3 차원 휴먼 자세 추정을 위한 다시점 준지도 학습

김도엽, 장주용
광운대학교
dyubkim@kw.ac.kr, jychang@kw.ac.kr

Multi-view semi-supervised learning for 3D human pose estimation

Do Yeop Kim, Ju Yong Chang
Kwangwoon University

요 약

3 차원 휴먼 자세 추정 모델은 다시점 모델과 단시점 모델로 분류될 수 있다. 일반적으로 다시점 모델은 단시점 모델에 비하여 뛰어난 자세 추정 성능을 보인다. 단시점 모델의 경우 3 차원 자세 추정 성능의 향상은 많은 양의 학습 데이터를 필요로 한다. 하지만 3 차원 자세에 대한 참값을 획득하는 것은 쉬운 일이 아니다. 이러한 문제를 다루기 위해, 우리는 다시점 모델로부터 다시점 휴먼 자세 데이터에 대한 의사 참값을 생성하고, 이를 단시점 모델의 학습에 활용하는 방법을 제안한다. 또한, 우리는 각각의 다시점 영상으로부터 추정된 자세의 일관성을 고려하는 다시점 일관성 손실함수를 제안하여, 이것이 단시점 모델의 효과적인 학습에 도움을 준다는 것을 보인다.

1. 서론

3 차원 휴먼 자세 추정(3D human pose estimation) 방법은 크게 다시점 모델(multi-view model)과 단시점 모델(single-view model)로 구분될 수 있다.

한 자세에 대한 여러 카메라 시점의 영상을 입력으로 사용하는 다시점 모델은 단시점 모델보다 정확한 자세 추정이 가능하다. 그 이유는 다시점 모델이 3 차원 휴먼 자세 추정 시 깊이 모호성(depth ambiguity) 문제와 영상 시점에 따른 가리워짐(occlusion) 문제에 강인한 모델을 다시점 영상으로부터 학습할 수 있기 때문이다.

단시점 모델은 단일 시점의 영상 입력으로부터 3 차원 휴먼 자세를 추정하는 방법으로 최근 딥러닝의 발전과 함께 큰 성능 증가를 보였다. 그러나 여전히 다시점 모델에 비하여 깊이 모호성 문제와 가리워짐 문제에 취약하다. 단시점 모델의 성능 개선은 다양한 시점과 자세를 포함하는 대량의 정제된 데이터를 필요로 한다. 그러나 3 차원 자세에 대한 참값(GT; ground-truth)을 제공하는 데이터를 획득하는 일은 일반적으로 많은 시간과 비용을 필요로 한다.

본 논문에서 우리는 3 차원 자세 GT 가 제공되지 않는(unlabeled), 캘리브레이션된 다시점 데이터셋을 가정하고, 이러한 데이터셋을 활용하여 단시점 모델의 성능을 개선하는 방법을 제안한다. 기본적인 아이디어는 사전 학습된 다시점 모델[1]을 unlabeled 다시점 데이터셋에 적용하고, 그 추정 결과를 다시점 영상들에 대응하는 3 차원 휴먼 자세에 대한 의사 참값(P-GT; pseudo-GT)으로서 단시점 모델의 학습에 활용하는 것이다. 또한 우리는 다시점 영상에 대한 단시점 모델의 자세 추정 결과들에 일관성을 부여하는 다시점 일관성 손실함수(multi-view consistency loss)를 제안한다. 이는 단시점 모델의 3 차원 깊이 추정 성능과 가리워짐 발생 시 휴먼 자세 추정 성능을 개선한다.

우리는 제안하는 3 차원 휴먼 자세 추정 방법을 정량적,

정성적으로 평가한다. 그리고 평가 결과로부터 다시점 모델로부터 획득된 P-GT 가 단시점 모델의 학습 및 성능 개선에 활용될 수 있음을 보인다.

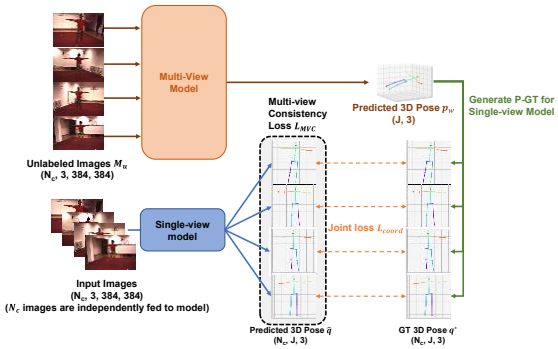

그림 1. 제안하는 방법의 개요

2. 제안하는 방법

본 연구에서 제안하는 단시점 모델의 성능 개선을 위해 unlabeled 다시점 영상 데이터셋을 활용하는 방법의 대략적인 절차는 다음과 같다. 첫 번째, 사전 학습된 다시점 모델로부터 P-GT 를 생성한다. 두 번째, GT 를 포함하는 labeled 데이터셋으로 사전 학습된 모델로 단시점 모델을 초기화 한다. 세 번째, P-GT 를 포함하는 unlabeled 다시점 영상 데이터셋 및 GT 를 포함하는 labeled 데이터셋을 함께 사용하여 단시점 모델을 추가 학습한다. 이 때 제안하는 다시점 일관성 손실함수와 함께 단시점 모델을 최적화한다.

2.1. 단시점 모델의 사전 학습

본 연구에서 3 차원 휴먼 자세 추정을 수행하는 단시점 모델은 ImageNet[2]으로 사전 학습된 ResNet-50[3]을 백본(backbone)으로 하는 적분 회귀(integral regression) 모델[4]이다. 단시점 모델을 구성하기 위해 ResNet-50 모델에서 global average pooling 층과 fully connected 층을 3 개의 연속된 deconvolution 층과 하나의 1x1 convolution 층으로 바꿔 fully convolutional 네트워크 F 를 만든다. 각 deconvolution 층의 필터 크기와 stride 는 각각 4 와 2 로 설정한다. 단시점 모델은 F 의 출력 텐서에 soft-argmax[4] 연산을 적용하여 3 차원 휴먼 자세를 구성하는 관절 좌표를 획득한다.

구체적으로 네트워크 F 는 입력 영상 $I \in \mathbb{R}^{3\times384\times384}$ 을 입력 받아 J개 관절에 대한 3 차원 히트맵(heatmap) $\boldsymbol{H}_j \in \mathbb{R}^{96\times96\times96}$ $(j = 1, \dots, J)$ 를 출력한다. 그 후, $\boldsymbol{H}_j$ 에 soft-argmax 를 적용하여 각 관절의 3 차원 좌표 $\boldsymbol{q}_j = [q_{j,x}, q_{j,y}, q_{j,z}]^T \in \mathbb{R}^3$를 획득한다.

Soft-argmax 연산은 히트맵을 확률 분포로 만들고 기댓값(expectation)을 계산하여 공간 좌표를 획득하는 방법으로 다음 식은 j 번째 관절에 대한 3 차원 히트맵 $\boldsymbol{H}_j(x,y,z)$를 확률 분포 $\widetilde{\boldsymbol{H}}_j(x,y,z)$로 만들기 위해 softmax 연산을 적용하는 과정을 나타낸다:

$$\widetilde{\boldsymbol{H}}_j(x,y,z) = \frac{e^{\boldsymbol{H}_j(x,y,z)}}{\sum_{z'}\sum_{y'}\sum_{x'} e^{\boldsymbol{H}_j(x',y',z')}}. \quad (1)$$

그 후 $\widetilde{\boldsymbol{H}}_j(x,y,z)$에 다음과 같은 기댓값 연산을 적용하여 특징점 좌표 $\boldsymbol{q}_j$를 획득한다:

$$\begin{cases} q_{j,x} = \sum_{x'}\sum_{y'}\sum_{z'} x' \widetilde{\boldsymbol{H}}_j(x',y',z') \\ q_{j,y} = \sum_{x'}\sum_{y'}\sum_{z'} y' \widetilde{\boldsymbol{H}}_j(x',y',z') \\ q_{j,z} = \sum_{x'}\sum_{y'}\sum_{z'} z' \widetilde{\boldsymbol{H}}_j(x',y',z') \end{cases}. \quad (2)$$

우리는 식 (2)를 모든 관절에 적용하여 3 차원 휴먼 자세 $\boldsymbol{q} = \{\boldsymbol{q}_1, \dots, \boldsymbol{q}_J\}$를 획득한다.

단시점 모델의 사전 학습을 위해 우리는 모델이 출력한 $\boldsymbol{q}$에 labeled 데이터셋의 GT 에 기반한 L1 손실 함수를 적용한다.

2.2. P-GT 데이터셋 생성

우리는 unlabeled 다시점 영상 데이터셋을 활용하기 위하여 사전 학습된 다시점 모델 G 로부터 J개 관절들의 좌표 $\boldsymbol{p} = \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_J\}$를 추정한다. 다시점 모델 G 는 [1]의 algebraic triangulation 모델을 사전 학습하여 사용한다.

$$\boldsymbol{p} = G(\boldsymbol{M}_u), \quad (3)$$

여기서 $\boldsymbol{M}_u = \{\boldsymbol{m}_c\}_{c=1}^{N_c} \in \mathbb{R}^{N_c\times3\times384\times384}$는 unlabeled 다시점 영상 데이터셋의 한 sample 이다. $\boldsymbol{m}_c$와 N_c는 각각 시점 c 에서 촬영된 영상과 한 자세를 관찰하는 서로 다른 시점의 개수를 의미한다. 여기서 3 차원 휴먼 자세 $\boldsymbol{p}$ 는 월드 좌표계(world coordinate system)에서 정의된다.

$\boldsymbol{p}$를 단시점 모델의 학습에 활용하기 위해 우리는 $\boldsymbol{p}$를 각 시점 c 의 영상 $\boldsymbol{m}_c$ 에 대응하는 히트맵 공간으로 정규화(normalization)한다. 이는 $\boldsymbol{p}$ 의 $\boldsymbol{m}_c$ 에 대한 원근 투영(perspective projection), 픽셀 좌표에 대한 정규화, 그리고 깊이 좌표에 대한 정규화 과정들로 이루어진다.

다음 식은 관절 좌표 $\boldsymbol{p}_j$ 를 $\boldsymbol{m}_c$ 에 투영하여 2 차원 좌표를 획득하고, 픽셀 좌표에 대해 정규화 하는 과정을 나타낸다:

$$\boldsymbol{p}_{j,c} = [p_{j,c,x}, p_{j,c,y}, p_{j,c,z}]^T = \boldsymbol{R}_c\boldsymbol{p}_j + \boldsymbol{t}_c, \quad (4)$$

$$[p_{j,c,u}, p_{j,c,v}, 1]^T = \boldsymbol{K}_c\boldsymbol{p}_{j,c}, \quad (5)$$

$$[q_{j,c,x}, q_{j,c,y}]^T = \left[\frac{p_{j,c,u}}{4}, \frac{p_{j,c,v}}{4}\right]^T. \quad (6)$$

$\boldsymbol{R}_c \in SO(3)$ 와 $\boldsymbol{t}_c \in \mathbb{R}^3$ 는 카메라 c 의 외부 파라미터를, 그리고 $\boldsymbol{K}_c \in \mathbb{R}^{3\times3}$ 는 내부 파라미터를 나타낸다. 식 (4)를 통해 월드 좌표계에서 정의된 $\boldsymbol{p}_j$ 는 카메라 좌표계에서 정의된 관절 좌표 $\boldsymbol{p}_{j,c}$ 로 변환된다. $\boldsymbol{p}_{j,c}$ 는 식 (5)를 통해

픽셀 좌표 $[p_{j,c,u}, p_{j,c,v}]^T$ 로 투영된다. 마지막으로 식 (6)을 통해 우리는 영상 크기와 히트맵 크기 사이의 비율인 4 를 사용하여 정규화된 2 차원 좌표 $[q_{j,c,x}, q_{j,c,y}]^T \in \mathbb{R}^2$를 얻을 수 있다. 또한 우리는 $\boldsymbol{p}_{j,c}$ 의 깊이 좌표 $p_{j,c,z}$ 를 다음과 같이 정규화한다:

$$q_{j,c,z} = \left(\frac{(p_{j,c,z} - p_{root,c,z})}{1000} + 1\right) * 0.5 * 96. \quad (7)$$

각 관절의 깊이 좌표를 정규화 하기 위하여 우리는 먼저 휴먼 객체의 크기가 $2000\,mm^3$ 이하임을 가정한다. $p_{j,c,z}$ 에서 골반 관절의 깊이 값인 $p_{root,c,z}$ 를 빼서 골반 관절을 기준으로 상대적으로 정의되는 깊이 값을 얻는다. 이러한 깊이 값은 $[-1000, 1000]$ 의 범위에 존재하므로 우리는 추가적으로, 정규화된 깊이 값이 히트맵의 뎁스 축 범위인 $[0, 96]$ 내에 존재하도록 만든다. 이러한 과정은 식 (7)에 나타나 있으며, 이를 통해 우리는 정규화된 깊이 값 $q_{j,c,z}$ 를 획득한다. 결국 시점 c 에 대응하는 히트맵 공간에서 정의되는 P-GT 관절 좌표 $\boldsymbol{q}^*_{j,c}$는 다음과 같다:

$$\boldsymbol{q}^*_{j,c} = [q_{j,c,x}, q_{j,c,y}, q_{j,c,z}]. \quad (8)$$

식 (4)-(8)을 각 시점 c 에 적용하여 우리는 unlabeled sample $\boldsymbol{M_u}$에 대한 P-GT $\{\boldsymbol{q}^*_c\}_{c=1}^{N_c}$를 획득할 수 있고, 이를 활용하여 우리는 P-GT 데이터셋 $\boldsymbol{M} = \{\boldsymbol{m}_c, \boldsymbol{q}^*_c\}_{c=1}^{N_c}$ 을 구성한다.

2.3. 다시점 일관성 손실함수

제안하는 방법은 다시점 영상 데이터를 사용하여 단시점 모델을 학습한다. 단시점 모델이 한 자세에 대응하는 다시점 영상을 입력 받는 경우, 모델에 의해 추정된 각 시점에서의 휴먼 자세들은 일관된 자세를 취해야 한다. 따라서 우리는 학습된 모델로 하여금 이러한 조건을 만족시키게끔 하기 위해 다시점 일관성 손실함수를 제안한다. 다시점 일관성 손실함수는 각 시점에 대해 추정된 관절 좌표들을 월드 좌표계 기준으로 변환하고, 그 결과 자세들 사이의 L1 손실 함수들의 합으로 정의된다. 히트맵 공간으로 정규화된 관절 좌표를 월드 좌표계로 변환하는 과정은 식 (4)-(8)의 역 연산으로 수행된다. 이제 다시점 일관성 손실함수 L_{MVC}는 다음과 같다:

$$L_{MVC} = \frac{1}{J}\sum_{j=1}^{J}\sum_{c\neq c'}\left\|\Pi_c(\hat{\boldsymbol{q}}_{j,c}) - \Pi_{c'}(\hat{\boldsymbol{q}}_{j,c'})\right\|_1, \quad (9)$$

여기서 $\hat{\boldsymbol{q}}_{j,c}$와 $\hat{\boldsymbol{q}}_{j,c'}$는 각각 시점 c와 c'에서 단시점 모델에 의해 추정된 정규화된 관절 좌표를 나타내며, Π 는 정규화된 관절 좌표를 월드 좌표계 기준으로 변환하는 함수이다.

우리는 사전 학습된 단시점 모델을 미세 조정(fine-tuning) 하기 위한 손실 함수 L을 다음과 같이 정의한다:

$$L = \alpha L_{coord} + \beta L_{MVC}, \quad (10)$$

여기서 L_{coord}는 단시점 모델의 출력 $\hat{\boldsymbol{q}}$에 적용하는 P-GT 및 GT 에 기반한 L1 손실 함수들의 합으로 정의된다. α와 β 는 각 손실 함수의 영향력을 결정하는 가중치이다. 단시점 모델의 미세 조정을 위해 우리는 식 (10)을 최소화한다.

3. 실험 결과

3.1. 데이터셋, 평가 방법, 구현 세부사항

본 연구는 제안하는 방법을 학습, 평가하기 위하여 대규모의 3 차원 휴먼 자세를 포함하는 Human3.6M[5] 데이터셋을 사용한다. Human3.6M 데이터셋에서 각 휴먼 객체(subject, 이하 S)은 15 가지의 동작을 수행하며 각 휴먼 객체가 동작을 수행하는 비디오를 4 개의 서로 다른 시점의 카메라로 촬영한다. 우리는 기존 연구들[6, 7]의 학습 및 평가 방법에 따라 11 명 중 5 명(S1, S5, S6, S7, S8)의 인물에 대한 데이터를 학습 데이터셋으로 사용한다. 이 중 3 명(S1, S5, S6)의 데이터는 labeled 데이터셋으로, 2 명(S7, S8)의 데이터는 unlabeled 데이터셋으로 가정한다. 나머지 2 명(S9, S11)의 데이터는 평가 데이터셋으로 사용한다. 우리는 평가 데이터셋을 64 프레임 마다 서브 샘플링(sub-sampling)하여 사용한다.

우리는 단시점 모델의 성능을 정량적으로 평가하기 위해 MPJPE 와 PA-MPJPE 를 측정하여 보고한다. MPJPE 는 평가 데이터셋에서 단시점 모델에 의해 추정된 관절과 그 GT 사이의 유클리드 거리를 나타낸다. PA-MPJPE 는 추정된 관절과 GT 사이에 Procrustes alignment[8]를 수행한 후 MPJPE 를 구한 값이다. MPJPE 와 PA-MPJPE 의 단위는 mm이다.

다시점 모델과 단시점 모델의 사전 학습은 labeled 데이터셋으로 수행된다. P-GT 데이터셋의 생성시 학습의 효율성을 위하여 사전 학습된 다시점 모델을 적용한 자세 추정 결과를 오프라인으로 저장한다. 그 후 단시점 모델의 미세 조정 시 사용한다.

다시점 모델의 사전 학습에는 labeled 데이터셋이 사용되며, 에포크(epoch) 수, 배치 크기(batch size), 학습률(learning rate)은 각각 6, 8, 10^{-5} 로 설정한다. 단시점 모델의 사전 학습 또한 같은 학습 데이터셋이 사용되며, 에포크 수, 배치 크기, 학습률은 각각 20, 32, 10^{-4} 이다. 두 모델의 사전 학습에 사용된 optimizer 는 Adam[9] 이다.

단시점 모델의 미세 조정을 위해 우리는 labeled 데이터셋과 P-GT 데이터셋으로 9 에포크 동안 학습한다.

표 1. Human3.6M 평가 데이터셋에 대한 제안하는 방법과 baseline 방법들의 성능 비교

	Base		L1-only		Ours	
	MPJPE (mm)	PA-MPJPE (mm)	MPJPE (mm)	PA-MPJPE (mm)	MPJPE (mm)	PA-MPJPE (mm)
Cam1	107.78	84.72	78.34	62.9	**76.92**	**59.36**
Cam2	99.7	79.11	106.12	86.41	**98.19**	**77.66**
Cam3	116.7	93.13	79.27	62.22	**76.26**	**57.86**
Cam4	96.57	76.76	93.76	74.79	**84.71**	**67.15**

이 때 배치 크기와 학습률은 각각 6, 10^{-4} 로 설정한다. Optimizer 로는 Adam 을 사용한다. 손실 함수의 가중치는 $\alpha = 1$과 $\beta = 0.1$로 설정한다. 제안하는 모델은 Pytorch[10] 프레임워크를 사용하여 구현되었다.

3.2 정량적 결과

우리는 제안하는 방법이 단시점 모델의 성능 개선에 도움을 주는 것을 보이기 위하여 2 가지 baseline 모델들과 제안하는 방법(Ours)을 정량적으로 비교한다. Baseline 모델로는 다음의 2 가지 방법을 사용한다. 첫 번째는 GT 데이터셋으로 학습된 단시점 모델(Base)이고, 두 번째는 다시점 일관성 손실 함수를 적용하지 않고 P-GT 와 각 시점의 추정 결과에 L1 손실 함수만을 적용한 모델(L1-only)이다.

표 1 은 baseline 방법들과 제안하는 방법의 정량적 성능 비교를 보여준다. 우리는 다시점 데이터에 대한 성능 개선을 보이기 위하여 Human3.6M 평가 데이터셋의 각 카메라 시점에 대하여 평가 결과를 제시하였다. 우리는 먼저 L1-only baseline 이 Base 보다 높은 성능을 보임을 알 수 있다. 이 결과는 기존의 다시점 모델이 생성하는 P-GT 가 학습에 도움이 될 만한 정확도를 가짐을 보여준다. 또한 제안하는 방법은 모든 카메라 시점에 대해 두 baseline 모델보다 높은 성능을 보인다. 우리는 이 결과로부터 P-GT 데이터셋과 다시점 일관성 손실함수가 단시점 모델의 3 차원 휴먼 자세 추정 성능을 실질적으로 개선할 수 있음을 확인하였다.

3.3 정성적 결과

그림 2 는 Base 모델과 제안하는 방법의 자세 추정 결과와 그 GT 를 시각적으로 보여준다. 그림 2 의 입력 영상들은 상대적으로 어려운 자세를 포함하고 있는데, 각 영상 위에 투영된 자세 추정 결과들을 통해 우리는 Base 모델에 비하여 제안하는 방법이 GT 에 보다 가까운 자세를 추정함을 알 수 있다.

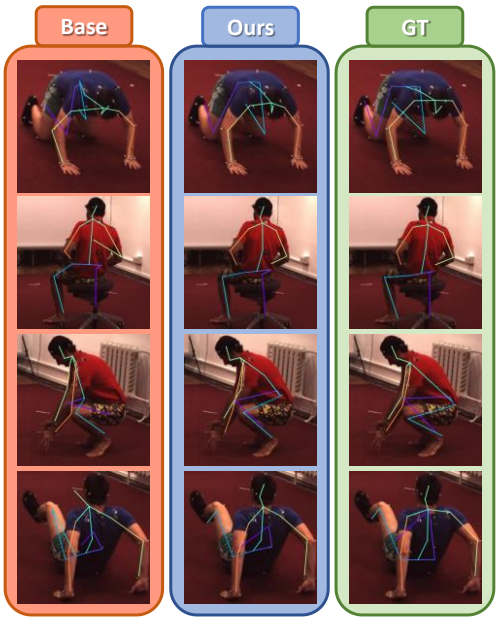

그림 2. Baseline, 제안하는 방법, GT 의 정성적 비교

4. 결론

본 연구는 휴먼 객체의 3 차원 자세 추정을 위한 단시점 모델의 성능을 개선하기 위해 캘리브레이션 된 unlabeled 다시점 데이터셋을 활용하는 준지도 학습 방법을 제안한다. 제안하는 방법은 다시점 데이터에 다시점 모델을 적용하여 P-GT 를 생성하고, 이를 단시점 모델의 미세 조정에 활용한다. 또한 우리는 다시점 입력 영상에 대한 3 차원 휴먼 자세 추정의 일관성을 고려하는 다시점 일관성 손실 함수를 제안한다. 실험을 통해 우리는 기존의 사전 학습된 다시점 모델에 의해 생성된 P-GT 와 다시점 일관성 손실 함수가 단시점 모델의 성능을 정량적, 정성적으로 향상시킴을 확인하였다.

감사의 글

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00348, 2D/3D 영상 통합 분석을 이용한 클라우드 기반 무인점포 환경 대응형 영상보안시스템 개발)

참고문헌

[1] K. Iskakov, E. Burkov, V. Lempisky, and Y. Malkov, "Learnable triangulation of human pose," ICCV, 2019.
[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," NIPS, 2012.
[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition," CVPR, 2016.
[4] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, “Integral human pose regression,” ECCV, 2018.
[5] C. Ionescu, D. Papava, V. Olaru, and 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, 2013.
[6] S. Park and N. Kwak, "3D human pose estimation with relational networks," BMVC, 2018.
[7] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," CVPR, 2017.
[8] J. C. Gower, "Generalized procrustes analysis," Psychometrika, vol. 40, no. 2, 1975.
[9] D. P. Kingma and L. J. Ba, “Adam: A method for stochastic optimization,” ICLR, 2015.
[10] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” NIPS Workshops, 2017.