College of Engineering Chengannur
Department of Computer Engineering
M. Tech. Computer Science (Image Processing)
03CS6901 Seminar I
Abstract of Proposed Seminar Topic

Deep Spatial Transformation for Pose-Guided Person Image Generation and Animation

CHN20MT007 Praveena K M

December 17, 2020

Keywords: Image Spatial Transformation, Image Animation, Pose-guided Image Generation, Convolutional Neural Networks (CNNs),

Abstract

Pose-guided person image generation and animation aim to transform a source person image to target poses. These tasks require spatial manipulation of source data. we deal with the conditional generation task where the target images are the spatial deformation versions of the source images. Such deformation can be caused by object motions or viewpoint changes. This task is the core of many image/video generation problems.For example, pose-guided person image generation transforms a person image from a source pose to a target pose while retaining the source appearance details. The corresponding pose-guided image animation task further models the temporal consistency and generates a video from a still source image according to a driving target pose sequence.These tasks can be tackled by reasonably reassembling the input data in the spatial domain.

However, Convolutional Neural Networks (CNNs) lack the ability to spatially transform the input features in a parameter efficient manner. One important property of CNNs is the equivariance to transformation, which means that if the input spatially shifts, then the output shifts in the same way. This property can benefit tasks requiring reasoning about images such as segmentation detection , etc. However, it limits the networks by the lack of ability to quires spatially rearranging the input data. In order to enable spatial transformation capabilities of CNNs, Spatial Transformer Networks (STN) introduces a Spatial Transformer module to standard neural networks. This module regresses transformation parameters and warps the input features using a global affine transformation. However, the global affine transformation is not sufficient in representing the complex deformations of non-rigid objects.

The attention mechanism is able to transform information beyond local regions. It gives networks the ability to build long-term dependencies by allowing networks to use non-local features. However, for spatial transformation tasks in which target images are the deformation results of source images, each output position a clear one-to-one relationship with the source position. Theref each output feature is only related to a local region of the so features. In other words, the attention coefficient matrix betw the source and target should be a sparse matrix instead of a d matrix.

Flow-based operation forces the attention coefficient matrix t a sparse matrix by sampling a very local source patch for each put position. It predicts 2D coordinate offsets for the target feat specifying the sampling source positions. However, networks st gle to find reasonable sampling locations when warping the in at the feature level . Possible explanations for this phenome are that: (1)The input features and flow fields change simult ously during the training stage. Their parameter update proce are mutually constrained, which means that the input features not obtain reasonable gradients without correct flow fields and versa. (2) The commonly used Bilinear sampling method prov poor gradient propagation

In order to obtain meaningful flow fields, some flow-based m ods warp input data at the pixel level.However, this operation its the networks to be unable to generate new content. Meanwh large motions are difficult to be extracted due to the requiren of generating full-resolution flow fields Some methods warp the put at the feature level by pre-calculating the flow fields using ditional 3D models or generate dense flow fields from sparse point representation . However, they do not solve the problems straightforward manner, which leads to an insufficient transfor tion representation capability.

Here, we propose a differentiable Global-Flow Local-Atten (GFLA) framework to solve the problems. This framework en CNNs to reasonably sample and reassemble source features with using any labeled flow fields. Specifically, this network can be vided into two parts: Global Flow Field Estimator and Local Ne Texture Renderer. The Global Flow Field Estimator is respons for extracting the long-term dependencies between sources and gets. It estimates flow fields that assign a local source feature pa for each target position. The Local Neural Texture Renderer the extracted flow fields to sample the vivid source neural textu In order to warp sources at the feature level, proposes several

geted solutions to deal with the analyzed problems. First, a Sampling Correctness loss is proposed to constrain flow fields to sample semantically similar regions. This loss helps with the convergence by providing flow fields with additional gradients that are not related to the input source features. Then, a content-aware sampling method is proposed to avoid the poor gradient propagation of the Bilinear sampling. Experiments show that our framework is able to spatially transform the information in an efficient manner. Ablation studies demonstrate that the proposed improvements are helpful for the convergence.

The image-based pose transformation can be further extended for the pose animation task by coherently rendering an input skeleton video. However, most existing models directly apply image transformation methods for this task and generate each video frame independently. This operation does not take the correlations of adjacent frames into consideration, which leads to temporally inconsistent results. In order to obtain coherent results, we make additional efforts to model the temporal dynamics. Thus the input skeleton sequences extracted by popular pose estimation models are always inconsistent. Since these models predict result poses in an image-based manner and do not consider the temporal information of videos, obvious noise can be observed in their results. Therefore, we propose a Motion Extraction Network to extract clean skeleton sequences from the corresponding noise data. Meanwhile,the improved GFLA model to generate video clips in a recurrent manner, allows the model to explicitly extract the correlations between adjacent frames. The main contributions can be summarized as:

• A GFLA model is proposed for deep spatial transformation. Experiments on the pose-guided person image generation task show that the model is able to spatially transform the source neural textures in an efficient manner.

• The temporal consistency is further modeled for the person image animation task.

• The novel view synthesis and face image animation demonstrate that our model can be flexibly applied to other tasks requiring spatial transformation.

References

[1] Yurui Ren;Ge Li;Shan Liu;Thomas H. LiPose. Deep spatial transformation for pose-guided person image generation and animation,. *IEEE Transactions on Image Processing*, 29:8622 – 8635, August 2020.

[2] Q. Sun B. Schiele T. Tuytelaars L. Ma, X. Jia and L. Van Gool. Pose guided person image generation,. *Advances in Neural Information Processing Systems*, page 406–416, 2017.

[3] J. Chen T. H. Li Y. Ren, X. Yu and G. Li. "progressive pose attention transfer for person image generation",. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, june 2019.

[4] J. Chen T. H. Li Y. Ren, X. Yu and G. Li. "deep image spatial transformation for person image generation",. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 7690–7699, 2020.

[5] P. Esser and E. Sutter. "a variational u-net for conditional appearance and shape generation",. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8857–8 june 2018.

[6] J. L. Hennessy and D. A. Patterson. *Computer Architectu Quantitative Approach*. Morgan Kaufmann Publishers Inc. Francisco, Second edition, 1996.