Supplementary | Crowd-Guided Ensembles: How Can We Choreograph Crowd Workers for Video Segmentation?

OVERVIEW

The supplementary material is split into parts:

This document includes additional figures and details of the evaluation that could not fit within the main paper.

The naive segmentation results are provided in the "`results`" directory where each subdirectory corresponds to a sequence of DAVIS. Scribbles will be available publicly but take too much space to be within the supplementary.

Instructions for crowdsourced tasks are available in printable format in `segmentation_instructions.pdf` and `scribble_instructions.pdf`.

Web interfaces for crowdsourced tasks are provided in the "`tasks`" directory. An internet connection is not required to browse them. They were tested with Firefox 52 on Ubuntu 16.04 and Firefox 51 on Mac OS X 10.10.5.

- `index.html` — the main segmentation task.
- `videoReview.html` — the scribbling task.

In the segmentation task, a segmentation result is provided as `car-example.json` which you can load from within the interface (data > load > browse file).

Video sessions are available in the "`sessions`" directory. They include a segmentation session, a scribble session, a summary review session to accept/reject in batches, as well as a short display of the bonus review session. Each video is encoded with H.264 such that it should be readable in most modern players including browsers.

ADDITIONAL EVALUATIONS AND FIGURES

Additional Results for F and T Metrics

In the main paper, we only showed scribble evaluation figures for the J metric because of space constraints. We compare these here with the F and T metrics where appropriate.

Figure 1 confirms that the joint distribution of brush sizes and result qualities are somewhat similar for J and F metrics. The boundary accuracy F seems to show a slightly higher concentration of better qualities with small brushes.

The evaluation of tradeoffs for varying scribble replications is provided in Figure 2. J and F metrics show similar trends.

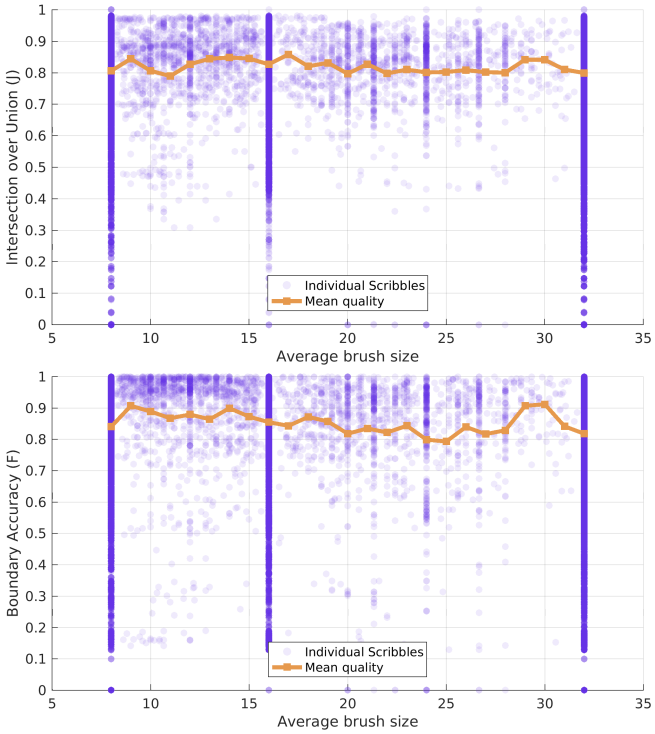

Figure 1: Joint distribution of brush sizes and corresponding result qualities. The available brushes were of size $b = 8, 16, 32$ and we display here the average brush size for a scribble that led to some result for all sequences with sampling $P = 25$.

We also included full evolutions for the *pxor* strategy and the figure shows that it is clearly less stable. In some cases, it is actually detrimental and requires enough scribbles before they can be merged in a useful manner. The T metric is better when lower, and thus the Y axis trend is reversed. It also shows that we can achieve higher temporal stability using propagation methods instead of the full segmentation. The trends beyond iteration $P = 3$ cannot be easily extrapolated, so we do not include them.

Figure 3 confirms similar trends between metrics with respect to the brush regularization effects for both *pxor* and *wmaj* merging strategies. Expectedly, using a larger regularization ($f > 1$) is detrimental for both techniques because it uniformly lessens the confidence in the worker's brush strokes. For the corrective fixing, it likely introduces false positive regions w.r.t. the worker intents.

ACM ISBN 978-1-4503-5620-6/18/04…$15.00
DOI: https://doi.org/10.1145/3173574.3173685

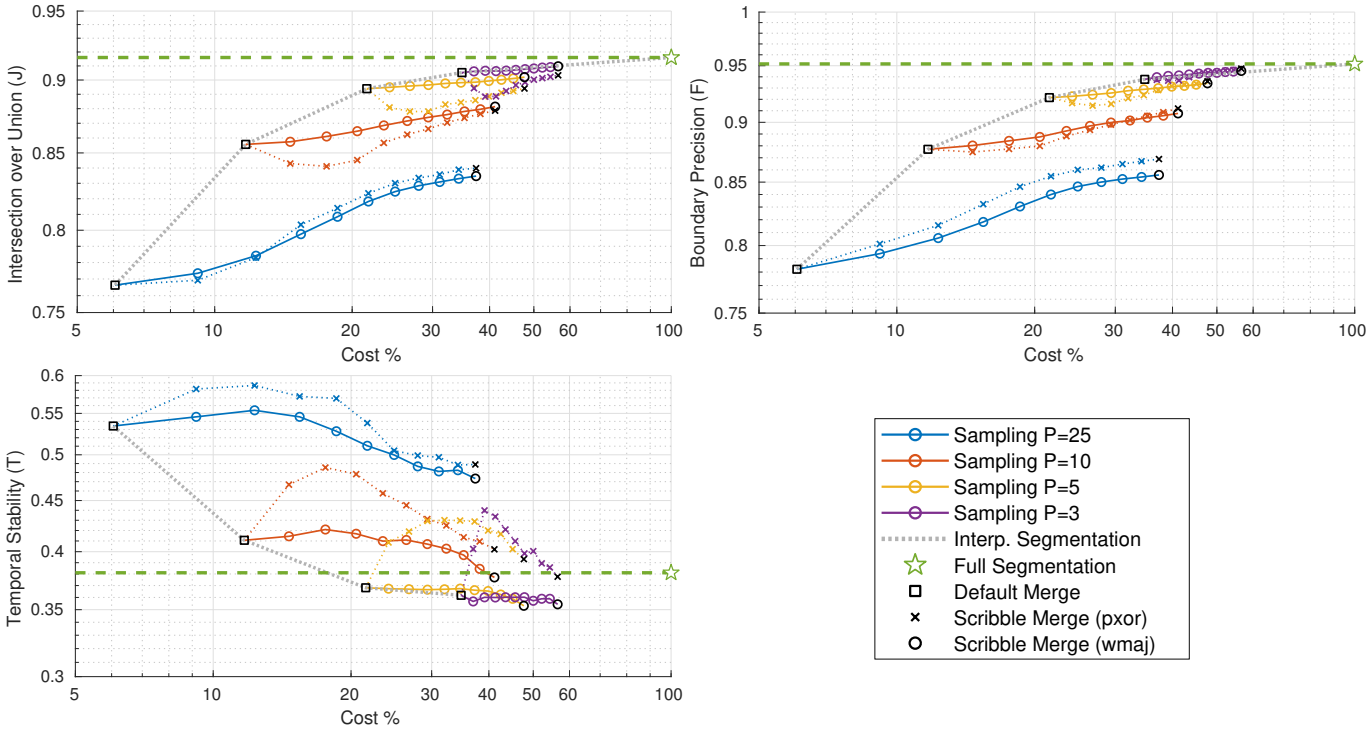

Figure 2: Evolution of the propagated quality with respect to the acquisition cost for different sampling intervals. The successive markers correspond to merging propagations using pixelwise majority (square), increasing scribbles replications using *wmaj* (circles) and *pxor* (crosses). The costs are represented as percentages of the naive full segmentation cost (star).

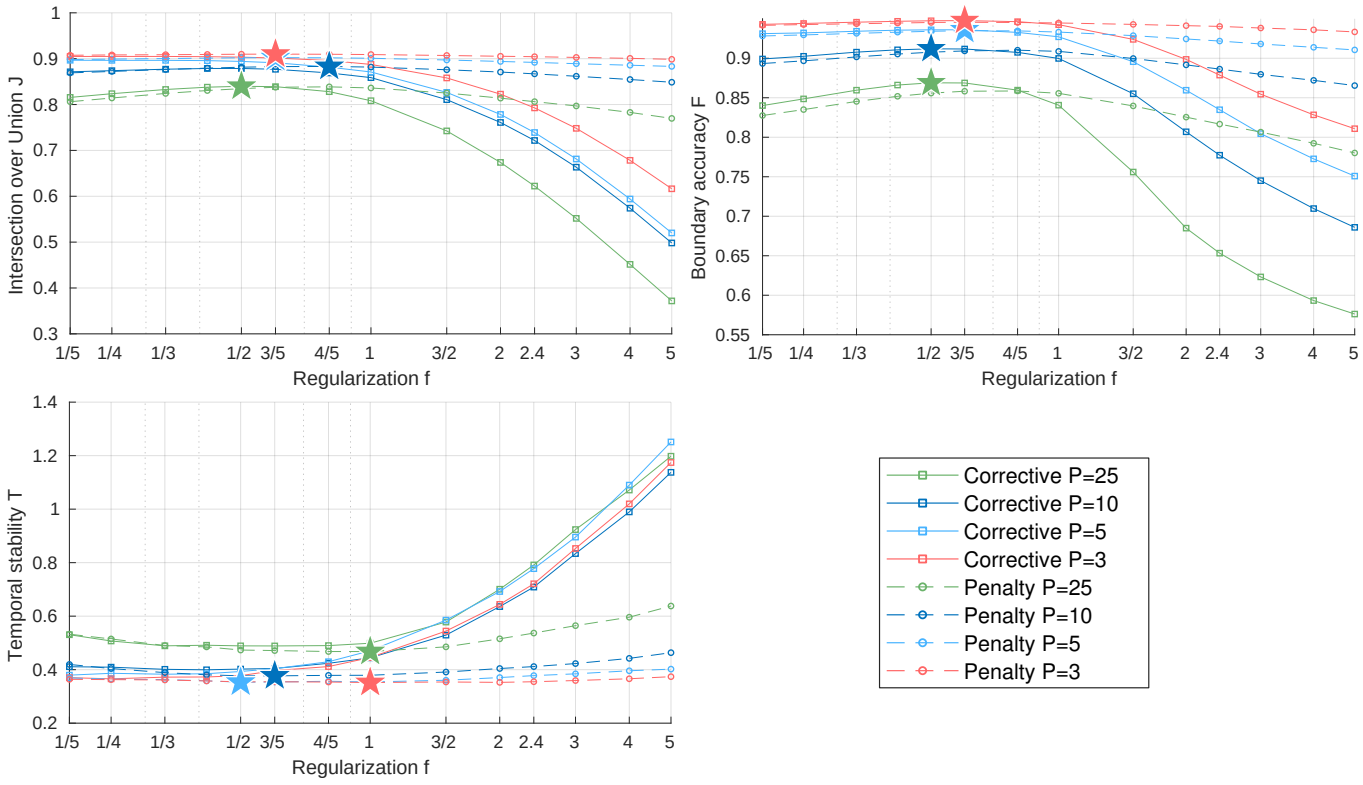

Figure 3: Impact of the brush regularization f on scribbles at different samplings P for both the corrective (*pxor*) and penalty-based merging (*wmaj*) strategies. For each sampling value, the best regularization and method are marked with a star.

As for smaller regularization ($f < 1$), we decided to to not allow smaller brushes to avoid workers focusing on segmentation fixing since we need scribbles to take less time than segmentations. This decision was justified by some workers who contacted us to ask for smaller brush sizes stating for example "[I] need smaller brush tools and also a pen tool".

The need for a *pen tool* illustrates that crowd workers can have preconceptions that go against our own intention. Many workers have been accustomed to segmentation given the large need of image segmentation in the machine learning communities. As a result, our new scribble task was implicitly labeled by some workers as a mean to directly correct the propagated segmentations. Our aim was to avoid such direct corrective fixing, and thus we limited the brush sizes.

This results in workers not being able to take care of very fine details by themselves as intended. However, a region that is collectively viewed as bad by multiple workers is more likely to really be defective. This implies that we can go beyond the limit of the brush sizes we impose by using the collective decision of the crowd, without needing the individual workers to use a smaller brush size. We interpret the effective smaller regularization as a potential confirmation that the collective decision is actually more precise than individual limited brush strokes.

Figure 4 also confirms that J and F trends are similar when increasing the number of propagation methods.

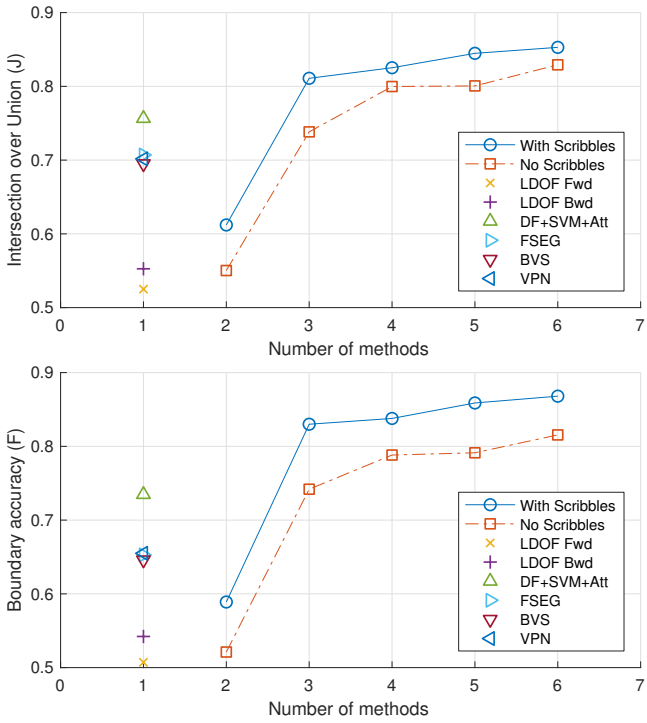

Figure 4: Evolution of the quality with increased number of propagation methods. The methods being used are, in increasing order: LDOF forward, LDOF backward, DF+SVM+Att, FSEG, BVS, VPN.

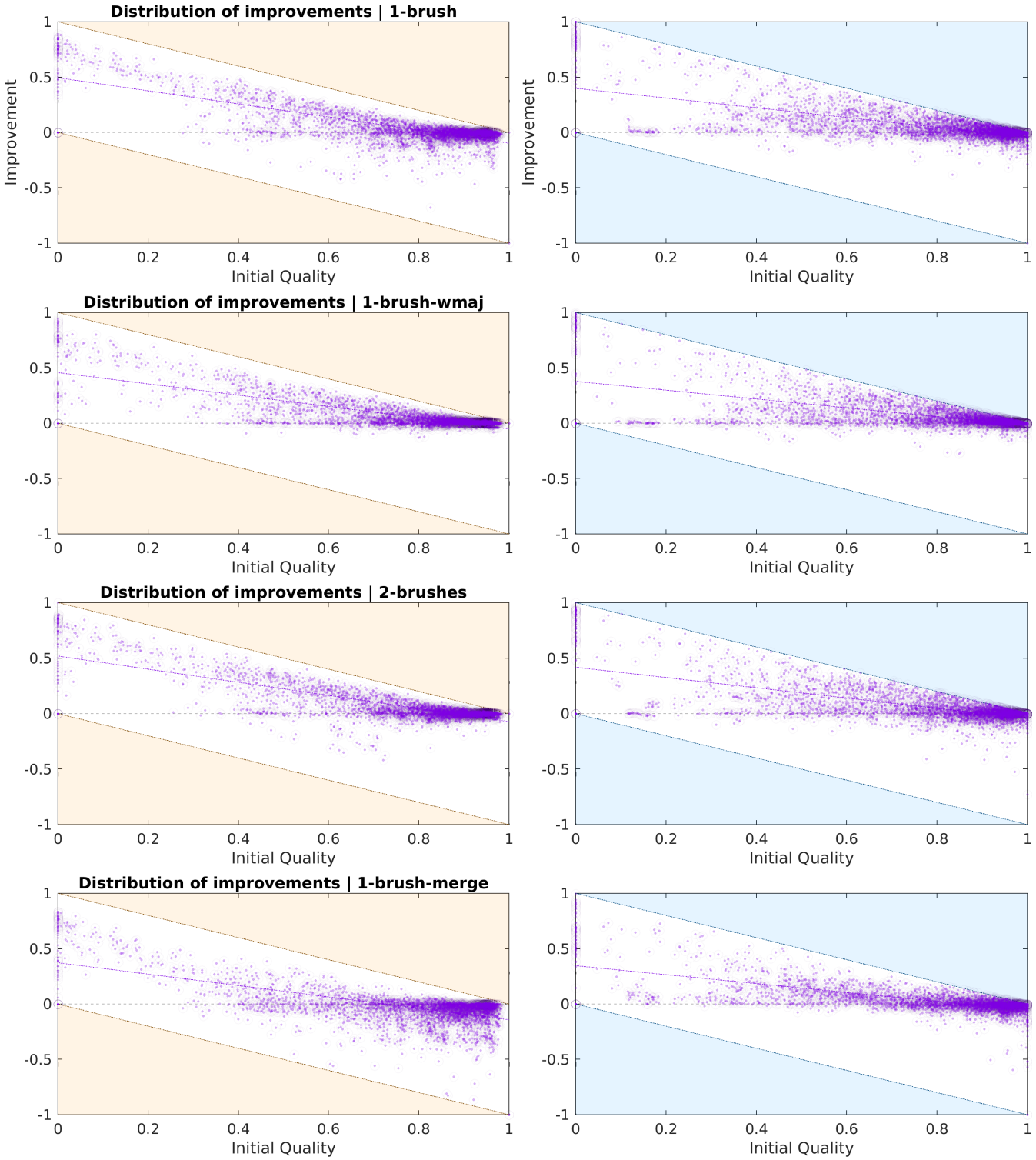

Figure 5: Distributions of scribble quality improvement. Each line corresponds to a different merging strategy: from top to bottom, *pxor*, *wmaj*, *2-brushes* and *pxor* merging before scribbles are acquired. The left column corresponds to J quality whereas the right column is for F. Each scatter point corresponds to a set of scribbles for a single frame of DAVIS. The quality metrics J and F are confined within the interval $[0;1]$. The colored regions correspond to improvements that are not possible by definition: when starting with $J = 0$, no negative improvement can happen whereas starting with $J = 1$ can only lead to stagnation or some degradation. The middle dashed line $y = 0$ corresponds to no improvement or degradation. The blue dotted line following the data is a linear fit to the datapoints.

Distribution of Scribble Impacts

Figure 5 provides a look at the full distribution of scribble impacts in terms of how much quality improvement each individual scribble set led to. This is done with each scribble merging strategy *pxor*, *wmaj* and *2-brushes* as well as *pxor* when merging before acquiring the scribbles.

Merging before acquiring scribbles has the worst J distribution and it does indeed lead to quality degradation as shown in the paper when not using brush regularization (all these plots assume $f = 1$). Using *pxor* is less stable as can be seen with the several negative outcomes, especially compared to the more stable *wmaj* and *2-brushes* which have very few negative outcomes.

Time Analysis

In the paper, we show the cumulative timings from Figure 6 including each replication of segmentation and scribbles for the varying samplings $P = 25, 10, 5, 3$ and compared with the full segmentation. We describe here how these timings were computed.

For the segmentation tasks, we had already originally a log of all the actions with timestamps, thus we directly used it for the time analysis. For the scribble tasks, we unfortunately did not have such log for the main analysis, and thus we recomputed the timings using a time model we inferred from a sparse set of new tasks with a full timestamped log of actions.

In both cases, we did not consider the active time as the difference between acceptance of the task and submission because this time was often very large for no good reason. Some workers seemingly stack tasks and then complete them, resulting in the difference between acceptance and submission having many outlier timings. Instead, we considered the recorded mouse actions and use only the cumulated active time.

Scribble Timing Model

For our scribble tasks, we re-computed the timings of our initial full-scale evaluation by modeling the time of brush strokes. We model the scribble brush strokes with two types of strokes: *single-dot* strokes, and *multi-dots* strokes. Furthermore, we created a different model for each brush size $b = 8, 16, 32$ as smaller brush sizes require more attention and this seems to lead to longer brush strokes for similar amounts of mouse displacement.

For the multi-dots strokes, we model a stroke as a sequence starting with a *press* event, followed by consecutive repeated mouse *moves*, and finished with a *release* event. The exact timings depend on the worker using the interface, their computer speeds, environments and operating system. We do not have access to most of this, and instead model the general average case, which we use to infer the original timings of our large-scale scribble experiment.

To create our model, we sent a new scribble task for each sequence of DAVIS for each sampling $P = 25, 10, 5, 3$, recording the worker's mouse events while scribbling.

See Table 1 for our effective brush stroke model parameters. For single-dot brush strokes, our model is defined by the average cumulated press and release times. For multi-dots brush strokes, it consists of an average press time (which we found to be mostly negligible), an average release time, as well as a linear model fitting the time it takes to move given the distance (defined by an offset Δt_0 and a time factor α).

As we expected, the timings are smaller the larger the brush size is, which we interpret as a need for higher attention when the brush is smaller, leading to longer timings.

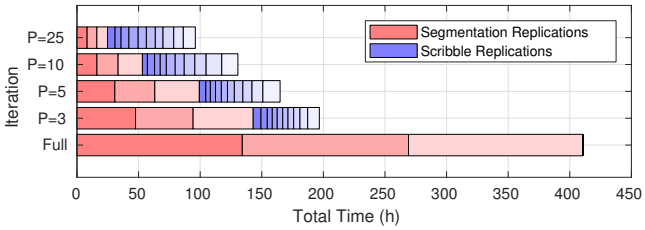

Figure 6: Cumulative and parallel timings of the segmentation and scribble replications for sampling intervals $P = 25, 10, 5, 3$ as well as the full segmentation. The cumulative timings show that even though scribbles are shorter, the need to apply them on a large portion of the full sequences (i.e. all frames minus the segmentations), makes them actually quite expensive. A full replication $R = 10$ leads to a total scribble time larger than the segmentation time.

Metric	Brush Size 8	16	32	Unit
Average single-dot time	142	125	123	ms
Average press time	0.18	0.16	0.15	ms
Average release time	169	126	69	ms
Move time offset Δt_0	26.6	27.1	26.2	ms
Move time factor α	4720	2233	489	ms/npx
Average move time	37.9	35.0	29.9	ms
Observations	2.2M	2.9M	4.0M	·

Table 1: Brush stroke model parameters. The units *ms* correspond to milliseconds, and *npx* are distances in a normalized image viewport with boundaries within $[0, 1]^2$.