KICSP
한국 융합신호처리학회

융합신호처리학회논문지 Vol. 24, No.1 : 69~75 March. 2023
DOI : 10.23087/jkicsp.2023.24.1.009

아웃페인팅 기반 반려동물 자세 추정에 관한 예비 연구

이규빈[1], 이영찬[1], 유원상[1,2*]
[1]선문대학교 정보통신공학과 인공지능 영상처리 연구실
[2]선문대학교 바이오빅데이터융합전공, 유전체기반바이오IT연구소, 산업기술연구소

A Pilot Study on Outpainting-powered Pet Pose Estimation

Gyubin Lee[1], Youngchan Lee[1], Wonsang You[1,2*]
[1]Artificial Intelligence and Image Processing Laboratory (AIIP Lab),
Department of Information and Communication Engineering, Sun Moon University
[2]Division of Bio Bigdata, Genome-based BioIT Convergence Institute,
Research Institute for Industrial Technology, Sun Moon University

요 약 최근 동물 행동 분석 및 건강관리 분야를 중심으로 딥러닝 기반 동물 자세 추정 기법에 대한 관심이 높아지고 있다. 그러나 기존 동물 자세 추정 기법은 영상에서 신체 부위가 가려지거나 존재하지 않을 경우 좋은 성능을 보이지 않는다. 특히 꼬리나 귀가 가려진 경우, 반려견의 행동 및 감정 분석의 성능에도 심각한 영향을 미친다. 본 논문에서는 이러한 다루기 힘든 문제를 해결하기 위해, 이미지 아웃페인팅 네트워크를 자세 추정 네트워크에 연결하여 이미지 외부에 존재하는 반려견의 신체를 복원한 확장된 이미지를 생성하여 반려견의 자세를 추정하는 단순하면서도 새로운 접근방법을 제안하였고, 제안된 방법의 실현가능성을 검토하는 예비 연구를 수행하였다. 이미지 아웃페인팅 모델로는 CE-GAN과 트랜스포머 기반의 BAT-Fill을 사용하였고, 자세 추정 모델로는 SimpleBaseline을 사용하였다. 실험 결과, 크롭된 입력 이미지에서 반려견의 자세를 추정하였을 때보다, BAT-Fill을 사용하여 아웃페인팅된 확장 이미지에서 반려견의 자세를 추정하였을 때 자세 추정의 성능이 향상되었다.

• 주제어 : 반려동물 영상, 동물 자세 추정, 가림 현상, 이미지 아웃페인팅, 트랜스포머, 딥러닝

Abstract In recent years, there has been a growing interest in deep learning-based animal pose estimation, especially in the areas of animal behavior analysis and healthcare. However, existing animal pose estimation techniques do not perform well when body parts are occluded or not present. In particular, the occlusion of dog tail or ear might lead to a significant degradation of performance in pet behavior and emotion recognition. In this paper, to solve this intractable problem, we propose a simple yet novel framework for pet pose estimation where pet pose is predicted on an outpainted image where some body parts hidden outside the input image are reconstructed by the image inpainting network preceding the pose estimation network, and we performed a preliminary study to test the feasibility of the proposed approach. We assessed CE-GAN and BAT-Fill for image outpainting, and evaluated SimpleBaseline for pet pose estimation. Our experimental results show that pet pose estimation on outpainted images generated using BAT-Fill outperforms the existing methods of pose estimation on outpainting-less input image.

• Key Words : Pet image, Animal pose estimation, Occlusion, Image outpainting, Transformer, Deep learning

Received 14 March 2023, Revised 20 March 2023, Accepted 28 March 2023

*** Corresponding Author** Wonsang You, 70, Sunmoon-ro 221beon-gil, Tangjeong-myeon, Asan, Chungnam, South Korea.
E-mail: wyou@kaist.ac.kr

Ⅰ. 서론

딥러닝 기반의 자세 추정(pose estimation) 기법 연구는 초기에는 주로 사람 영상을 중심으로 발전되어 왔지만, 최근에는 동물의 행동 분석이나 동물 건강관리 분야에 대한 관심이 높아지면서 동물 영상에서 동물 자세 추정 방법도 연구되고 있다[1]. 기존 동물 자체 추정 기법들은 신체 부위의 일부가 다른 객체에 가려지거나 이미지 내에 존재하지 않을 때 자세 추정의 정확도가 크게 저하된다. 또한, 대부분의 자체 추정 모델이 정답 신체가 존재하지 않을 때는 해당 부위의 예측값의 정보를 무시한다. 영상에서 귀와 꼬리 등 주요 부위의 가림 현상(occlusion)이 발생하는 경우, 정확한 자세 추정에 어려움이 있을 뿐만 아니라 반려견의 행동 또는 감정을 파악하는 기계학습 모델의 성능에도 심각한 영향을 미칠 수 있다. 이러한 가림 현상을 극복하기 위해 다양한 방법이 제안되었지만, 신체 부위가 이미지 밖에 존재하는 경우는 여전히 난제로 남아 있다. 또한, 가림 현상을 극복하는 방법들은 대부분 사람 영상에서 다루어진 반면, 동물 영상에 대한 해결 방법은 논의가 부족한 실정이다.

본 연구는 반려동물의 신체 부위의 일부가 이미지 밖에 있는 경우 이미지 아웃페인팅(image outpainting) 기법을 사용하여 자세 추정의 성능을 향상시키는 단순하면서도 새로운 접근방법을 제안하였다. 제안된 프레임워크는 주어진 영상의 경계 외부 영역을 맥락에 맞게 확장하고, 아웃페인팅된 확장 이미지에서 영상 외부에 존재하는 신체 부위의 위치를 예측하는 것을 가능하게 한다.

제안된 모델은 기존의 비교적 단순한 딥러닝 모델의 결합으로 구성되어 있고 제한된 데이터셋을 사용하여 훈련되었기 때문에 만족할만한 수준의 정량적 성능에 이르지 못하지만, 본 연구의 학술적 기여는 아웃페인팅 기반의 자세 추정 방법의 실현 가능성을 보여주었다는 점에 의의가 있다.

Ⅱ. 관련연구

2.1 이미지 아웃페인팅

이미지 아웃페인팅은 영상복원 기법 중 하나로, 영상 외부의 영역을 자연스럽게 채우는 기술이다. 이미지 인페인팅과 유사하지만, 생성되는 영역이 모든 측면으로 넓어 성능 개선에 어려움이 있다[2]. 생성적 적대 신경망(generative adversarial network, GAN)의 등장 이래로 출력 이미지 품질 향상을 위해 이를 사용한 모델들이 제안되었다. GAN은 인코더-디코더 구조로 이루어지며 평균 제곱 오차와 GAN 손실만을 이용한다[3]. 이후 LSTM(long short term memory) 기반의 RCT(Recurrent Content Transfer) 모델이 제안되었다[4]. 이미지의 외부 영역을 이미지 인페인팅 방법으로 전환하여 경계영역을 재정렬하는 방법도 제안되었다[5].

2.2 동물 자세 추정

최근 동물 자세 추정 기술이 발전하고 있지만, 복잡한 환경에서도 높은 정확도를 가지고 있는 사람 자세 추정에 비하면 성능이 여전히 미흡하다. 동물 자세 추정을 위한 알고리즘의 대부분은 사람 자세 추정 알고리즘을 확장하는 방식으로 구성된다. 데이터의 증대를 위해 합성 데이터를 사용하거나, 사람으로 학습된 데이터를 사용하는 전이 학습 기법도 제안되었다[6-7]. 인코더-디코더에 3D 컨볼루션 레이어를 추가하여 시공간적 피쳐를 추출하여, 연속된 이미지에서 시간에 따른 자세의 움직임을 추적하는 방법도 제안되었다[8].

Ⅲ. 실험방법

제안된 모델은 Fig. 1과 같이 이미지 아웃페인팅 과정과 자세 추정 과정의 두 단계로 구성된다. 이미지 아웃페인팅 과정에서는 입력된 이미지로부터 이미지 아웃페인팅 기법을 통해 이미지 외곽부위를 복원한다. 자세 추정 과정에서는 생성된 아웃페인팅 이미지상에서 반려견의 보이지 않는 신체 부위 위치를 추정한다.

3.1 이미지 아웃페인팅 모델

첫 번째 과정에서는 이미지 아웃페인팅을 통해 이미지의 외곽 영역을 복원하여 이미지 외부에 위치한 반려견의 신체 부위의 텍스처를 복원한다. 본 연구에서는 이미지 아웃페인팅 모델로서, 컨텍스트 인코더 기반의 CE-GAN과 트랜스포머 기반의 BAT-Fill 두 가지 GAN 모델에 대해 실험하였다[10-11].

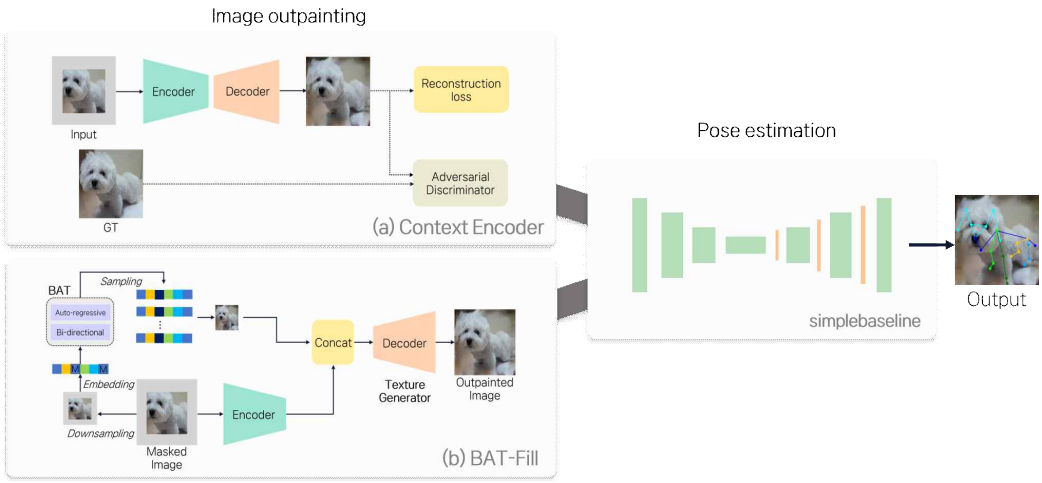

Fig. 1. 제안된 아웃페인팅 기반 자세추정 모델의 네트워크 구조

3.1.1 Context Encoder GAN

CE-GAN은 컨텍스트 인코더(Context encoder) 기반의 생성적 적대 신경망(GAN) 모델이다[10]. 컨텍스트 인코더 기반의 생성자(generator) 네트워크에서는 192×192 크기의 이미지를 아웃페인팅하여 256×256 이미지로 확장한다. 생성자의 컨텍스트 인코더는 마스킹된 입력을 6개의 컨볼루션 레이어를 통해서 반복적으로 다운샘플링한다. 디코더는 디컨볼루션 레이어로 구성되어 컨텍스트 인코더에서 다운샘플링된 정보를 복원한다. 판별자(discri- minator) 네트워크는 실제 이미지 또는 생성된 이미지가 실제일 확률을 추정한다. 판별자는 생성된 부분만 보는 것이 아닌 아웃페인팅된 전체 이미지 영역에 적용된다.

3.1.2 BAT-Fill

BAT-Fill은 양방향 자기회귀 트랜스포머(bidirectional autoregressive transformer, BAT)와 생성적 적대 신경망(GAN) 네트워크로 구성되어 있다[11]. 마스킹된 이미지를 낮은 해상도로 다운샘플링 한 뒤, 다운샘플링된 이미지를 트랜스포머(BAT)에 입력하여 저해상도로 복원한다. 생성자는 복구된 저해상도 이미지 특징을 입력으로 받아 입력 이미지의 훼손되지 않은 영역을 참조하여 고해상도 텍스처를 합성하고 결과를 생성한다.

3.2 2D 자세 추정 모델

반려견의 2차원 자세 추정을 위한 모델로서 SimpleBaseline을 사용하였다. SimpleBaseline은 2차원 사람 자세 추정 모델로서 ResNet을 기반으로 한다[12]. Top-down 방식으로서 객체 검출 후 단일 자세를 추정하는 과정으로 이루어진다. 2D 자세 추정 네트워크는 ResNet의 끝부분에 3개의 디컨볼루션 레이어를 더한 구조이다. 고해상도 피처맵과 각 관절 부위에 대한 히트맵(heatmap)을 얻기 위한 디컨볼루션 레이어는 업샘플링과 컨볼루션 파라미터를 결합하여 구성된다. 관절의 위치는 원본 이미지와 뒤집힌(flip) 이미지의 평균 히트맵에서 예측된다.

3.3 학습 데이터

실험을 위해 동물 데이터세트 Animal- Pose, AP-10K와 반려견 데이터셋 DogPose를 사용하였다. Animal-Pose는 개, 고양이, 소, 말, 양의 카테고리를 가진 4,000장 이상 규모의 데이터셋으로서 총 20개의 키포인트가 정의되어 있다[7]. AP-10K은 23개의 동물 카테고리를 가진 10,015장의 이미지로 구성되며 17개의 키포인트를 가지고 있다[9].

DogPose는 동물 자세 추정 모델 개발을 위해 자체 구축한 반려견 이미지 데이터셋이다. DogPose 데이터셋은 11개의 행동 카테고리(경계, 놀이, 먹이, 배변, 수

면, 훈련 등)로 분류된 반려견 동영상으로부터 추출되었다.

각 동영상에서 추출된 1,240장과 꼬리의 움직임으로 이루어진 33장의 이미지를 더해 총 1,273장의 이미지로 구성되었다. DogPose의 23개 키포인트의 목록이 Table 1에 요약되어 있다.

Table 1. DogPose 데이터셋 키포인트 목록

번호	정의	번호	정의
0	nose	13	left front paw
1	left eye	14	right front paw
2	right eye	15	tailbase
3	left earbase	16	tailend
4	right earbase	17	left knee
5	left earend	18	right knee
6	right earend	19	left ankle
7	throat	20	lright ankle
8	withers	21	left hind paw
9	left elbow	22	right hind paw
10	right elbow		
11	left wrist		
12	right wrist		

※ 빨강색은 Animal-Pose와 AP-10K 모두 공통되는 부분, 파랑색은 Animal-Pose와 공통되는 부분을 의미함.

반려견 자세 추정 모델을 학습하기 위한 데이터로서 Animal-Pose와 AP-10K 데이터세트에서 반려견이 포함된 영상만을 추린 후 DogPose와 함께 통합하여 사용하였다. Animal-Pose는 20개, AP-10K는 17개, DogPose는 23개로 키포인트의 개수가 서로 다르지만, DogPose 키포인트의 포맷으로 재정의하였다. AP-10K 데이터세트에서 left/right hip은 Table 1의 left/Right knee와 동일하게 정의되며, 이미지에 없거나 정답 키포인트가 없는 부위의 경우에는 모두 0의 값으로 일괄 설정하였다.

통합된 데이터세트의 주석은 MS COCO 형식으로 가공하였다. COCO 주석은 이미지 이름과 크기, ID뿐만 아니라 반려견 개체의 경계 상자 좌표와 신체 부위의 가시(hidden or not) 정보를 포함한다.

3.4 성능 지표

이미지 아웃페인팅에 대한 정량적 평가 지표로 최대 신호 대 잡음 비(peak signal-to noise ratio, PSNR), 구조적 유사도(structural similarity index measure, SSIM)을 사용한다. PSNR과 SSIM 산출 방법 식은 (1), (2)와 같다.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad (1)$$

$$SSIM(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \quad (2)$$

2차원 반려견 자세 추정 결과의 정량적 평가 지표는 평균 정밀도(average precision, AP)와 평균 재현율(average recall, AR)를 사용한다. 자세 추정에서의 평균 정밀도와 재현율 계산에는 식 (3)과 같이 정의되는 객체 키포인트 유사도(object keypoint similarity, OKS)를 사용한다.

$$OKS = \sum_i [\exp(\frac{-d_i^2}{2s^2k_i^2})\delta(v_i > 0)/\sum_i \delta(v_i > 0)] \quad (3)$$

식(3)에서 d_i는 정답 키포인트와 추정된 키포인트 사이의 유클리디안 거리이고 v_i는 정답의 가시성 여부(visibility flag), s는 객체 분할 영역의 제곱근이며 k_i는 상수이다. AP.5는 OKS에서 임곗값 0.5, AP.75는 임곗값이 0.75이다.

Ⅳ. 실험결과

4.1 실험환경

실험 과정에서 사용된 영상은 총 3,782장으로 학습 3,000장, 검증 300장, 평가 482장으로 나뉘어 사용되었다. 영상은 모두 반려견의 경계 상자 크기로 크롭되었으며, 크롭된 이미지는 모두 256×256 크기로 재조정되어 이미지 아웃페인팅 네트워크에 사용하였고, 반려견 자세 추정 네트워크의 입력으로 아웃페인팅 결과 영상이 사용되었다.

반려견 자세 추정 네트워크는 ImageNet으로 사전 학습된 ResNet101을 백본으로 하여 실험되었다. 학습 데이터의 증대를 위해 이미지 회전 및 확대 등의 데이터 증강 기법이 적용되었다. Context encoder와 BAT-Fill 모델에서 배치 크기(batch size)는 128, 최적화 기법으로는 Adam, 학습률은 0.001, 반복 횟수(epoch)는 140번으로 설정하였다.

4.2 아웃페인팅 성능평가

482장의 테스트셋을 사용하여 이미지 아웃페인팅 모델의 성능을 비교한 결과는 Table 2와 같다, BAT-Fill의 아웃페인팅 결과가 Context encoder보다 높은 수치를 보인다(Context encoder GAN vs. BAT-Fill: PSNR= 30.58<31.04; SSIM=0.63<0.71).

Table 2. 이미지 아웃페인팅의 정량적 평가결과

Method	PSNR	SSIM
Context encoder	30.58	0.63
BAT-Fill	**31.04**	**0.71**

Fig. 2는 이미지 아웃페인팅의 정성적 결과를 보여준다. Context encoder에 비하여 BAT-Fill을 사용하였을 때 신체 부위가 더욱 잘 복원되고 있음을 알 수 있다. 그러나 원본 이미지와 비교해서 이미지 복원이 여전히 정확하지 못하고 시각적 품질도 충분하지 않음을 보여준다.

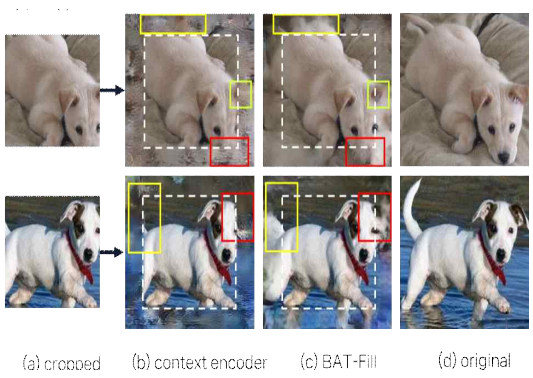

Fig. 2. 아웃페인팅 결과의 정성적 비교

4.3 자세추정 성능평가

Fig. 3은 크롭한 영상을 입력하여 아웃페인팅하고, 확장된 이미지에서 추정된 반려견 자세 키포인트를 정답 키포인트(GT)와 비교한 예시를 보여준다. 아웃페인팅된 이미지에서 반려견의 자세를 추정한 결과는 원본 이미지의 정답 키포인트와 유사함을 확인할 수 있다.

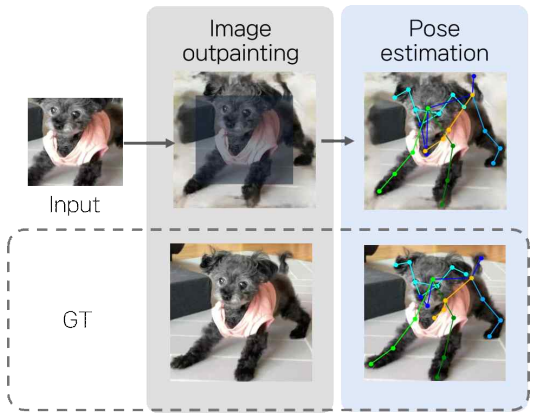

Fig. 3. 아웃페인팅 기반 반려견 자세추정 예시

Table 3은 CE-GAN 및 BAT-Fill의 두 아웃페인팅 모델을 사용하여 복원된 확장 이미지에 대하여 SimpleBaseline 모델을 사용하여 반려견의 자세를 추정한 결과에 대한 정량적 평가 결과를 요약한다.

Table 3. 2D 반려견 자세추정 결과의 정성적 비교

Data	Outpaint	AP	AP.5	AP.75	AR	AR.5	AR.75
GT	none	0.713	0.921	0.761	0.765	0.932	0.805
Crop	none	0.382	0.777	0.338	0.499	0.824	0.500
	CE-GAN	0.453	0.783	0.461	0.544	0.822	0.564
	BAT-Fill	0.564	0.854	0.592	0.643	0.877	0.663

정답 이미지(GT)에서 아웃페인팅 없이 자세 추정한 성능을 기준 지표로 삼았을 때, 크롭한 이미지로부터 아웃페인팅된 이미지의 평균 정밀도(AP)와 평균 재현율(AR)이 기준지표에 도달하지는 못하였다. 다만, BAT-Fill 모델을 사용하여 아웃페인팅한 이미지에서 반려견의 자세를 추정한 결과는 아웃페인팅 없이 자세 추정한 경우나 CE-GAN 아웃페인팅 기반의 자세 추정한 경우에 비하여 뚜렷하게 높은 성능을 보였다.

Fig. 4는 자세 추정 결과의 모델간 정성적 비교를 보여준다. CE-GAN보다 BAT-Fill을 사용했을 때 정성적 자세 추정의 품질이 향상되었음을 확인할 수 있다. 이는 아웃페인팅의 전처리 과정이 2차원 자세 추정의 성능을 향상시키고 있음을 보여준다.

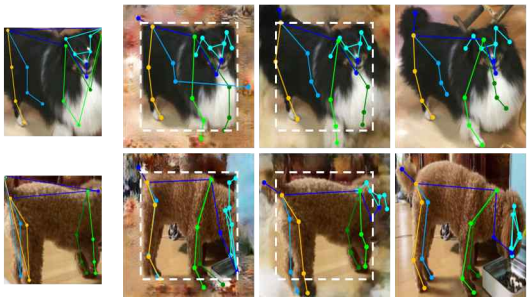

Fig. 4. 아웃페인팅 기반 반려견 자세추정 결과 비교

Ⅴ. 결론

본 연구에서는 이미지 아웃페인팅 기법을 사용하여 이미지 외부의 보이지 않는 반려동물 신체 부위를 예측함으로써 자세 추정의 성능을 향상시키는 기법을 제안하고 제안된 방법의 실현 가능성을 테스트하였다.

실험 결과, 아웃페인팅된 이미지에 자세 추정 방법을 적용하였을 때 아웃페인팅을 하지 않고 원래의 크롭된 입력 이미지에서 자세 추정을 하였을 때보다 성능이 향상되었다. 특히, CE-GAN보다 트랜스포머 기반의 BAT-Fill 모델을 사용하여 아웃페인팅하였을 때 더욱 좋은 성능을 보였다.

하지만, 제안된 방법은 자세 추정의 성능이 이미지 아웃페인팅 네트워크의 결과에 따라 크게 좌우된다는 단점이 있다. 아웃페인팅 네트워크와 자세 추정 네트워크를 종단 간 학습 방법으로 훈련하는 것도 이러한 문제를 완화시킬 수 있다.

본 연구는 비교적 단순한 아웃페인팅 및 자세 추정 모델을 사용하여 아웃페인팅 기반의 자세 추정의 실현 가능성을 테스트하는데 목적이 있다. 향후 본 예비 연구를 바탕으로, 다양한 가려짐 상황에서 효과적으로 반려동물의 자세를 추정할 수 있도록 새로운 구조의 아웃페인팅 기반 자세 추정 모델을 제안하고 발전시켜 나갈 계획이다.

ACKNOWLEDGMENTS

본 연구는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 기본연구(2022R1F1A1075204), 4단계 두뇌한국21 사업(4단계 BK21 사업) 및 지자체-대학 협력기반 지역혁신사업(2022RIS-004), 중소기업벤처부의 재원으로 수행된 2021년도 창업성장기술개발사업(S3228660)의 연구결과로 수행되었음.

REFERENCES

[1] M. Chae, C. Lee, “Implementation of Prevention and Eradication System for Harmful Wild Animals Based on YOLO,” The Journal of Korea Institute of Convergence Signal Processing, vol.23, no.3, pp. 137-142, 2022.

[2] P. Gao, X. Yang, R. Zhang, J. Y. Goulermas, Y. Geng, Y.Yan, K. Huang, (2023). “Generalised image outpainting with U-Transformer.” Neural Networks, 162, pp.1-10.

[3] M., Sabini, G. Rusak, “Painting outside the box: Image outpainting with gans.” arXiv arXiv:1808.08483, 2018

[4] Z. Yang, J. Dong, P. Liu, Y. Yang, S. Yan, “Very long natural scenery image prediction by outpainting,” in Proc. IEEE/CVF, 2019, pp. 10561-10570.

[5] K.. Kim, Y. Yun, K.-W. Kang, K. Kong, S. Lee, S.-J. Kang, “Painting outside as inside: Edge guided image outpainting via bidirectional rearrangement with progressive step learning,” in Proc. IEEE/CVF, 2021, pp. 2122-2130.

[6] C. Li, G. H. Lee, “From synthetic to real: Unsupervised domain adaptation for animal pose estimation,” in Proc. IEEE/CVF, 2021, pp. 1482-1491.

[7] J. Cao, H. Tang, H. S. Fang, X. Shen, C. Lu, & Y. W. Tai, “Cross-domain adaptation for animal pose estimation,” in Proc. IEEE/CVF, 2019, pp. 9498-9507.

[8] H. Russello, R. van der Tol, & G. Kootstra, (2022). “T-LEAP: Occlusion-robust pose estimation of walking cows using temporal information.” Computers and Electronics in Agriculture, 192, 106559.

[9] H. Yu, Y. Xu, J. Zhang, W. Zhao, Z. Guan, & D. Tao, “Ap-10k: A benchmark for animal pose estimation in the wild,”, NeurIPS 2021 Datasets and Benchmarks Track (Round 2), 2021.

[10] B. V. Hoorick, “Image outpainting and harmonization using generative adversarial networks” arXiv arXiv:1912. 10960, 2019.

[11] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, Miao, “Diverse image inpainting with bidirectional and autoregressive transformers.” in Proc. ACM International Conference on Multimedia, 2021, pp. 69-78.

[12] B. Xiao, H. Wu, & Y. Wei. “Simple baselines for human pose estimation and tracking.” in Proc. ECCV, 2018, pp. 466-481.

저자소개

이 규 빈 (Gyubin Lee)

2023년 2월 : 선문대학교 정보통신공학과(공학사)
2023년 3월~현재 : 선문대학교 정보통신공학과(석사과정)
관심분야 : 영상처리, 딥러닝

이 영 찬 (Youngchan Lee)

2022년 2월 : 선문대학교 정보통신공학과(공학사)
2022년 3월~현재 : 선문대학교 정보통신공학과(석사과정)
관심분야 : 영상처리, 딥러닝

유 원 상 (Wonsang You)

2003년 서강대 전자공학과 학사
2008년 KAIST 공학석사
2013년 독일 마그데부르크대 박사
2009-2012 라이프니츠 연구소
2013-2019 미국국립아동병원
2020년~현재 선문대학교 정보통신공학과 조교수
관심분야 : 영상처리, 인공지능, 의료영상, 뇌공학