Denoising Diffusion Probabilistic Models for Styled Walking Synthesis

Edmund J. C. Findlay
edmund.findlay@durham.ac.uk
Durham University
Durham, UK

Haozheng Zhang
haozheng.zhang@durham.ac.uk
Durham University
Durham, UK

Ziyi Chang
ziyi.chang@durham.ac.uk
Durham University
Durham, UK

Hubert P. H. Shum*
hubert.shum@durham.ac.uk
Durham University
Durham, UK

ABSTRACT

Generating realistic motions for digital humans is time-consuming for many graphics applications. Data-driven motion synthesis approaches have seen solid progress in recent years through deep generative models. These results offer high-quality motions but typically suffer in motion style diversity. For the first time, we propose a framework using the denoising diffusion probabilistic model (DDPM) to synthesize styled human motions, integrating two tasks into one pipeline with increased style diversity compared with traditional motion synthesis methods. Experimental results show that our system can generate high-quality and diverse walking motions.

CCS CONCEPTS

• **Computing methodologies → Animation**; • **Motion processing**; • **Machine learning**;

KEYWORDS

motion synthesis, motion style, diffusion models

ACM Reference Format:
Edmund J. C. Findlay, Haozheng Zhang, Ziyi Chang, and Hubert P. H. Shum. 2018. Denoising Diffusion Probabilistic Models for Styled Walking Synthesis. In *Proceedings of The 15th Annual ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG'23).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Human motion synthesis and motion style transfer are two important problems in many computer graphics and animation applications. Traditional motion-based motion synthesis aims to learn a sequence of possible poses when provided with an initial pose, partial sequence of poses or a control signal. Existing deep learning-based methods have shown promising performance in motion synthesis. For example, Holden et al. [Holden et al. 2017] propose a Phase-Functioned Neural Network for generating smooth cyclic human behaviour using a cyclic function which takes the phase as an input. [Holden et al. 2016] proposes a fast motion synthesis model using a deep learning network, which maps high-level parameters to a motion by learning the motion data embedding in a non-linear manifold. These approaches enable animators only use high-level instructions rather than low-level details to produce animation.

Motion style transfer is a technique that transfers the visual style of human motion to one taken from another [Li et al. 2017]. There is an enduring interest in generating motions in a variety of styles for animation and computer games since it is expensive to capture all desired styled motions. Conventional methods use dynamic models [Xia et al. 2015] or analyse the frequency elements of the motion to transfer the motion style [Yumer and Mitra 2016]. Recent studies achieve better performance using the deep neural network [Aberman et al. 2020; Yumer and Mitra 2016]. These methods usually require the style and/or content motions as the input for training. However, there is still a lack of a system that can synthesise human motion while transferring motion style.

Most estixting human motion synthesis or motion style transfer approaches usually focus on a single task and cannot handle two tasks simultaneously. In addition, previous motion synthesis results offer high-quality motions but typically suffer in diversity. For the first time, we propose a framework that applies denoising diffusion probabilistic modelling to support styled human motion synthesis and transfer, integrating two tasks into one pipeline with increased diversity. Training by adding Gaussian noise to the input via a sequence of variances, DDPM can leverage connections to energy-based models and noise conditional score-matching, such that it predicts the value of the noise added as output. We synthesize new motions by iteratively removing the predicted noise starting from random Gaussian noise, and our design is able to generate diverse and convincing synthetic walking motions with different styles (e.g. Figure 1).

Our main contribution is that, to the best of our knowledge, we propose the first end-to-end system applying DDPM for human motion, specifically the walking motion synthesis task.

*Corresponding author.

MIG'23, Nov. 03–05, 2022, Guanajuato, Mexico
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

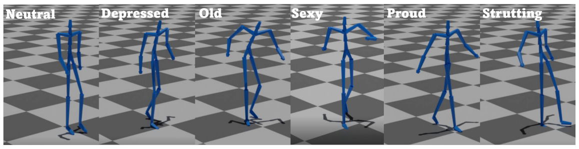


Figure 1: Our synthesized walking motions with different styles.

2 METHODOLOGY

2.1 Denoising Diffusion Probabilistic Model

Ho et al. [Ho et al. 2020] proposed denoising diffusion probabilistic models (DDPM) for high-quality image generation while keeping the diversity. DDPMs add and remove the Gaussian noise $\sqrt{\beta}\epsilon \sim \mathcal{N}(0, \beta\mathbf{I})$ via two processes where β is a predefined noise schedule. One is the forward process to add noise by $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$ and the other is the reverse process to remove the noise with the estimation of the mean value by $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \beta_t\mathbf{I})$ where θ are the parameters of the network. Instead of directly optimizing the prediction of the mean value, Ho et al. [Ho et al. 2020] proposed to simplify the loss as $\mathcal{L}_{ddpm} = \mathbb{E}_{t,x_0,\epsilon}\,||\epsilon - \epsilon_\theta(x_t, t)||_2^2$ where ϵ_θ is the prediction of the added noise.

2.2 DDPM for Walking Motion Synthesis

Given motion data x_0, we follow the two processes proposed by Ho et al [Ho et al. 2020]. Additionally, we condition our the reverse process of DDPM with the content embedding c and style embedding s by $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c, s), \beta_t\mathbf{I})$. Accordingly, the DDPM loss is formulated as $\mathcal{L}_{ddpm} = \mathbb{E}_{t,x_0,\epsilon}\,||\epsilon - \epsilon_\theta(x_t, t, c, s)||_2^2$. The predictions of foot contact and root position are also constrained by squared error losses. We also propose to use a discriminator D to enforce learning of the style and content information via the adversarial loss:

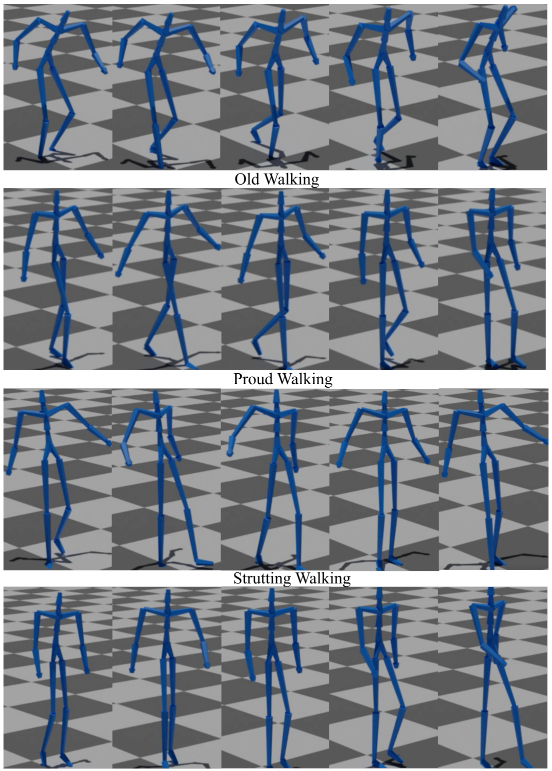
$$\mathcal{L}_{adv} = \frac{1}{2}||D(x_0) - 1||_2^2 + \frac{1}{2}||D(x_0') - 0||_2^2.$$

where x_0' is the predicted motion and x_0 is the original motion. We use such losses as our overall loss function to learn our DDPM for walking motion synthesis.

3 EXPERIMENTS

We train our models on an NVIDIA RTX 3090 GPU and used 32-bit floating-point arithmetic. Motions are from two publicly available datasets: Xia et al. [Xia et al. 2015] and HumanAct12 [Guo et al. 2020]. We use $T = 100$ as the total steps for DDPM. The synthesis starts from an isotropic noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ with conditions on the content and style. After T DDPM steps, we can obtain the synthesized walking motion.

Figure 2 briefly presents the synthesized walking motions with different styles. Our model shows the controlled synthesis by the style and content embeddings. Additionally, our model is potential to be used for style transfer by replacing the original style embedding with a desired style embedding while maintaining the content embedding.



Old Walking

Proud Walking

Strutting Walking

Neutral Walking

Figure 2: Synthesized walking examples with different styles.

4 CONCLUSION AND FUTURE WORK

This paper is the first to use DDPM for walking motion synthesis by conditions and adversarial loss. We show good results of the styled synthesized walking motions.

In future, we plan to synthesize other types of motions with DDPM. Starting from this work on walking motion synthesis, we are going to explore deeper on the potential of using DDPM to synthesize a wider range of motions with high quality and diversity.

REFERENCES

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired Motion Style Transfer from Video to Animation. *ACM Trans. Graph.* 39, 4 (2020).

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2Motion: Conditioned Generation of 3D Human Motions. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-Functioned Neural Networks for Character Control. *ACM Trans. Graph.* 36, 4 (2017).

Daniel Holden, Jun Saito, and Taku Komura. 2016. A Deep Learning Framework for Character Motion Synthesis and Editing. *ACM Trans. Graph.* 35, 4 (2016).

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal Style Transfer via Feature Transforms. (2017). arXiv:arXiv:1705.08086

Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. 2015. Realtime Style Transfer for Unlabeled Heterogeneous Human Motion. *ACM Transactions on Graphics* 34 (07 2015), 119:1–119:10.

M. Ersin Yumer and Niloy J. Mitra. 2016. Spectral Style Transfer for Human Motion between Independent Actions. *ACM Trans. Graph.* 35, 4 (2016).