Supplementary Material
The Animation Transformer: Visual Correspondence via Segment Matching

Evan Casey[1] Víctor Pérez[1] Zhuoru Li[2]
[1]Cadmium [2]Project HAT

In the following pages, we present additional background information, experimental details, qualitative examples of AnT in action, user study results, as well as visualizations and analysis of the learned attention patterns.

1. Industry standards in animation

To understand the motivation behind AnT, it is important to consider how hand-drawn animation is produced at studios today. The vast majority of animation is produced at HD (1080 x 1920) or beyond resolution on digital drawing tablets or scanned in from pencil drawings. Once it is converted to a uniform line drawing (also known as a clean line), artists colorize by clicking on each individual line enclosure (segment) with a uniform color and flood filling it with a color. This painstakingly laborious process can take on the order of minutes per frame to do manually for complex animation.

The traditional style of flood filling each line enclosure has been around since the dawn of animation and continues to be the de facto standard because it allows the artist to quickly color many images in a short amount of time. For an assistive colorization tool to be effective in this domain, it is crucial that it integrates easily with this workflow and thus produce predictions at the level of segments. By doing so, this also enables the artist to manually intervene and correct any mistakes with the existing flood fill tool they are accustomed to.

2. Pitfalls of pixel-based approaches

One approach is to combine the output of a pixel-based model with flood fill segmentation information and choose the maximally occurring color in each segment. We explored this approach and highlight several issues that can occur.

In Figure 1 we use the popular open-source colorization model PaintsChainer and provide a color hint for every segment. The model is trained with an MSE loss in RGB space, so it learns to predict colors that are close to the user-provided color palette. When multiple colors are used, it quickly starts mixing the reference colors and diverging from the user-specified color palette.

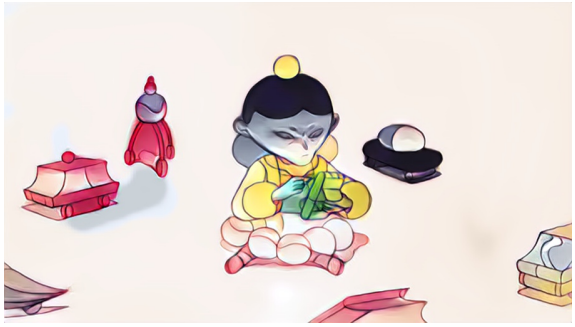

Figure 1: **Color mixing in PaintsChainer.**

To overcome the color mixing issue we can train a pixel model with categorical cross entropy loss by discretizing the input color palette into a compact label space. We use this approach with the correspondence network in Deep Exemplar Video Colorization (see Figure 2). The resulting output stays true to the provided color palette, but the model loses important details due to input downsampling and max-pooling in the CNN backbone (both of which are necessary to compute pixel attention on a 16GB GPU). We use DEVC in our benchmarks, but convert the pixel output to segment labels for evaluation.

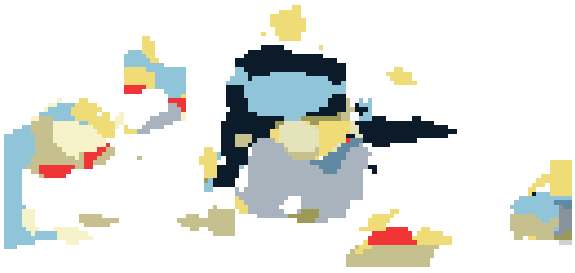

Figure 2: **Raw output of the correspondence subnetwork of DEVC.**

3. Choice of evaluation metrics

Given that our task is to output predictions at the level of segments, how do we measure performance? Existing

metrics for pixel-based tracking and colorization tasks are not suitable: a practical metric would roughly approximate how many corrections an animator would need to make to correct any inaccurate predictions. Since the artists make corrections at the level of segments, this begs the need for segment-level evaluation metrics. Thus, we define two evaluation metrics that are specifically suited for the task: **Accuracy** and **Mean IoU**. We describe each of these in detail and discuss their connection with other evaluation metrics in colorization and tracking.

Accuracy is defined as the percentage of correct segment-level label predictions averaged over all segments in each of the target sequences. In colorization, this is somewhat analogous to MSE in RGB color space – we want to predict the right color label and penalize incorrect colors. However, unlike in photo-realistic colorization we are predicting from a discrete set of labels.

Mean IoU is defined as the mean Intersection-over-Union for each segment averaged over all segments in the target sequence. In the video segmentation context, our Mean IoU metric is analogous to Region Similarity $\mathcal{J}$. However, instead of measuring the similarity between pixel regions we are measuring in the level of segments.

4. User study

To evaluate our approach we conducted a user study. We ask professional artists to colorize sequences (see Figure 3) from the real dataset without and with the assistance from AnT. In the test with Ant, we colorized the sequences with AnT then asked users to check and correct incorrect parts in the results. All tests were done in professional software, and we record users' interactions and work time. The summary result is shown in Table 1. We can see that AnT significantly increases the work efficiency.

5. Qualitative results

Comparison with other methods In this section, we show results of our proposed approach (AnT) to: DEVC, Lazy Brush, EBSynth, Style2paints. LazyBrush fails to handle large movements but fills segments with a uniform color, making it suitable for animation workflows. EBSynth similarly degrades with large movements but is not segment-aware so it blends pixels together. Style2paints is not suitable for animation colorization task.

Additional results: In Figure 5 we show qualitative examples of a variate set of sequences colorized with AnT and DEVC. In the same way that previous qualitative examples, these colorization sequences have been created following a recursive propagation of colors, using each colorized image as input for the next generation (as described in figure 8 in main body). AnT presents superior performance especially when dealing with ambiguous segments and occlusions. In Figure 6 we show results from line drawing with gaps.

6. Inspecting Attention in AnT

In Figure 7 we present the attention patterns formed in the attention layers of the Transformer module at different stages. The visualizations are created for the case where target segment features are updated, i.e. self-attention is computed between segments from the target image and cross-attention aggregates segment information from the reference image to each target segment. The opacity of green lines represents the attention weight between a target segment and each segment from the contrary image. For example, in the first-row of self-attention, the segment A has small attention weights towards multitude of other target segments while in the last row of cross-attention its attention is mainly focused on the correct correspondence from the reference image.

Selected segments: We have chosen two segments where our model correctly found correspondences in situations where more than just visual information was necessary. In these cases, the spatial and structural information provided by the positional encoder and the Transformer was key to disambiguate correct correspondences from wrong matches. We show the robustness of AnT to occlusions with segment A and its ability to find the correct correspondences in ambiguous scenarios with B (which shares visual resemblance with its neighboring segments).

Attention patterns: From our experiments, we can appreciate how attention focuses on gathering information from lots of segments from the contrary images in early layers. We argue that segment representations get benefited from attending a large number of segments all around the image to get a sense of the global structure of the scene and its relative distances with other segments. Towards the later layers, attention gets progressively narrowed towards the most important elements to represent each segment. This is important to disambiguate between similar segments. For example, in the latest row of the cross-attention layer, both segments still gather information from enclosures close to them such as the hand for A or other pills for B.

Case	Human				AnT+Human				Interactions	Time
	Mouse click	Key down	Interactions	Time (s)	Mouse click	Key down	Interactions	Time (s)	**(AnT+Human / Human)**	**(AnT+Human / Human)**
A	180	629	809	174.10	31	16	47	56.16	5.81%	32.26%
B	402	1013	1415	429.53	93	206	299	119.57	21.13%	27.84%
C	365	1497	1862	369.09	131	700	831	140.85	44.63%	38.16%
D	605	897	1502	550.23	138	551	689	169.93	45.87%	30.88%
E	2151	5058	7209	1826.52	90	167	257	169.70	3.56%	9.29%
F	280	849	1129	237.22	79	270	349	91.41	30.91%	38.53%

Table 1: **User study result.** Comparison of user effort on colorization task without/with assistance from AnT. "Mouse click" interactions mainly include switching, moving, and zooming the canvas, filling, and picking colors. "Key down" interactions mainly include toggling tools, file operations, undo/redo.

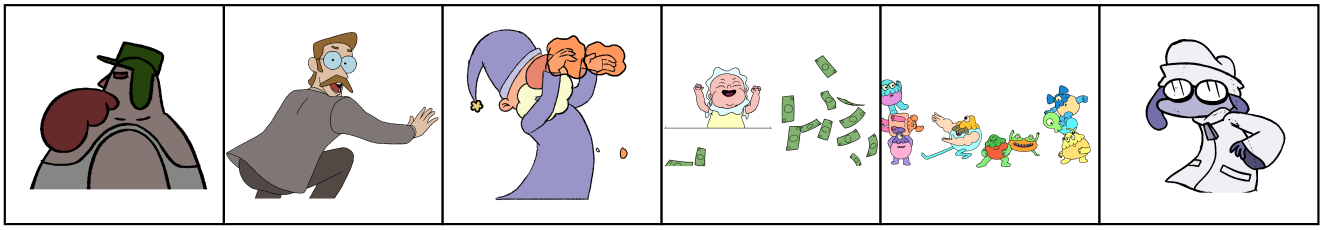

Figure 3: **Samples from sequences used in the user study.**

Figure 4: **Comparison with other methods.**

Figure 5: **Qualitative results for AnT and DEVC.** Zoom in to view in more detail.

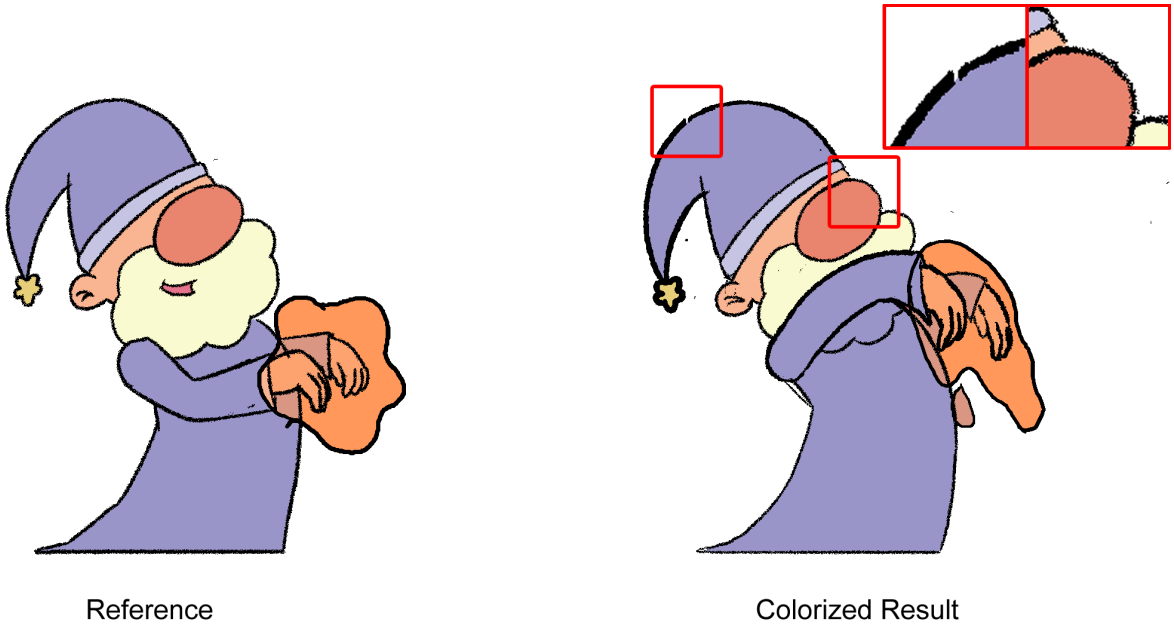

Figure 6: **Our method can handle line drawings with gaps.** Zoom in to view in more detail.

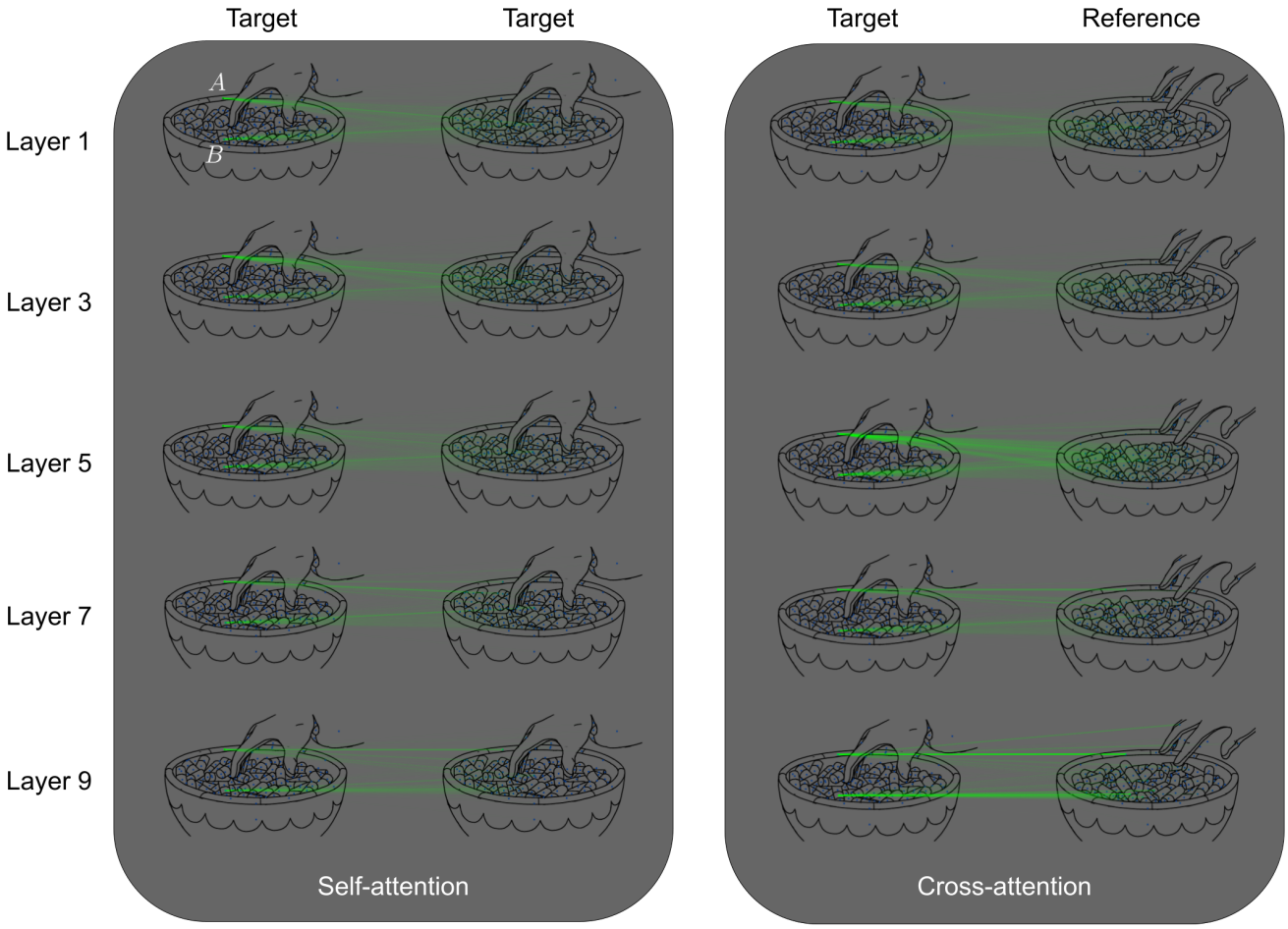

Figure 7: **Self- and cross- attention layer visualizations for two segments.** The locations of segments A and B are shown in the top left-hand corner.