Self-supervised video decomposition for video editing

Supervisors: Javier Vazquez-Corral, Danna Xue (CVC/UAB)

Video editing [1-3] presents unique challenges compared to image editing, as it requires modeling frame-to-frame relationships and addressing temporal variations such as motion and illumination changes (see Figure 1). This complexity arises from the added dimension of time, making it challenging for users to maintain consistency across all frames while applying edits. Simplifying video editing to the level of image editing would be more convenient, allowing users to make edits to individual frames with ease. This approach would enable seamless propagation of editing results across the entire video while handling spatiotemporally varying components.

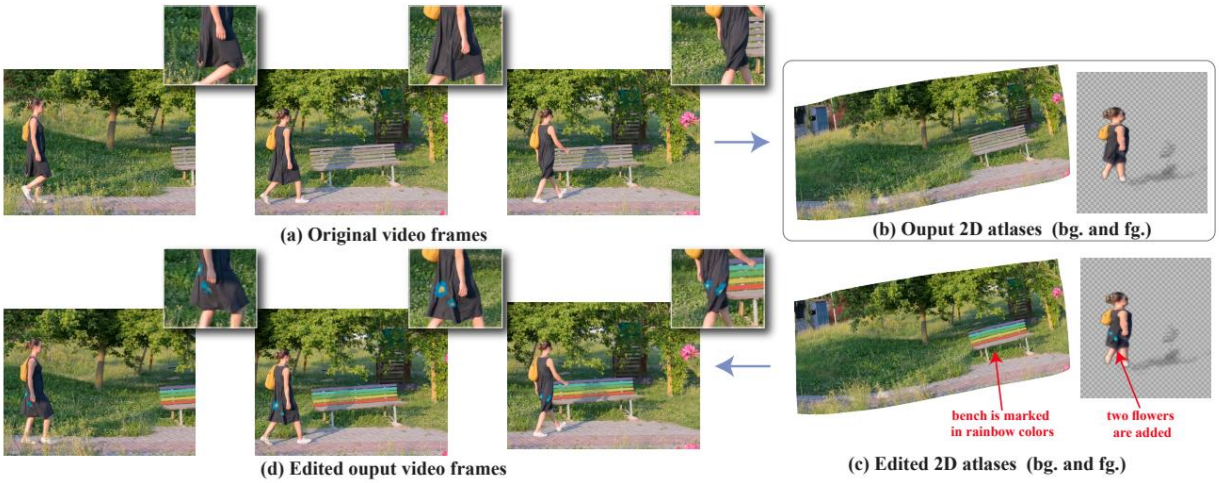

Figure 1. An example of the workflow of video decomposition for video editing (Image from [1]). Given a natural video as input (a), the model estimates a set of layered 2D atlases (b), each providing a unified and interpretable parameterization of an object's or background's appearance throughout the video. Any edits made to the atlases (c) to be consistently and automatically propagated to all frames of the video (d).

In this project, our goal is to learn an implicit neural representation model based on this video from a given input sequence. The model can decompose the video into different layers, with each layer containing information about an object with independent motion (foreground) or the scene (background). The model is trained in an unsupervised manner without any ground truth mask for the objects.

Project key points:

- Build up the self-supervised video decomposition algorithm framework.
- Reduce the importance of high-quality initial masks by introducing image feature priors, e.g. CLIP or DINO features.
- Optimize the processing of multi-object videos.

- Apply a new motion estimation model [4].
- Enhance the quality of reconstructed images by adaptively learning hash encoding parameters and network model parameters using a hypernetwork [5].
- Create a simple image editing interface (optional).

References:

[1] Kasten, Yoni, et al. "Layered neural atlases for consistent video editing." *ACM Transactions on Graphics (TOG)* 40.6 (2021): 1-12.

[2] Ye, Vickie, et al. "Deformable sprites for unsupervised video decomposition." *CVPR,* 2022.

[3] Chan, Cheng-Hung, et al. "Hashing Neural Video Decomposition with Multiplicative Residuals in Space-Time." ICCV, 2023.

[4] Chugunov, Ilya, et al. "Neural Spline Fields for Burst Image Fusion and Layer Separation." *arXiv preprint arXiv:2312.14235* (2023).

[5] Sen, Bipasha, et al. "HyP-NeRF: Learning Improved NeRF Priors using a HyperNetwork." *Advances in Neural Information Processing Systems* 36 (2024).