International Journal of Computer Vision (2023) 131:644–658
https://doi.org/10.1007/s11263-022-01722-5

Bipartite Graph Reasoning GANs for Person Pose and Facial Image Synthesis

Hao Tang[1] · Ling Shao[2] · Philip H. S. Torr[3] · Nicu Sebe[4]

Received: 12 September 2021 / Accepted: 9 November 2022 / Published online: 8 December 2022

Abstract
We present a novel bipartite graph reasoning Generative Adversarial Network (BiGraphGAN) for two challenging tasks: person pose and facial image synthesis. The proposed graph generator consists of two novel blocks that aim to model the pose-to-pose and pose-to-image relations, respectively. Specifically, the proposed bipartite graph reasoning (BGR) block aims to reason the long-range cross relations between the source and target pose in a bipartite graph, which mitigates some of the challenges caused by pose deformation. Moreover, we propose a new interaction-and-aggregation (IA) block to effectively update and enhance the feature representation capability of both a person's shape and appearance in an interactive way. To further capture the change in pose of each part more precisely, we propose a novel part-aware bipartite graph reasoning (PBGR) block to decompose the task of reasoning the global structure transformation with a bipartite graph into learning different local transformations for different semantic body/face parts. Experiments on two challenging generation tasks with three public datasets demonstrate the effectiveness of the proposed methods in terms of objective quantitative scores and subjective visual realness. The source code and trained models are available at https://github.com/Ha0Tang/BiGraphGAN.

Keywords GANs · Bipartite graph reasoning · Person pose synthesis · Facial expression synthesis

1 Introduction

In this paper, we focus on translating a person image from one pose to another and a facial image from one expression to another, as depicted in Fig. 1a. Existing person pose and facial image generation methods, such as Ma et al. (2017); Ma and Sun (2018); Siarohin et al. (2018); Tang et al. (2019c); AlBahar and Huang (2019); Esser et al. (2018); Zhu et al. (2019); Chan et al. (2019); Balakrishnan and Zhao (2018); Zanfir et al. (2018); Liang et al. (2019); Liu et al. (2019); Tang et al. (2019c); Zhang et al. (2020) typically rely on convolutional layers. However, due to the physical design of convolutional filters, convolutional operations can only model local relations. To capture global relations, existing methods such as Zhu et al. (2019); Tang et al. (2019c) inefficiently stack multiple convolutional layers to enlarge the receptive fields to cover all the body joints from both the source pose and the target pose. However, none of the above-mentioned methods explicitly consider modeling the cross relations between the source and target pose.

Rather than relying solely on convolutions/Transformers in the coordinate space to implicitly capture the cross relations between the source pose and the target pose, we propose to construct a latent interaction space where global or long-range (can also be understood as long-distance, which means that the distance between the same joint on the source pose and the target pose very long) reasoning can be performed directly. Within this interaction space, a pair of source and target joints that share similar semantics (e.g., the source left-hand and the target left-hand joints) are represented by a single mapping, instead of a set of scattered coordinate-specific mappings. Reasoning the relations of multiple different human joints is thus simplified to modeling those between the corresponding mappings in the

Communicated by Martin Fergie.

✉ Hao Tang
hao.tang@vision.ee.ethz.ch

[1] Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland

[2] Terminus AI Lab, Terminus Group, Beijing, China

[3] Department of Engineering Science, University of Oxford, Oxford, UK

[4] Department of Information Engineering and Computer Science (DISI), University of Trento, Trento, Italy