 사람의 얼굴을 복원하여 사람 인식률을 높이는 것이 목적이므로 기존의 사람 인식 데이터셋을 활용하여 새롭게 데이터셋을 제작하였다. 제작을 위하여 우선적으로 COCO데이터에서 사람이 인식된 사진들을 모아서 기존의 데이터셋을 만들어 주었다. 기존의 데이터셋 에 딥러닝 기반 얼굴 검출 알고리즘인 Dual Shot Face Detector(DSFD)[13]를 적용하여 얼굴을 인식시킨 뒤 얼굴 영역을 downsampling하여 비식별화된 데이터셋을 제작하였다.

본 논문에서 사용되는 비식별화는 DSFD[13]를 통해 얻은 bounding box 영역에 대해 linear interpolation에 의해 해당 영역의 사이즈를 줄이고, nearest-neighbor interpolation 의해 다시 사이즈를 키움에 따라 pixelated 비식별화를 수행할 수 있었다.

3.2 Training RDID-GAN

비식별화된 사람을 복원하기 위해 두 개의 generator G,F와 두개의 discriminator D_X,D_Y를 활용하였다. 두 개의 generator G,F는 각각 비식별화된 사람의 얼굴을 복원하는 기능과 원본 영상을 다시 비식별화하는 기능을 가지고 있어 서로 반대의 기능을 수행한다. 그리고 D_X,D_Y는 각 입력 이미지가 해당하는 데이터셋 X, Y에 포함될 확률을 예측하는 기능을 가지고 있다. D_X,D_Y는 서로 동일한 네트워크 구조를 갖고 있고 G,F도 서로 동일한 네트워크 구조를 가지고 있어 출력 결과가 서로 동일하다. 두 개의 generator와 discriminator가 서로 적대적으로 학습된다. 이러한 학습 방법에 의해 적용되는 adversarial loss는 다음과 같다.

$$L_{adv} = E_{x\sim X}[\log(D_X(x))] + E_{y\sim Y}[\log(1-D_X(F(y)))] + E_{y\sim Y}[\log(D_Y(y))] + E_{x\sim X}[\log(1-D_Y(G(x)))] \quad (1)$$

그리고 generator의 생성 결과와 ground truth와 생성 결과 비교를 위해 L1 loss를 활용하였다. 먼저 generator의 생성 결과와 ground truth와의 비교를 위한 L_{dist} loss는 다음과 같다.

$$L_{dist} = E_{x\sim X,y\sim Y}[\|y-G(x)\| + \|x-F(y)\|] \quad (2)$$

Generator는 서로 반대의 기능을 하기 때문에 cycle-consistency loss[7]를 적용하여 학습에 활용하였다. 따라서 cycle-consistency loss[7] L_{cyc}는 다음과 같이 주어진다.

$$L_{cyc} = E_{x\sim X}[\|x-F(G(x))\|] + E_{y\sim Y}[\|y-G(F(x))\|] \quad (3)$$

따라서 최종적으로 loss function L_{total}은 다음과 같이 주어진다.

$$L_{total} = \lambda_1 L_{adv} + \lambda_2 L_{dist} + \lambda_3 L_{cycle} \quad (4)$$

이때 $\lambda_1,\lambda_2,\lambda_3$는 각 loss들에 대한 가중치 값이다. 실험에서는 $\lambda_1,\lambda_2,\lambda_3$를 각각 1, 10, 10으로 설정해주었다. 따라서 최종 목표는 최적의 discriminator와 generator인 D_X^*,D_Y^*,G^*,F^*를 찾는 것이며 이는 다음과 같이 구할 수 있다.

$$D_X^*,D_Y^*,G^*,F^* = \arg\min_{G,F}\max_{D_X,D_Y} L_{total} \quad (5)$$

3.3 Network Architectures

본 논문에서는 그림 1(a)와 같이 generator G,F를 사용하였다. generator G,F는 down sampling layer와 up sampling layer 사이에 9개의 residual block[14]으로 이루어져 있다. 각 residual block[14]은 그림 1(c)와 같이 구성하였다. 그리고 layer에는 channel 및 spatial attention을 통해 모자이크 된 부분에 집중할 수 있도록 CBAM[12]으로 이루어진 attention module을 추가하여 성능 향상을 유도하였다. Discriminator는 그림 1(b)와 같이 디자인하였다. 이는 CycleGAN[9]에서 사용한 discriminator와 동일한 구조로 local-level의 판별을 통해 생성 이미지의 품질 향상에 도움을 준다.

입력 이미지를 generator에 넣으면 3개의 down sampling layer를 통과하게 된다. 각 layer는 convolutional layer, instance normalization layer, attention module, ReLU의 순서로 만들어져 있다. Down sampling layer를 지난 feature들은 9개의 residual block을 지나게 된다. Residual block에서는 attention module을 통해 생성 이미지의 품질 향상을 유도하고 skip connection을 추가하여 학습을 유도하였다. Residual block들을 통과한 feature는 up sampling layer를 통과하게 된다. Upsampling layer는 2개의 transposed convolutional block을 통과하게 되고 마지막으로 convolutional Layer와 tanh의 활성화 함수를 통과하여 새로운 reconstructed output image를 생성한다. 생성된 이미지는 discriminator를 통과하게 되는데 5개의 convolutional block을 통과하게 된다. 그림 1(b)에서 볼 수 있듯이 attention module을 추가하여 네트워크 학습을 유도하였다.

4. Experiments

4.1 Implementation Details

기존의 사람 인식 데이터셋으로 COCO[1]의 인물사진을 사용하였다. 우선 이 사진들에 있는 사람들의 얼굴을 모자이크하였고 모자이크된 사진과 되지 않은 사진의 pair를 만들어 학습에 사용하였다. 이는 CycleGAN을

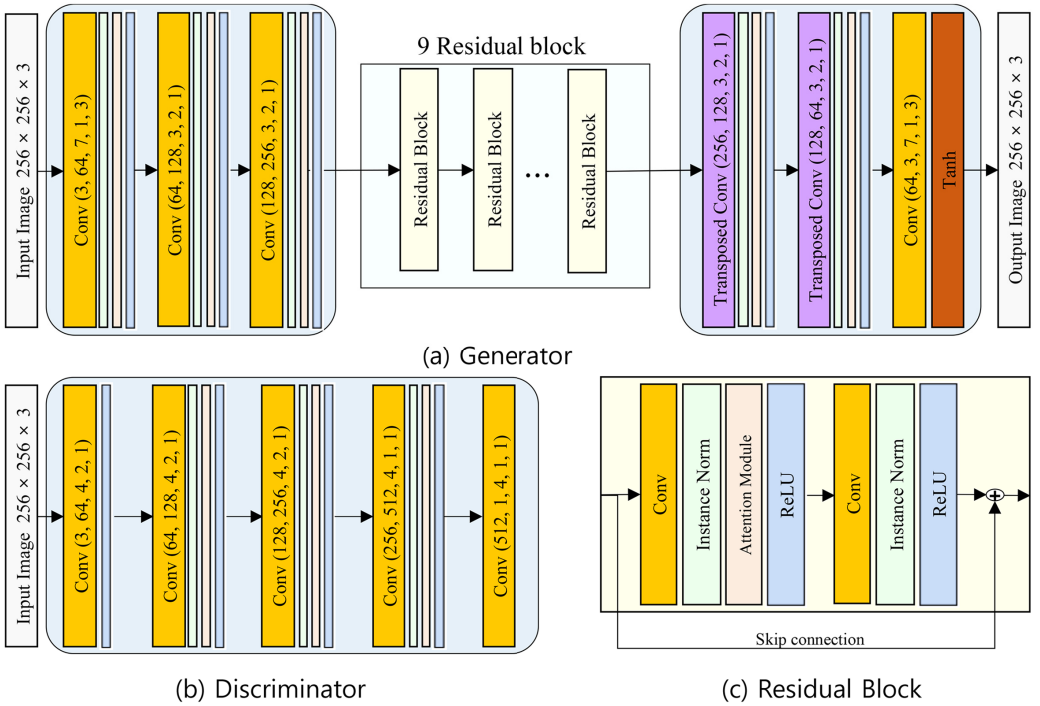

그림 1 RDID-GAN의 Generator와 Discriminator의 구조도. Conv block의 괄호 안의 숫자는 순서대로 input channel 수, output channel 수, kernel size, stride, padding을 의미한다.

Fig. 1 Network architecture of RDID-GAN generator and discriminator. The numbers inside the parenthesis in the Conv block denote input channel number, output channel number, kernel size, stride, and padding respectively. The width and height of the kernel size are identical

비롯하여 Pix2Pix와 같은 다양한 model을 사용할 수 있도록 만들기 위함이다.

Input image의 차원을 $256 \times 256 \times 3$으로 만들어 준 뒤 optimizer로 Adam을 사용했고, learning rate는 초기에 0.002로 설정하였다. 학습 파라미터로 batch size는 32로 100 epoch까지 학습하였다. 마지막으로 성능 측정 방식으로는 AP를 사용하여 측정하였다.

4.2 Qualitative Results

실험결과는 정성적인 방법과 정량적인 방법 2가지를 확인하였다. 정성적인 방법이란 다양한 방법들을 통하여 reconstructed된 image들의 qualitative results들을 비교하는 것이다. 그림 2의 사진들의 열에서 왼쪽부터 오른쪽 순서대로 Mosaic, CycleGAN[9], Pix2Pix[8], RDID-GAN(Ours)와 ground truth(GT)를 비교하였다. CycleGAN[9]의 경우 학습할 때 GT 레이블을 사용하지 않기 때문에 비식별화 된 부분을 변환하는데 어려움을 보였다. 반면, Pix2Pix[8]의 경우 비식별화된 부분을 복원하려고 하였지만 얼굴의 세부적인 부분을 복원하는데 어려움이 있었다. RDID-GAN의 경우 비식별화된 부분을 인식하고 얼굴의 세부적인 부분을 복원하여 생성 결과가 다른 알고리즘 대비 좋아지는 것을 확인할 수 있었다. 특히, 얼굴의 세부적인 사항들이 복원이 되지만 기존의 GT와는 차이가 있어 개인정보 침해의 우려를 줄일 수 있다는 것 또한 자명하다.

4.2 Quantitative Results

정량적인 방법으로는 앞서 설명한 Keypoint RCNN을 이용해 사람 검출 성능을 AP로 측정해 보았다. 이는 사람이 얼마나 검출되었는지에 관한 결과라고 볼 수 있다. 표 1의 bbox detection AP를 확인해 보면 RDID-GAN의 bbox detection AP는 53.384로 다른 알고리즘에 비해서 높게 나오는 것을 확인할 수 있다. Keypoint detection AP 또한 RDID-GAN의 값은 60.607로 다른 알고리즘에 비해서 높게 나오는 것을 확인할 수 있다. 이를 통해 RDID-GAN을 사용하게 되면 기존의 방법들에 비해 더 합당하게 복원되는 것을 확인할 수 있다.

5. Conclusion

RDID-GAN을 사용하게 되면 기존의 방법들과 비교

하였을 때 좋은 성능을 보여준다는 것을 정량적 및 정성

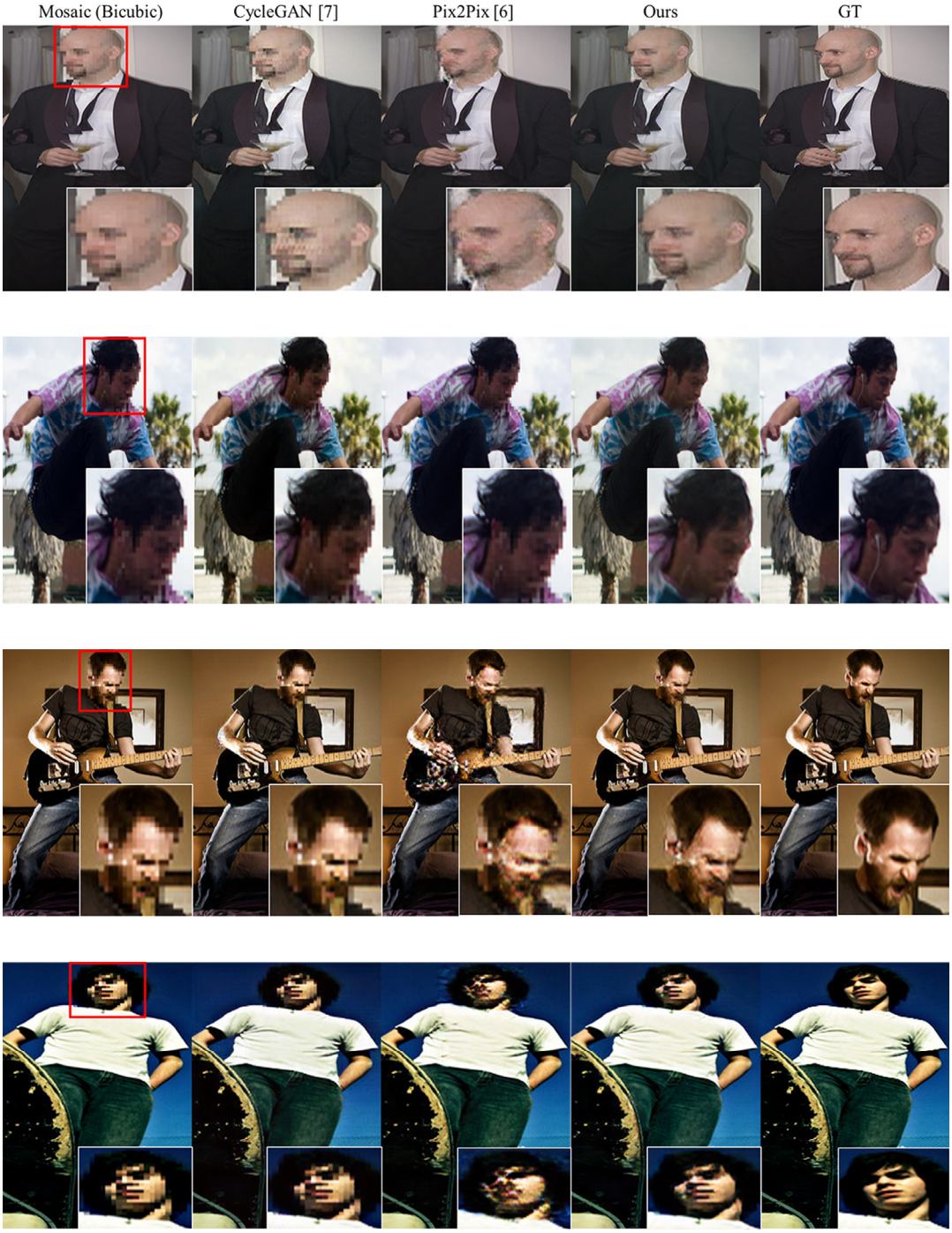

그림 2 RDID-GAN과 다른 방법들 간의 Qualitative Results 비교

Fig. 2 Comparison of the qualitative results between RDID-GAN and other methods. Displayed from left to right: mosaic (Bicubic), Pix2Pix [8], CycleGAN [9], RDID-GAN(Ours), and ground truth (GT)

표 1 AP를 이용한 정량적 성능 비교표
Table 1 Comparison of bbox and keypoints detection AP

Methods	Bbox	keypoints
GT	55.441	65.496
Ours	53.384	60.607
Pix2Pix [8]	53.303	59.838
CycleGAN [9]	52.956	56.227
Mosiac	52.168	53.377

적으로 확인하였다. 이는 RDID-GAN의 아이디어인 face에 관한 attention과 loss를 추가한 것이 의미가 있었다는 사실을 말해준다. RDID-GAN은 현실 세계의 CCTV로부터 수집한 데이터가 혹시라도 있을 수 있는 개인 정보 침해의 우려를 줄일 수 있다. 특히, 현실세계에서 발생하는 여러 사회 문제를 해결하기 위한 데이터셋을 구축하는데 많은 도움을 줄 수 있을 것이라 기대한다.

References

[1] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," *Proc. of the European Conference on Computer Vision*, 2014.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *Proc. of the Computer Vision and Pattern Recognition*, 2009.

[3] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," *Proc. of the 2015 IEEE/CVF International Conference on Computer Vision*, 2015.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proc. of the 2017 IEEE/CVF International Conference on Computer Vision*, 2017.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative Adversarial Nets," *Proc. of the NeurIPS*, 2014.

[6] C. Ledig et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," *Proc. of the 2017 IEEE/CVF International Conference on Computer Vision*, 2017.

[7] J. Yu et al., "Free-Form Image Inpainting with Gated Convolution," *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019.

[8] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proc. of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[9] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proc. of the 2017 IEEE/CVF International Conference on Computer Vision*, 2017.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[11] J. Park, S. Woo, J. Lee, and I. S. Kweon, "BAM: Bottleneck Attention Module," *Proc of the British Machine Vision Conference*, 2018.

[12] S. Woo, J. Park, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *Proc. of the European Conference on Computer Vision*, 2018.

[13] L. Jian, et al., "DSFD: dual shot face detector," *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

오 원 석

2022년 고려대학교 화공생명공학과(학사 예정). 관심분야는 기계학습, 딥러닝, 인공지능, 시각지능

배 강 민

2017년 고려대학교 전기전자전파공학부(학사). 2019년 KAIST 전기및전자공학부(석사). 2019년~현재 한국전자통신연구원 시각지능연구실 연구원. 관심분야는 기계학습, 딥러닝, 인공지능, 시각지능

배 유 석

1995년 경북대학교 전산학과(학사). 1997년 경북대학교 컴퓨터과학과(석사). 2011년 경북대학교 컴퓨터과학과(박사). 1996년~현재 한국전자통신연구원 시각지능연구실 책임연구원. 관심분야는 기계학습, 딥러닝, 인공지능, 시각지능, 빅데이터 분석, 분산 병렬 컴퓨팅