RESCON 2025 CMS 32

TEMPORAL AWARE TEXT-DRIVEN STYLE TRANSFER FOR MOTION-BASED VIDEO TRANSFORMATION

A.D.D.S. De Silva* and R. Siyambalapitiya

Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka
*addsdesilva8@gmail.com

Video stylisation plays a crucial role in creative domains such as virtual reality, game development, content creation, and filmmaking. However, existing traditional video style transfer methods often produce flickering and popping due to inconsistent frame stylization. They typically rely on reference style images or computationally heavy modules such as neural atlas layers, making them unsuitable for real-time use. This research introduces a deep learning-based framework for text-guided style transfer on single or multiple objects in videos, focusing on temporal consistency and low computational cost. Two approaches are proposed; combining You Only Look Once version 8 (YOLOv8) for object detection, Deep Simple Online and Realtime Tracking (DeepSORT) for tracking, and the Segment Anything Model (SAM) for precise segmentation. Stylisation is performed using CLIPStyler, which applies descriptive text prompts as style instructions, making the process fully text-driven and independent of reference images. A custom dataset of elephants was created and annotated for training and evaluation. In the first approach, stylisation was applied only to segmented object frames, which were then blended with unaltered background frames. This method is efficient but slightly affects background quality. To address this, the second approach first generated a fully stylised video and then used segmentation masks to isolate stylised objects, merging them with the original background frames. This preserved background clarity while maintaining object stylisation. Quantitative evaluation produced strong results: Intersection over Union (IoU) = 0.9689, dice coefficient = 0.9689, F1-score = 0.88, precision = 0.90, and recall = 0.87. A user study with 30 participants, including professional videographers, found that over 90% agreed the style aligned with the text, object shapes were preserved, and background artefacts were minimal. These results demonstrate the framework's effectiveness for object-level stylisation. Future work will explore advanced detection models, improved segmentation with SAM 2, zero-shot object detection, and voice-based style control.

Keywords: CLIPStyler, Temporal consistency, Text-guided style transfer, Video object segmentation