The Animation Transformer: Visual Correspondence via Segment Matching

We thank the reviewers for their thoughtful feedback. We are encouraged they found our work to be worthy of pursuit (**R1**), intuitive (**R2**) and interesting (**R4**), our idea to represent hand-drawn images as segments valuable (**R3**), and our use of the cycle-consistency loss for self-supervision to be reasonable and effective (**R2**, **R3**, **R4**). We address reviewer comments below and will incorporate all feedback.

@R2,R4 - Time complexity and memory complexity: A single forward pass of AnT takes on average **76ms (13 FPS)** on a Nvidia Tesla V100 GPU. Using M and N to denote the number of reference and target segments, each cross attention layer AnT has to make $\mathcal{O}(MN)$ comparisons and each self attention layer AnT has to make $\mathcal{O}(M^2 + N^2)$ comparisons. By comparison, a forward pass of DEVC takes on average **147ms (6 FPS)**. Memory-wise, DEVC has to make $\mathcal{O}((HW)^2)$ comparisons, where H,W are the spatial dimensions of the CNN features. We were limited to using a batch size of 3 for DEVC, whereas we could use a batch size of 64 for AnT, yielding much faster training. Our leak-proof filling method implemented in OpenGL takes on average **1.4s** on the same hardware, yielding a total inference speed of **2.16s** for AnT or **2.87s** for DEVC.

@R1,R2 - Applicability to shading and/or textures: In this work we focused on the use case of flat coloring since it covers the vast majority of uses in the cel animation production. However, the contents of each region need not be limited to flat colors – if shaded regions are drawn, they can be regarded as normal fill regions and processed by our method. Similarly, texture masks can be copied across regions. In future work, optical flow predicted from the matching output could be used to warp textured regions.

@R4 - HD Results: Figures 12, 13, 14 (supplementary information) show HD results of our approach on wide variety of settings. We accidentally downscaled the images in the main submission file and will fix that in the revised version.

@R1 - Why operate on raster images instead of vector? We put a lot of thought into this question and ultimately came out in favor of raster due its flexibility – a vector image can always be converted to raster whereas existing raster-to-vector approaches often break down on production scale raster artwork. We also surveyed 118 professional cel animators and found that over 50% spend more time drawing in raster than drawing in vector.

@R1 - Geometric representations are better than pixels: We agree that geometric formats such as vector would be a rich representation for this task, however the goal of our approach is to be able operate on raster images (for reasons described above). That said, AnT avoids many of the pitfalls of operating directly on raw pixels by exploiting the presence of segments and using a transformer to learn the structural relationships between segments.

@R1 - Artists leave gaps in their artwork: We use a leak-proof filling method to handle accidental gaps which can handle most commercial animations. While not every artist closes gaps, our user tests indicate that this covers a very wide range of artwork types. We will provide additional samples in the revised version on different artwork styles.

@R1 - Applicability to in-betweening: As shown in [1], segment matching can be used to help with optical flow prediction and in-betweening.

@R2 - Significance of self-attention: Self-attention aggregates information about salient relationships between segments in the same image, which helps to disambiguate between potential matches in the cross attention layers. For example, if two eyes are above a nose in the reference image they should also have a similar layout in the target image.

@R2 - Concatenating the embeddings: In high dimensional space the CNN and positional features form two smaller, distinct subspaces that are approximately orthogonal to each other and can be manipulated independently of each other by a learned transformation. This allows us to reduce the number of trainable parameters in the transformer.

@R2,R3,R4 - Notation and typos: We agree with these suggestions and will make corrections in the revised version. $\mathbf{z}$ are the intermediate features in the transformer, $\mathbf{f}$ are the final features used in the correlation matrix.

@R4 - New segments in the target frame: Our method only matches segments across the reference and target frames, so it cannot colorize new segments that appear in the target frame. In future work we plan to incorporate multiple reference frames to solve this issue.

@R2 - Comparison with other pixel based approaches: We will add a comparison with [33] in the revised version. However, [5] does not include enough information and code for reproduction.

@R2 - How does categorical cross-entropy loss allow for one-to-many mappings? Loss is computed for the color prediction of each region in the target image. A color from a single region in the reference image can be copied to multiple regions in the target image.

References

[1] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C. C. Loy, and Z. Liu. Deep animation video interpolation in the wild. In *CVPR*, 2021. 1