KPE: Keypoint Pose Encoding for Transformer-based Image Generation

Soon Yau Cheong Armin Mustafa Andrew Gilbert

UNIVERSITY OF SURREY

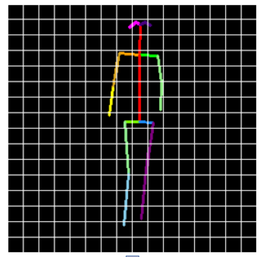
s.cheong@surrey.ac.uk

github.com/soon-yau/kpe

BMVC 2022

Introduction – Redundancy in Pose Representation

Existing method of pose representation transformer-based image generation contain is inefficient. VQGAN[1] breaks a skeleton image into grid, flatten them and encode into discrete tokens. We can see:
- most tokens are black background and do not contain useful pose information. This unnecessary increase computational quadratically for transformer.
- the 2D spatial information is lost after flattening. Unlike convolutional networks, it turns out transformers do not need 2D spatial locality of pixels to learn human pose.



Our main contributions:

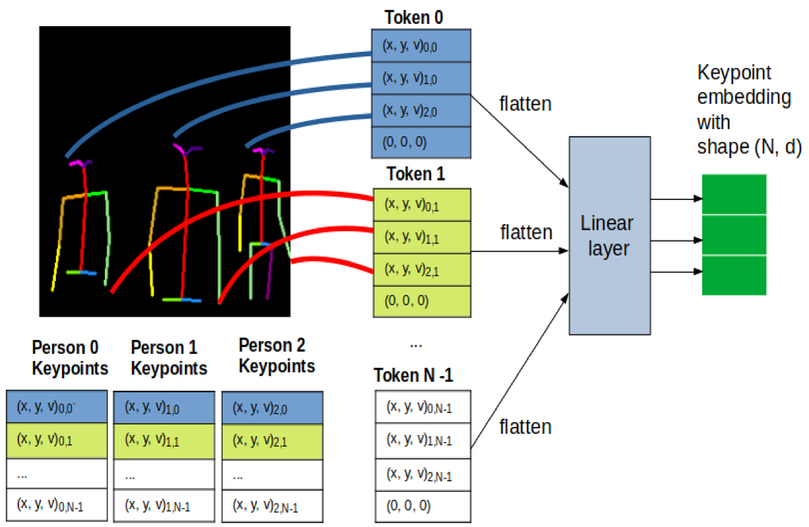
1. ***Keypoint Pose Encoding (KPE)*** to use the low-dimensional keypoint directly. This results in over 10x more memory efficiency and it is over 73% faster to train and inference!

2. ***People Count Error (PCE)*** a novel method to detect people image error.

3. State-of-the-art **text and pose** conditioned image **generative transformer.**.

Methodology – Keypoint Pose Encoding

A single 2D keypoint is defined as a tuple of (x, y, v) where x and y are Cartesian coordinates and v is the visibility score. Instead of expanding keypoints into skeleton image, we encode the keypoints directly into tokens. Figure below show how we encode keypoints of 3 people into N=25 tokens, resulting massive reduction from 16x16=256 tokens of skeleton image.



Text-Pose-to-Image Model

We implemented KPE into DALL-E[2] architecture to generate image with text and pose conditions.

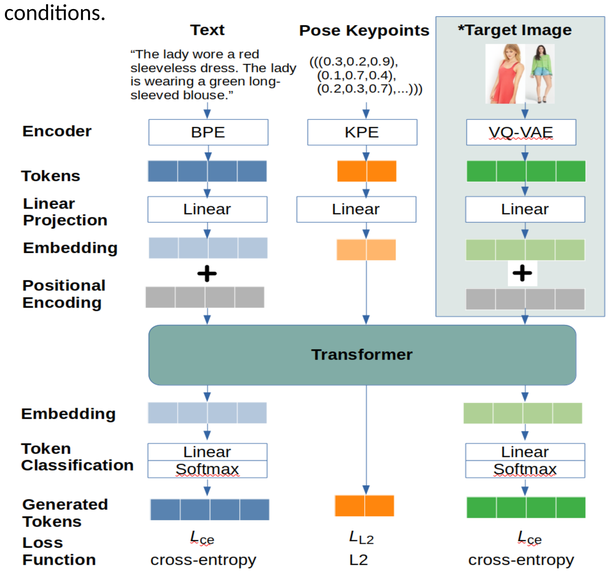


*Image encoding is only used in training, we only use text and pose in inference.

People Count Error (PCE)

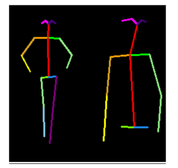
Errors in generated people image often manifest as having missing or additional body parts. ***PCE*** makes use of rich human anatomy knowledge embodied in pose estimation models e.g. OpenPose [3] to detect such error. Let's say we tell the model to generate a person but it ends having 3 arms. OpenPose knows a person can only have 2 arms and allocate the third arm to additional person. As a result, OpenPose think there are 2 people in the image instead of the supposedly one. This discrepancy in people count flags images errors.

(a) gt=2, h=3 (b) gt=2, h=3 (c) gt=2, h=3 (d) gt=1, h=2 (e) gt=3, h=2

where *gt* is the expected people count, and *h* is the people count detected by OpenPose.

High Fidelity Image with Accurate Pose

We created a multiperson text, pose and image dataset from Deepfashion [4] and train our model to generate samples below – consistent and accurate partial and multiscale poses, with people appearances matching text prompt.



+ "the lady is wearing a yellow long-sleeved dress. the lady is wearing a multi-color long-sleeved tee."

Qualitative Results

Our method KPE outperforms in all metrics. We also show that using pose as additional condition improves the image quality and reduce image errors.

Pose Method	DALL-E	DALL-E+VQGAN	KPE (Ours)
Number of pose tokens ↓	-	256	**25**
Relative inference speed ↑	**1.73**×	1.0×	**1.73**×
FID ↓	22.11	21.81	**20.39**
PCE ($\times 10^{-3}$) ↓	8.2	1.2	**0.6**
CLIPSIM ↑	**0.27**	**0.27**	**0.27**
IS ↑	2.912	3.027	**3.034**
OKS ↑	0.598	**0.970**	**0.970**
Mask-SSIM ↑	0.265	0.420	**0.424**

References

[1] Patrick Esser, Robin Rombach, and Björn Ommer. *Taming transformers for high-resolution image synthesis*. CVPR, 2021.
[2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever., *Zero-shot text-to-image generation*. ICML, 2021.
[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. *Openpose: Realtime multi-person 2d pose estimation using part affinity fields*. PAMI, 2019.
[4] Ziwei, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Liu Tang. *Deepfashion: Powering robust clothes recognition and retrieval with rich annotations*, CVPR, 2016.