논문 2024-19-02

3차원 자세 추정 기법의 성능 향상을 위한 임의 시점 합성 기반의 고난도 예제 생성

(Hard Example Generation by Novel View Synthesis for 3-D Pose Estimation)

김 민 지, 김 성 찬*
(Minji Kim, Sungchan Kim)

Abstract : It is widely recognized that for 3D human pose estimation (HPE), dataset acquisition is expensive and the effectiveness of augmentation techniques of conventional visual recognition tasks is limited. We address these difficulties by presenting a simple but effective method that augments input images in terms of viewpoints when training a 3D human pose estimation (HPE) model. Our intuition is that meaningful variants of the input images for HPE could be obtained by viewing a human instance in the images from an arbitrary viewpoint different from that in the original images. The core idea is to synthesize new images that have self-occlusion and thus are difficult to predict at different viewpoints even with the same pose of the original example. We incorporate this idea into the training procedure of the 3D HPE model as an augmentation stage of the input samples. We show that a strategy for augmenting the synthesized example should be carefully designed in terms of the frequency of performing the augmentation and the selection of viewpoints for synthesizing the samples. To this end, we propose a new metric to measure the prediction difficulty of input images for 3D HPE in terms of the distance between corresponding keypoints on both sides of a human body. Extensive exploration of the space of augmentation probability choices and example selection according to the proposed distance metric leads to a performance gain of up to 6.2% on Human3.6M, the well-known pose estimation dataset.

Keywords : Human pose estimation, Novel view synthesis, Occlusion

Ⅰ. 서 론

인간의 자세를 추정하는 문제 (Human pose estimation, HPE)는 증강현실 (Augmented reality, AR), 모니터링, 무인 감시를 포함한 다양한 영역에 적용될 수 있어 많은 관심을 받고 있다. HPE의 핵심은 머리, 손목, 골반과 같은 인체의 중요한 부위인 키포인트 (keypoint)를 정확하게 찾아내는 것이다. 현재까지 HPE 기법들은 최신 최적화 및 특징 추출기법들로 인해 괄목할 만한 성능 향상을 이루었지만, 여전히 조도, 크기와 같은 조건과 복잡한 신체 구조로 인해 어려운 문제로 인식된다. 이는 시각적인 정보만으로는 인체의 복잡하고 다양한 굴곡과 움직임을 포괄적으로 판단하기 어려운 경우가 있기 때문이다.

HPE의 성능 저하는 라벨링의 부정확성, 폐색 (occlusion) 등으로 인한 내부적인 요인 및 의상, 피사체 크기, 성별 등의 외부적인 요인에 의해 발생한다. 특히 폐색은 카메라의 시야에 신체 일부만 포착되는 경우, 물체가 특정 신체 부위나 팔다리를 가리는 경우, 다리를 꼬는 등 특정 신체 부위가 다른 신체 부위를 가리는 경우 (self-occlusion) 등의 상황에서 발생하며 해결하기 어려운 문제로 남아있다. 폐색으로 인한 성능의 저하는 개발된 기법들이 폐색을 고려하지 않거나, 현재 HPE 분야에서 사용하는 대부분의 데이터세트에서 폐색이 발생하는 경우가 부족해 발생한다. 최근 [1]은 HPE 기법의 성능 향상을 위해 적대적 데이터 증강 방식을 제안했으며, [2, 3]은 애니메이션을 활용하여 다양한 시점에서 촬영된 데이터세트를 개발했다. [4]는 인체의 keypoint 주변의 배경 일부를 복사하는 마스킹 기법을 제안하여 폐색이 발생하는 문제를 해결했으며, [4]를 확장한 [5]는 적대적 패치의 활용을 제안하였다. 하지만 이러한 기법들은 자연스럽지 않은 이미지들을 생성할 수 있다는 한계점이 존재한다.

HPE 데이터세트를 구축하기 위해서는 큰 비용이 필요하기 때문에 데이터 증강 방식을 적용하는 시도 [1-5]들이 있지만, 기존의 데이터 증강 방식을 사용하는 경우 폐색에 대한 HPE의 성능 개선 효과가 제한적이라는 것이 실험적으로 알려졌다 [6]. 이와 달리 제안하는 기법은 자연스럽고 사실적인 이미지를 임의시점합성 (novel view synthesis)을 통해 생성하여 활용하고자 한다.

본 논문에서는 3차원 HPE 모델의 폐색과 관련된 성능

*Corresponding Author (s.kim@jbnu.ac.kr)
Received: Dec. 6, 2023, Revised: Dec. 11, 2023, Accepted: Dec 20, 2023.
M. Kim: The Department of Computer Science and Artificial Intelligence, Jeonbuk National University (Ph.D. Student)
S. Kim: The Department of Computer Science and Artificial Intelligence, Jeonbuk National University; Center for Advanced Image Information Technology, Jeonbuk National University (Prof.)
※ 이 논문은 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1A2C1011013).

ⓒ IEMEK J. Embed. Sys. Appl. 2024 Feb. 19(1) 9-17
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2024.19.1.9

향상을 위해 다양한 카메라 시점의 데이터를 활용하여 데이터를 증강하는 간단하면서도 효과적인 방법을 제안한다. 이때, 관찰되지 않은 카메라 시점의 이미지를 생성하기 위해 [7]에서 제안한 오토인코더 (autoencoder) 기반의 임의시점 합성 기법을 사용한다. 또한, 본 논문에서는 인체의 양쪽에 대칭적으로 존재하는 keypoint가 근접한 경우 자기폐색 (self-occlusion)이 자주 발생한다는 점을 활용하여 폐색이 발생하는 어려운 예제를 생성하는 기법을 제안한다.

본 연구의 의의는 다음과 같다.

1) 특정 신체 부위가 다른 신체 부위를 가리는 자기폐색이 발생하는 어려운 예제를 생성하기 위한 keypoint 기반의 방식을 제안하여 3차원 HPE 모델을 위한 데이터세트를 효과적으로 증강한다.

2) 기존 학습 과정을 수정하지 않고도 HPE 모델의 학습 과정에 쉽게 통합될 수 있는 데이터 증강 기법을 제안한다.

3) 제안한 기법을 다양한 증강 확률과 증강 기준을 기반으로 평가하여 추가 데이터 샘플을 활용하지 않고도 기존 데이터세트에 비해 MPJPE (mean per joint position error)를 최대 6.2%까지 크게 개선하였다.

Ⅱ. 관련 연구

1. 인간 자세 추정 (Human Pose Estimation, HPE)

초기 HPE 연구에서는 영상 내에서 인간을 감지하고 상체 자세를 추정하기 위해 그림 구조 [8, 9]를 사용했다. 하지만 그래프 라벨링의 추론이 어려워 복잡도가 증가하는 문제가 발생했다. [10]에서는 트리 구조의 그래프 형태를 사용한 Deformable Part Model (DPM)을 제안하여 폐색이 발생하지 않는 예제에서 높은 성능을 보였지만, 변수 간의 상관관계를 인지하지 못하는 사례가 발생했다. 특히, 트리 구조의 모델은 종종 왼쪽과 오른쪽의 keypoint를 혼동하는 경우가 발생하는데, 이 문제는 3차원 HPE의 경우 더욱 두드러지게 나타난다.

이후, HPE 데이터세트를 구축하기 위한 노력과 심층 신경망의 발전으로 HPE의 성능은 크게 향상되었다. [11]은 다중 시점을 사용하여 자세를 추정하는 방법을 제안했으며, [12]는 컨볼루션 신경망 (convolutional neural networks, CNNs)을 사용하여 사람의 관절 위치를 추정하였다. [13]은 자세 추정을 위한 multi score deep model을 제안했지만, 정확도가 높지 않았고, 이를 해결하기 위해 [14]는 학습 중 keypoint를 heatmap으로 표현하였다. 후속 연구 [15-20]는 대부분 더 큰 수용장 (receptive field)을 가진 심층 신경망을 기반으로 heatmap에서 자세 매칭을 수행하는 기법을 제시하였다. 이러한 접근 방식은 keypoint 간의 관계를 더 잘 학습하고 다양한 크기에서 능숙하게 자세를 추정하는 특징을 보인다 [21, 22]. 특히, [21]은 속도에서 좋은 성능을 보였으며, [22]는 크기의 다양성으로 인해 발생하는 문제를 해결하기 위해 고해상도와 저해상도 이미지를 모두 활용하는 기법을 제안하였다.

HPE 기법은 주로 하향식 (top-down)과 상향식 (bottom-up)으로 분류할 수 있다. 하향식은 이미지에서 먼저 사람 인스턴스를 찾은 후, keypoint를 추정한다. 이 방식은 정확도가 높고 확장성의 문제가 발생할 가능성이 적지만, 일반적으로 계산량이 많아 비실용적이다. 반면, 상향식은 이미지에서 모든 keypoint를 추정한 후, 이를 각각의 사람 인스턴스로 그룹화한다. 이 방식은 속도가 빨라 실시간 애플리케이션에 더 적합하지만, 사람 인스턴스의 크기가 다양한 경우 하향식보다 성능이 낮은 특징을 보인다.

2. 임의시점합성 (Novel View Synthesis, NVS)

NVS는 주어진 이미지에 대해 보이지 않는 카메라 시점에서 촬영된 새로운 이미지를 합성하는 것을 목표로 한다. 초기의 NVS는 주로 자동차나 가구와 같은 단순하고 외형의 변화가 없는 물체를 생성하는 데 중점을 두었다 [23-28]. 최근에는 인코더-디코더 (encoder-decoder) 구조에 기반을 둔 기법들이 제안되고 있는데, 이 구조에서는 잠재 공간에 회전각도 등의 정보를 통합하고 디코더가 새로운 시점에서 인코딩된 이미지를 합성하도록 학습된다 [29, 30]. 이외에도 [31]에서는 대규모의 시점 변환을 수행하기 위하여 recurrent network를 사용한 방법을 제안했으며, 옵티컬 플로우 (optical flow) [25, 28]와 깊이 맵 (depth map) [32]도 성능 개선을 위해 사용된다. [33]은 transforming autoencoder를 사용하여 다양한 시점의 이미지를 생성하였고, [34]는 3D-2D 투영 (projection)을 통해 2차원 정보를 활용하여 3차원 정보를 학습하는 프레임워크를 제안했다. [35]는 perspective Transformer를 이용하여 물체의 3차원 정보를 재구성하였다. [36]은 일련의 어텐션 (attention) 모듈을 포함하는 적대적생성망 (generative adversarial network)을 이용하여 자세 전이 기법을 제안하였다. [37]은 복셀 (voxel)을 이용하여 폐색에 초점을 맞춘 3차원 구조를 표현했다. 최근에는 자세의 전이와 사람의 이미지를 생성하는 연구들도 다양하게 이루어지고 있다 [36, 38-43]. 그러나 이러한 방법들은 어려운 자세를 합성하는 데 우선순위를 두지 않고, 사전에 정의된 자세로 이미지를 생성하는 한계점이 존재한다.

본 논문에서 제안하는 기법은 NVS 기법을 이용하여 보이지 않는 시점의 어려운 자세 데이터를 합성하고 이를 HPE 모델에 대한 추가 학습 예제로 사용하여 폐색 문제를 해결한다.

Ⅲ. 제안하는 기법

입력 이미지에서 3차원 keypoint를 예측하는 과정은 두 단계로 진행된다 [44-47]. 이미지가 입력되면, 첫 번째 단계에서는 입력 이미지의 2차원 keypoint를 추출한다. 두 번째 단계에서는 첫 번째 단계에서 추출된 2차원 keypoint를 입력받아 3차원 keypoint로 변환하기 위해 깊이 정보를 재구성한다.

그림 1은 제안하는 기법인 예측하기 어려운 합성된 이미지를 활용하여 학습 시 HPE 모델의 미니배치 (minibatch) 증강을 수행하는 전반적인 절차를 나타낸다. 제안하는 기법에서는 미니배치가 생성될 때마다 일정 확률에 따라 미니배치의 학습 샘플을 어려운 예제로 대체한다. 어려운 예제는

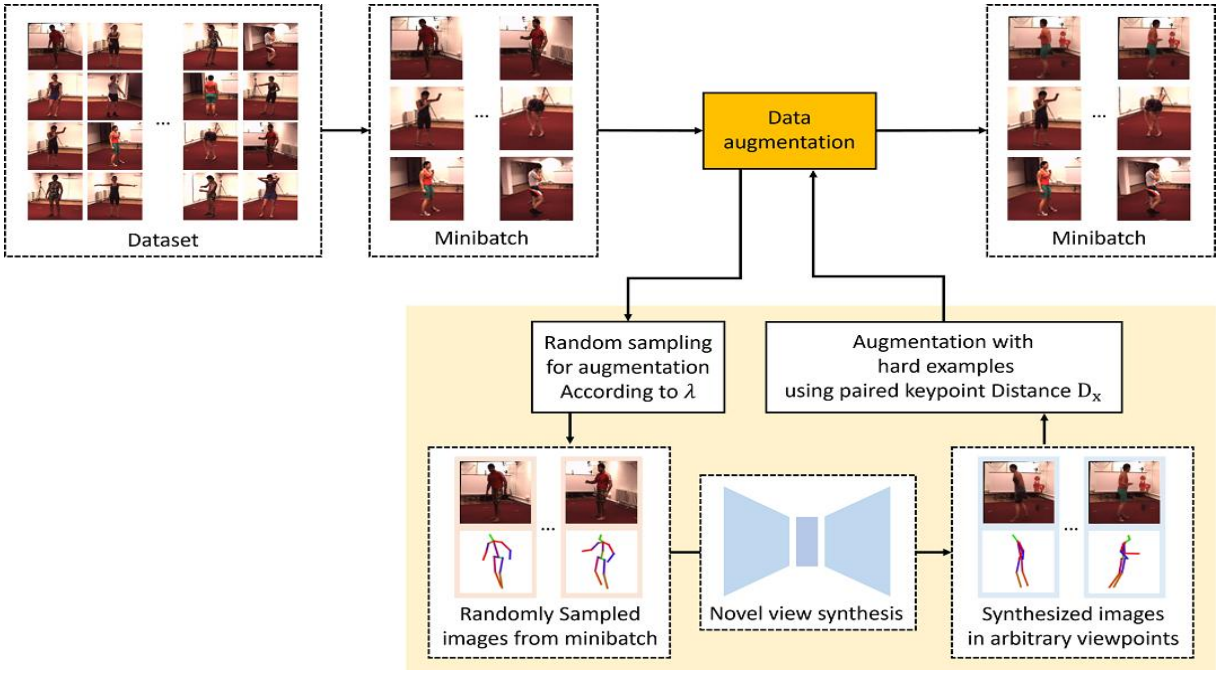


그림 1. 본 논문에서 제안하는 어려운 예제를 이용한 데이터 증강 방법
Fig. 1. Overall procedure of the proposed method for data augmentation using hard examples

예측 난이도를 측정하는 기준에 따라 NVS 기법을 통해 합성된 예제에서 선택된다.

1. 임의시점합성

본 논문에서 제안하는 기법은 합성된 어려운 자세의 이미지로 데이터세트를 보강하여 3차원 keypoint의 예측을 개선하는 것을 목표로 한다. 이미지에서 3차원 keypoint를 예측하는 과정은 두 단계로 진행되므로, 본 논문에서는 다양한 시점에서 합성된 이미지와 이에 해당하는 2차원 keypoint 데이터를 활용한다. 본 논문에서는 기존의 NVS 기법 [7]을 사용하여 주어진 입력 이미지에 대해 관찰되지 않은 시점의 이미지를 생성한 후, 사전에 정의한 기준에 따라 어려운 자세의 이미지를 선택한다. NVS 기법 [7]은 여러 시점에서 물체의 모양과 기하학적 정보를 모두 포함하는 잠재 공간을 학습하여 원본 데이터세트에 존재하지 않는 임의의 시점에서의 데이터를 합성한다.

본 논문에서 사용하는 NVS 기법 [7]은 인코더 $\theta_E : \mathbb{R}^{128\times128} \rightarrow \mathbb{R}^{512\times16\times16}$ 와 디코더 $\theta_D : \mathbb{R}^{512\times16\times16} \times \mathbb{R}^{128} \rightarrow \mathbb{R}^{128\times128}$ 로 구성된 오토인코더 구조를 사용하여 잠재 공간을 학습하고, 인스턴스의 가능한 모든 시점의 이미지를 합성하는 것을 목표로 한다. 학습되는 잠재 공간은 인스턴스의 외형 벡터 $v_a \in \mathbb{R}^{128}$ 와 기하학적 정보를 담은 벡터 $v_{3D} \in \mathbb{R}^{200\times3}$ 로 구성된다. NVS 모델은 인스턴스를 동시에 다양한 시점에서 촬영한 다중시점 데이터세트를 사용하여 학습된다. 이 NVS 모델은 azimuth (z축) R_z, elevation (y축) R_y, optical (x축) R_x의 세 축으로 구성된 회전 매트릭스 R과 표준 카메라 모델을 사용하여 다양한 시점의 이미지를 합성한다.

회전 매트릭스 R을 사용하여 시점 i에서 시점 j로 변환하는 과정을 $R^{i\rightarrow j}$라고 하면, 시점 i에서 촬영된 이미지 $x_i \in \mathbb{R}^{128\times128}$는 $R^{i\rightarrow j}$을 통해 다음과 같이 시점 j에서의 이미지 $\hat{x}_j$로 합성된다.

$$\hat{x}_j = \theta_D(\theta_E(x_i)) = \theta_D(R^{i\rightarrow j} \cdot v_{3D}, v_a). \quad (1)$$

인코더 θ_E는 이미지 x_i를 입력으로 받아 입력의 특징을 담은 잠재 공간 $\theta_E(x_i)$를 추출한다. 추출된 잠재 공간 $\theta_E(x_i)$는 외형 벡터 v_a와 기하학적 정보를 담은 벡터 v_{3D}로 나뉜다. 이후, v_{3D}는 $R^{i\rightarrow j}$을 이용해 시점 i에서 시점 j로 변환된다. 디코더 θ_D는 입력으로 v_a와 변환된 v_{3D}를 받아 시점 j에서의 이미지 $\hat{x}_j$를 합성한다.

2. 고난도 자세 영상 생성

인체 일부가 물체에 가려지는 폐색이나 인체의 한 부분이 다른 부분을 가리는 자기폐색은 HPE에서 keypoint를 정확하게 예측하는 것을 어렵게 만든다. 본 논문에서 제안하는 방법은 자기폐색이 발생하는 어려운 자세가 팔꿈치와 무릎과 같이 인체의 양쪽에 대칭적으로 존재하는 keypoint들이 서로 근접했을 때 발생한다는 점을 활용했다. 그림 2는 간단한 자세를 z축을 중심으로 회전시켰을 때, 인체의 양쪽에 대칭적인 keypoint들이 가까워지면서 예측이 더 어려운 자기폐색이 발생하는 자세로 변형되는 예시를 보여준다. 따라서, 본 논문에서는 주어진 자세에서 keypoint 예측의 난이도

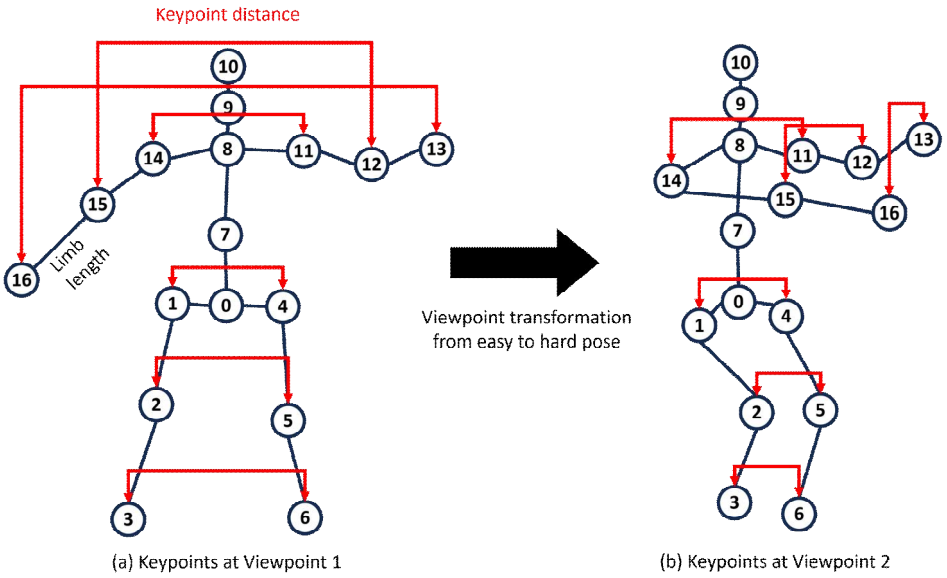


그림 2. 17개의 keypoint를 사용하는 사람 인스턴스
Fig. 2. Pictorial description of a human instance using 17 keypoints where limbs are represented by the black lines connecting corresponding keypoints

를 측정하기 위하여 (1, 4), (2, 5), (3, 6), (11, 14), (12, 15), (13, 16)과 같이 인체의 양쪽에 대칭적으로 존재하는 keypoint 쌍의 거리를 사용하는 방법을 제안한다.

입력 이미지 x가 주어졌을 때, 인체의 양쪽에 대칭적으로 존재하는 keypoint 쌍을 $S_x = \{(p_i, p_j)\}$라고 정의한다. 이때, S_x는 어깨, 팔꿈치, 손목, 골반, 무릎, 발목과 같이 인체의 양쪽에 대칭적으로 존재하는 6개의 keypoint 쌍으로 구성되며, p_i와 p_j는 각각의 keypoint를 나타낸다. 입력 영상 x가 어려운 자세를 포함할 가능성을 측정하는 기준은 S_x의 keypoint 사이의 평균 거리 D_x를 통해 계산되며, 이는 다음과 같이 정의된다.

$$D_x = \frac{1}{|S_x|} \sum_{(p_i p_j) \in S_x} \| p_i - p_j \| . \quad (2)$$

어려운 자세를 포함한 예제를 합성하기 위해 3차원 좌표 공간에서 임의의 축을 중심으로 입력 이미지 x를 회전시킬 때, 합성된 이미지들이 항상 자연스럽게 보이지는 않는다. 예를 들어 이미지를 x축이나 y축을 중심으로 회전하면 사람이 거꾸로 서 있는 합성 이미지를 얻을 수 있는데, 이는 데이터세트의 자세 분포와 일치하지 않을 가능성이 크다. 따라서 제안하는 방법에서는 이미지를 합성할 때, z축을 중심으로 한 회전만 고려하여 자연스러운 이미지를 합성하였다.

3. 구현 세부 사항

3.1 임의시점에서의 학습 데이터 생성

제안하는 기법에서는 [7]의 방법을 사용하여 다중시점의 이미지를 합성하며 HPE 모델의 평가를 위해 대표적인 다중시점 데이터세트인 Human3.6M [48]을 사용한다. Human3.6M 데이터세트의 학습 샘플에 포함된 사람 인스턴스는 항상 이미지의 중앙에 위치하지 않기 때문에, 주어진 샘플을 z축을 중심으로 회전시켜 새로운 시점에서의 이미지를 합성하는 경우 자연스러운 이미지를 합성할 수 없다. 이를 위해 본 논문에서는 인스턴스의 양쪽 골반 keypoint의 중심으로 정의되는 루트 키포인트 (root keypoint)가 이미지 좌표 공간의 원점에 위치하도록 샘플의 좌표를 조정하여 사용하였다. 또한, 학습의 효율성을 위해 Human3.6M 데이터세트의 4개의 다중시점 중 하나의 시점인 i에서 획득된 이미지 x_i만을 사용해 임의시점에서의 데이터를 오프라인으로 생성한다. 생성된 데이터는 시점 i에서 반시계 방향으로 10°씩 회전한 35개의 새로운 시점의 이미지를 합성하여 구성된다.

3.2 HPE 모델 학습

다음 단계는 HPE 모델 학습을 위한 증강 데이터세트를 구축하는 것이다. 제안하는 기법에서는 학습 샘플을 관찰되지 않은 시점의 어려운 샘플로 변환하여 원본 샘플보다 예측을 더 어렵게 만드는 방식으로 학습 샘플을 보강한다. 이를 위해 같은 인스턴스에서 keypoint 쌍의 거리 D_x가 원본 이미지보다 작은 임의의 시점에서의 합성 이미지들을 증강된 학습 샘플로 사용한다. 만약 원본 이미지보다 D_x가 작은 합성 이미지가 없는 경우, 학습 중에 합성 이미지로 대체하지 않고 원본 이미지를 그대로 사용한다. 이때, 원본 이미지를 어려운 예제로 대체하는 비율은 사전에 정의된 확률인 하이퍼 파라미터 (hyperparameter) λ에 따라 결정한다. 학습 샘플을 구성하는 인스턴스의 root keypoint를 이미지 좌표의 원점으로 정렬했기 때문에 테스트 샘플도 같은 정렬방식을 적용하였다.

표 1. Human3.6M 데이터세트에서 제안하는 기법의 정량적 성능 비교
Table 1. Results of an ablation study of the sampling strategy and the probability of augmenting hard examples in the Human3.6M dataset

Aug.Method		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Avg
No aug.		25.7	28.9	27.7	29.2	31.6	32.5	28.2	33.7	36.4	31.0	40.1	31.7	24.6	31.7	28.1	30.7
Part. Rot.	0.1 - alt - aug - 1	26.4	**28.7**	29.3	31.1	32.5	33.0	28.2	35.7	37.8	31.1	42.2	31.9	26.1	32.6	31.0	31.8
	0.3 - alt - aug - 1	**25.4**	29.3	28.7	30.7	32.9	**31.3**	28.4	34.7	39.3	31.7	41.2	32.0	**24.2**	33.3	28.2	31.4
	0.5 - alt - aug - 1	26.9	30.8	29.2	31.8	34.6	33.7	28.5	35.0	43.7	33.8	43.9	33.9	27.1	34.4	30.8	33.2
	0.1 - alt - aug - 2	**24.6**	**28.1**	27.8	**28.1**	**31.1**	**30.2**	**25.5**	**33.2**	37.3	**29.5**	40.8	**30.0**	**23.3**	**30.9**	**26.6**	**29.8**
	0.3 - alt - aug - 2	26.3	29.6	29.3	30.2	32.0	**31.9**	29.6	36.1	39.7	31.5	41.6	32.2	24.7	32.8	**27.9**	31.7
	0.5 - alt - aug - 2	27.4	30.7	32.2	32.6	36.7	33.0	30.7	37.5	43.1	33.3	46.1	34.1	27.5	35.4	31.4	34.1
	0.1 - proposed	26.5	**27.9**	29.6	30.6	31.7	**31.7**	**27.1**	34.6	**36.0**	**29.8**	41.1	**30.9**	25.4	**31.0**	28.8	30.9
	0.3 - proposed	**24.4**	**27.3**	28.8	**28.4**	**31.2**	**30.6**	**25.5**	36.0	37.2	**30.1**	**39.4**	**30.2**	25.2	**30.8**	28.5	**30.2**
	0.5 - proposed	**23.7**	**27.1**	**26.9**	**27.9**	**30.3**	**30.5**	**24.7**	**32.4**	36.4	**29.0**	**38.0**	**29.2**	**23.2**	**30.3**	**26.5**	**29.1**
Full Rot.	0.1 - alt - aug - 1	**24.0**	**27.3**	**27.6**	**27.5**	**30.7**	**30.5**	**25.4**	34.4	38.0	**29.7**	**38.5**	**29.5**	**23.2**	**29.8**	28.2	**29.6**
	0.3 - alt - aug - 1	**23.6**	**27.9**	**26.0**	**27.0**	**30.4**	**28.9**	**27.0**	33.7	37.0	**29.4**	**38.1**	**29.7**	**23.3**	**30.2**	**26.5**	**29.2**
	0.5 - alt - aug - 1	26.4	29.8	33.4	30.5	33.0	**31.8**	28.6	36.7	41.3	32.8	43.2	33.9	27.4	34.7	31.2	33.0
	0.1 - alt - aug - 2	26.3	29.6	30.2	30.4	33.3	**31.0**	29.5	35.1	38.8	32.0	42.4	**31.2**	25.2	32.2	**27.6**	31.7
	0.3 - alt - aug - 2	25.8	**28.8**	30.0	30.0	33.4	32.7	**28.1**	34.9	38.6	32.0	42.2	32.8	26.6	32.0	30.1	31.9
	0.5 - alt - aug - 2	25.8	29.6	30.1	31.1	33.4	**31.6**	29.7	35.6	41.7	33.1	43.8	33.7	27.3	35.1	30.5	32.8
	0.1 - proposed	**24.6**	**27.2**	**26.6**	**28.8**	**30.4**	**29.9**	**27.7**	**33.4**	37.0	**29.2**	**37.6**	**30.1**	**23.1**	**29.1**	**26.5**	**29.4**
	0.3 - proposed	**23.2**	**26.6**	**26.0**	**27.6**	**30.2**	**30.1**	**24.6**	**33.1**	**35.0**	**29.6**	**37.2**	**30.3**	**23.6**	**29.6**	**25.6**	**28.8**
	0.5 - proposed	**23.8**	**28.1**	27.8	**28.2**	31.8	**29.7**	**25.5**	35.5	38.3	**30.0**	**39.3**	**29.9**	**23.8**	**30.3**	**27.1**	**29.9**

Ⅳ. 실 험

1. 실험 설정

1.1 데이터세트

Human3.6M 데이터세트 [48]는 4대의 디지털카메라를 이용하여 수집되었으며, 11명의 전문 피사체 (여성 5명, 남성 6명)가 일상복을 입고 일상적인 동작을 수행하는 360만 개의 이미지와 자세 데이터로 이루어져 있다. 피사체의 체형 변화를 나타내는 체질량 지수는 17에서 29까지 변화가 매우 크다. 피사체들은 길 안내, 토론, 식사, 앉아서 활동하기, 인사하기, 사진 찍기, 자세 취하기, 쇼핑하기, 흡연하기, 기다리기, 걷기, 의자에 앉기, 전화 통화하기, 강아지 산책시키기, 함께 걷기 등 15가지의 동작을 수행한다. 자세 데이터는 10대의 모션 캡쳐 카메라로 획득되었으며, 17개의 관절의 keypoint로 구성되어 있다. 본 논문에서는 HPE 모델 학습을 위해 5명의 피사체 (1, 5, 6, 7, 8) 데이터를 사용하였으며, 2명의 피사체 (9, 11) 데이터를 테스트를 위해 사용했다.

1.2 성능 평가 기법

제안하는 데이터 증강 기법을 [49]의 HPE 모델 학습 과정에 추가한 후, [49]와 동일한 하이퍼 파라미터 설정을 사용해 HPE 모델을 처음부터 학습했다. 이때, 효율적인 학습을 위해 데이터세트의 입력 이미지에 대해 가능한 모든 회전 각도에서 합성된 이미지를 오프라인으로 미리 합성하여 학습 중 동일한 임의시점 영상들이 중복해서 생성되지 않도록 했다. 또한, 하이퍼 파라미터 λ와 어려운 예제를 선택하는 기준에 따른 기법 평가했다.

먼저, 어려운 예제를 샘플링하는 비율인 하이퍼 파라미터 λ에 따른 성능 평가를 위해 $\lambda \in \{0.1, 0.3, 0.5\}$에 따른 실험을 진행하였다. 다음으로, 미니배치에서 원본 학습 샘플을 대체할 어려운 예제를 선택하는 기준에 따른 성능 평가를 위해 세 가지 기준을 비교하였다. 비교한 세 가지 기준은 제안하는 방법 (proposed), 10° 간격으로 합성된 샘플 중 임의의 샘플을 무작위로 선택하는 방법 (alt-aug-1), 원본 데이터 샘플보다 D_x가 크더라도 합성된 샘플에서 D_x의 최솟값을 가진 샘플을 선택하는 방법 (alg-aug-2)이다. 추가적으로 전체 회전각에서 어려운 예제를 선택하는 것이 제한된 회전에서 어려운 예제를 선택하는 것보다 더 좋은 성능의 개선을 보이는지 조사하기 위하여 제한적인 회전 (Part. Rot.) 및 전체 회전 (Full Rot.) 방법을 비교하였다.

HPE 모델의 예측 성능을 정량화하기 위해 MPJPE (mean per joint position error) 지표를 이용한다. MPJPE는 예측된 keypoint와 ground truth keypoint 사이의 거리를 아래와 같이 측정한다.

$$MPJPE = \frac{1}{N}\frac{1}{J}\sum_{n=1}^{N}\sum_{j=1}^{J} \| (p_n^{(j)} - p_{\surd}. \quad (3)$$

N은 샘플의 갯수, J는 인스턴스의 keypoint 갯수를 나타낸다. p_n과 y_n은 각각 예측된 keypoint와 그에 해당하는 정답 (ground truth)을, p_{root}와 y_{root}는 각각 예측된 keypoint와 그에 해당하는 정답의 root keypoint를 나타낸다.

2. 정량적 성능 평가

표 1은 Human3.6M 데이터세트 [48]에서 제안하는 기법의 정량적 성능 평가를 수행한 결과를 보여준다. 각 예측의 정확도는 15개 동작들의 MPJPE 평균으로 계산되었다. 표의 첫 번째 행 (Aug Method)은 각 동작의 레이블을 나타내며, 1부터 15까지의 동작 레이블은 각각 길 안내, 토론, 식사, 인

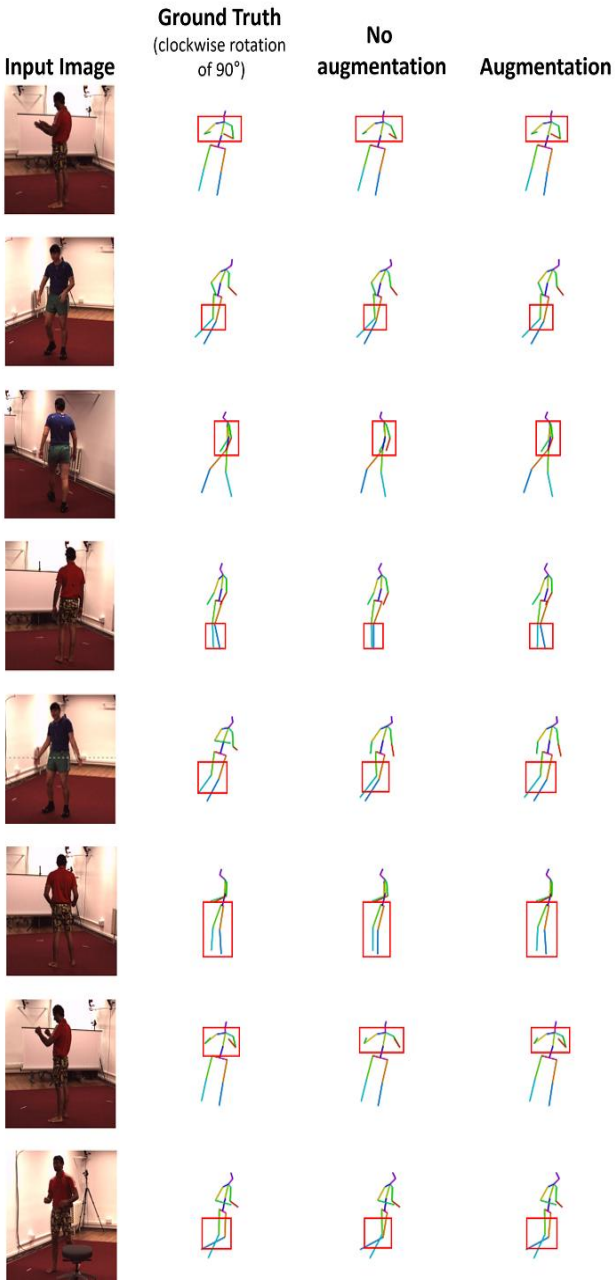


그림 3. 제안하는 기법의 정성적 성능
Fig. 3. Qualitative results of the proposed method

사, 전화 통화하기, 사진 촬영, 자세 취하기, 쇼핑하기, 앉아서 활동하기, 의자에 앉기, 흡연하기, 기다리기, 강아지 산책시키기, 걷기, 함께 걷기에 해당한다. 다음으로 두 번째 행(No aug.)은 원본 Human3.6M 데이터세트 [48]에 대한 HPE 모델의 성능을 나타낸다. 표 1의 첫 번째 섹션 "Part. Rot"은 제한된 회전 각도 [-30°, 30°]에서 어려운 예시를 선택한 결과에 해당하며, 두 번째 섹션 "Full Rot"은 전체 각도를 고려한 방법에 해당한다. 이후, 섹션별로 어려운 예제를 선택하는 세 가지 기준에 대한 비교를 진행하였다. 이때, 각 기준은 다시 이미지 증강 비율 λ에 따라 나누어 비교하였고, λ는 0.1, 0.3, 0.5의 3가지 경우로 정의하였다.

제안된 기법은 HPE 모델의 예측 정확도를 최대 6.2%까지 향상시켰으며, 대부분의 세부 동작들에서 일관된 성능의 향상을 보였다. 구체적인 성능 개선 사항들은 다음과 같다. 첫 번째, "Full Rot", 즉, 전체 회전 각도 범위에서 어려운 예제를 추출하여 증강하는 것이 제한된 각도의 경우 "Part. Rot" 보다 높은 성능 개선을 보였다. 그 이유는 고려하는 회전 각도의 범위가 넓을수록 자세에 더 큰 변화가 생기기 때문에 더 어려운 예제가 생성될 가능성이 크기 때문이다. 두 번째, 어려운 예제를 선택하는 기준이 고정된 경우, 어려운 예제를 샘플링하는 확률인 λ를 높이는 것이 HPE 모델의 예측 정확도 향상에 영향을 주었다. 이 또한 더 어려운 예제를 찾을 확률이 높아지는 측면에서 설명할 수 있다.

반면, 어려운 예제를 부적절하게 선택하면 예측 정확도가 저하되는 경향이 있으며, 이는 무작위 증강인 alt-aug-1과 D_x의 최솟값을 가진 어려운 예제를 무조건 선택하는 alt-aug-2 사이의 성능 격차를 통해 증명된다. 특히 alt-aug-2의 경우 HPE의 성능은 증강 기법을 수행하지 않은 경우보다도 오히려 낮았다.

3. 정성적 성능 평가

그림 3은 Human3.6M의 테스트 세트에서 원본 샘플을 이용하여 학습한 HPE 모델이 실패한 예제에서 우리가 제안한 방법으로 학습된 HPE 모델이 성공적으로 예측한 사례를 제시한다. 이때, 3차원 keypoint의 재구성, 즉, 깊이 정보의 재구성을 시각화하기 위해 테스트 이미지의 3차원 자세에 대한 실측 데이터를 시계 방향으로 90° 회전하여 표시하였다. 또한, 증강을 적용하지 않은 경우와 비교하여 제안한 방법의 예측 정확도가 뚜렷한 keypoint 영역을 경계 상자로 강조하여 표시하였다.

그림 4에서 제시된 실패 사례들은 다리, 상체, 팔에서 깊이 정보의 변화가 크다는 공통점이 있으며, 제안된 방법이 더욱 개선될 수 있음을 보여준다.

V. 결 론

폐색 문제는 3차원 HPE의 예측 정확도 향상에 결정적인 영향을 끼친다. 본 논문에서는 어려운 자세가 일반적으로 인체의 양쪽에 대칭적으로 존재하는 keypoint의 거리가 가까워질 때 발생한다는 사실에 착안하여 임의시점합성을 이용해 어려운 자세의 이미지를 합성하는 간단하면서도 효과적인 방법을 제시하고 이를 통해 폐색 문제를 개선했다. 또한, 입력 이미지의 시점에 따른 예측 난이도를 정량화할 수 있는 거리 지표를 제안했으며, 이를 HPE 모델의 학습 과정에 통합하여 입력 자세를 어렵게 만드는 새로운 시점의 이미지를 검색하여 활용하는 증강 기법을 제안하였다. 본 논문에서 Human3.6M 데이터 세트 기반의 다양한 실험을 통해 제시한 정성적 및 정량적 결과는 새로운 시점에서 합성된 어려운 예제로 보강된 데이터세트를 학습함으로써 HPE 모델을 효율적이고 효과적으로 개선할 수 있음을 보여준다.

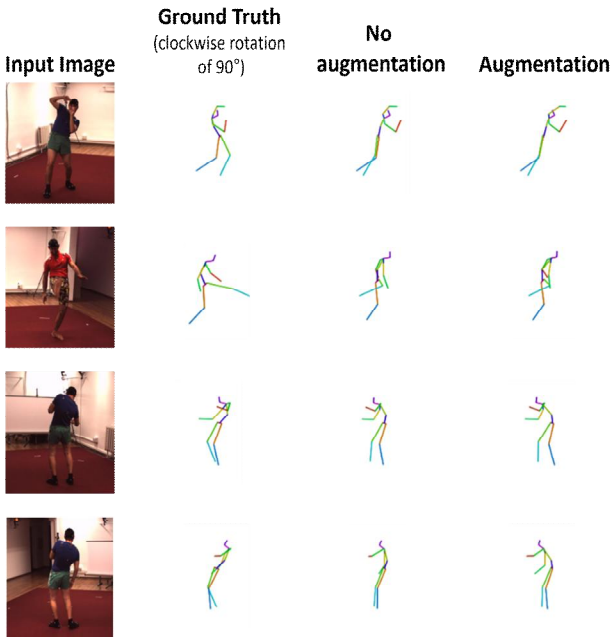


그림 4. 깊이의 급격한 변화로 인한 실패 사례
Fig. 4. Failures due to severe depth ambiguities

References

[1] J. Wang, S. Jin, W. Liu, W. Liu, C. Qian, P. Luo, "When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks," In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11855-11864, 2021.

[2] Y. Kwon, S. Petrangeli, D. Kim, H. Wang, H. Fuchs, V. Swaminathan, "Rotationally-consistent Novel View Synthesis for Humans," In Proceedings of the Proceedings of the 28th ACM International Conference on Multimedia, pp. 2308 - 2316, 2020.

[3] K. Olszewski, S. Tulyakov, O. Woodford, H. Li, L. Luo, "Transformable Bottleneck Networks," In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7648 - 7657, 2019.

[4] L. Ke, M. C. Chang, H. Qi, S. Lyu, "Multi-scale Structure-aware Network for Human Pose Estimation," In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp. 713 - 728, 2018.

[5] Y. Bin, X. Cao, X. Chen, Y. Ge, Y. Tai, C. Wang, J. Li, F. Huang, C. Gao, N. Sang, "Adversarial Semantic Data Augmentation for Human Pose Estimation," In Proceedings of the Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23 - 28, 2020, Proceedings, Part XIX 16. Springer, pp. 606 - 622, 2020.

[6] R. Pytel, O. S. Kayhan, J. C. van Gemert, "Tilting at Windmills: Data Augmentation for Deep Pose Estimation does not Help with Occlusions," In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 10568 - 10575, 2021.

[7] H. Rhodin, M. Salzmann, P. Fua, "Unsupervised Geometry-aware Representation for 3D Human Pose Estimation," In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp. 750 - 767, 2018.

[8] P. Buehler, M. Everingham, D. P. Huttenlocher, A. Zisserman, "Upper Body Detection and Tracking in Extended Signing Sequences," International Journal of Computer Vision, Vol. 95, pp. 180 - 197, 2011.

[9] B. Sapp, D. Weiss, B. Taskar, "Parsing Human Motion with Stretchable Models," In Proceedings of the CVPR 2011. IEEE, pp. 1281 - 1288, 2011.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part-based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, pp. 1627 - 1645, 2009.

[11] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, N. Navab, "Multiple Human Pose Estimation with Temporally Consistent 3D Pictorial Structures," In Proceedings of the Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13. Springer, pp. 742 - 754, 2015.

[12] A. Toshev, C. Szegedy, "Deeppose: Human Pose Estimation Via Deep Neural Networks," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653 - 1660, 2014.

[13] W. Ouyang, X. Chu, X. Wang, "Multi-source Deep Learning for Human Pose Estimation," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2329 - 2336, 2014.

[14] J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation," Advances in Neural Information Processing Systems, Vol. 27, 2014.

[15] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang, "Multi-context Attention for Human Pose Estimation," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1831 - 1840, 2017.

[16] G. Tian, Y. Yi, Z. Meng, Z. Li, J. Song, "PRM: Pose Recalibration Module for Action Recognition," In Proceedings of the The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021). Springer, pp. 757 - 766, 2022.

[17] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, "Learning Feature Pyramids for Human Pose Estimation," In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, pp. 1281 - 1290, 2017.

[18] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, J. Sun, "Learning Delicate Local Representations for Multi-person Pose Estimation," In Proceedings of the Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23 - 28, 2020, Proceedings, Part III 16. Springer, pp. 455 - 472, 2020.

[19] W. Tang, Y. Wu, "Does Learning Specific Features for Related Parts Help Human Pose Estimation?," In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1107 - 1116, 2019.

[20] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, "Cascaded Pyramid Network for Multi-person Pose Estimation," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103 - 7112, 2018.

[21] Z. Cao, T. Simon, S. E. Wei, Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291 - 7299, 2017.

[22] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, J. Wang, "High-resolution Representations for Labeling Pixels and Regions," arXiv preprint arXiv:1904.04514, 2019.

[23] D. Ji, J. Kwon, M. McFarland, S. Savarese, "Deep View Morphing," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2155 - 2163, 2017.

[24] T. D. Kulkarni, W. F. Whitney, P. Kohli, J. Tenenbaum, "Deep Convolutional Inverse Graphics Network," Advances in Neural Information Processing Systems, Vol. 28, 2015.

[25] E. Park, J. Yang, E. Yumer, D. Ceylan, A. C. Berg, "Transformation-grounded Image Generation Network for Novel 3D View Synthesis," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3500 - 3509, 2017.

[26] K. Rematas, C. H. Nguyen, T. Ritschel, M. Fritz, T. Tuytelaars, "Novel Views of Objects from a Single Image," IEEE Transactions on Pattern Analysis and Machine Intelligence , Vol. 39, pp. 1576 - 1590, 2016.

[27] R. Zhang, P. Isola, A. A. Efros, "Colorful Image Colorization," In Proceedings of the Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer, pp. 649 - 666, 2016.

[28] T. Zhou, S. Tulsiani, W. Sun, J. Malik, A. A. Efros, "View Synthesis by Appearance Flow," In Proceedings of the Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 - 14, 2016, Proceedings, Part IV 14. Springer, pp. 286 - 301, 2016.

[29] M. Tatarchenko, A. Dosovitskiy, T. Brox, "Single-view to Multi-view: Reconstructing Unseen Views with a Convolutional Network," CoRR abs/1511.06702, Vol. 1, No. 2, 2015.

[30] M. Tatarchenko, A. Dosovitskiy, T. Brox, "Multi-view 3D Models from Single Images with a Convolutional Network," In Proceedings of the Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 - 14, 2016, Proceedings, Part VII 14. Springer, pp. 322 - 337, 2016.

[31] J. Yang, S. E. Reed, M. H. Yang, H. Lee, "Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis," Advances in Neural Information Processing Systems, Vol. 28, 2015.

[32] J. Flynn, I. Neulander, J. Philbin, N. Snavely, "Deepstereo: Learning to Predict New Views from the World's Imagery," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5515 - 5524, 2016.

[33] G. E. Hinton, A. Krizhevsky, S. D. Wang, "Transforming Auto-encoders," In Proceedings of the Artificial Neural Networks and Machine Learning - ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21. Springer, pp. 44 - 51, 2011.

[34] D. Jimenez Rezende, S. M. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, N. Heess, "Unsupervised Learning of 3D Structure from Images," Advances in Neural Information Processing Systems, Vol. 29, 2016.

[35] X. Yan, J. Yang, E. Yumer, Y. Guo, H. Lee, "Perspective Transformer Nets: Learning Single-view 3D Object Reconstruction Without 3D Supervision," Advances in Neural Information Processing Systems, Vol. 29, 2016.

[36] Z. Zhu, T. Huang, M. Xu, B. Shi, W. Cheng, X. Bai, "Progressive and Aligned Pose Attention Transfer for Person Image Generation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, No. 8, pp. 4306 - 4320, 2021.

[37] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, M. Zollhofer, "Deepvoxels: Learning Persistent 3D Feature Embeddings," In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2437 - 2446, 2019.

[38] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. V. Gool, "Pose Guided Person Image Generation," Advances in Neural Information Processing Systems, Vol. 30, 2017.

[39] C. Lassner, G. Pons-Moll, P. V. Gehler, "A Generative Model of People in Clothing," In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, pp. 853 - 862, 2017.

[40] J. Walker, K. Marino, A. Gupta, M. Hebert, "The Pose Knows: Video Forecasting by Generating Pose Futures," In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, pp. 3332 - 3341, 2017.

[41] A. Siarohin, E. Sangineto, S. Lathuiliere, N. Sebe, "Deformable Gans for Pose-based Human Image Generation," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3408 - 3416, 2018.

[42] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, M. Fritz, "Disentangled Person Image Generation," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 99 - 108, 2018.

[43] P. Esser, E. Sutter, B. Ommer, "A Variational U-net for Conditional Appearance and Shape Generation," In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8857 - 8866, 2018.

[44] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, Y. Wang, "Motionbert: A Unified Perspective on Learning Human Motion Representations," In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15085 - 15099, 2023.

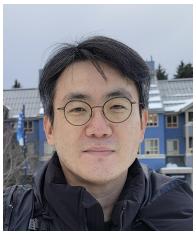
[45] J. Zhang, Z. Tu, J. Yang, Y. Chen, J. Yuan, "MixSTE: Seq2seq Mixed Spatio-temporal Encoder for 3D Human Pose Estimation in Video," In Proceedings of the CVF Conference

on Computer Vision and Pattern Recognition (CVPR), pp. 13222 - 13232, 2022.
[46] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding, "3D Human Pose Estimation with Spatial and Temporal Transformers," In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11656 - 11665, 2021.
[47] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, W. Gao, "P-stmo: Pre-trained Spatial Temporal Many-to-one Model for 3D Human Pose Estimation," In Proceedings of the European Conference on Computer Vision. Springer, pp. 461 - 478, 2022.
[48] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, "Human3. 6m: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, pp. 1325 - 1339, 2013.
[49] H. Ci, C. Wang, X. Ma, Y. Wang, "Optimizing Network Structure for 3D Human Pose Estimation," In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2262 - 2271, 2019.

Minji Kim (김 민 지)

2015 Computer Science and Engineering from Jeonbuk National University (B.S.)
2017 Computer Science and Engineering from Jeonbuk National University (M.S.)
2021~Computer Science and Artificial Intelligence from Jeonbuk National University (Ph.D student.)
Field of Interests: Artificial Intelligence, Computer Vision
Email: kmkmj927@jbnu.ac.kr

Sungchan Kim (김 성 찬)



1998 Metallurgical Engineering from Seoul National University (B.S.)
2000 Computer Science and Engineering from Seoul National University (M.S.)
2005 Computer Science and Engineering from Seoul National University (Ph.D)
2009~Department of Computer Science and Artificial Intelligence, Jeonbuk National University (Professor)
Field of Interests: Artificial Intelligence, Computer Vision
Email: s.kim@jbnu.ac.kr