STRIVE: Scene Text Replacement In Videos
Supplemental Section

Vijay Kumar B G [*1], Jeyasri Subramanian[2], Varnith Chordia[3], Eugene Bart[2], Shaobo Fang[4], Kelly Guan[5], and Raja Bala[3]

[1]NEC Laboratories, America
[2]Palo Alto Research Center
[3]Amazon
[4]Work done at PARC
[5]Stanford University

1. Video Demonstrations

We include a set of 13 videos comparing the performance of STRIVE with frame-wise SRNet on scenes from our three datasets: *Synthtext, Robotext, RealWorld* as summarized in Table 1. The videos can be viewed at `https://striveiccv2021.github.io/STRIVE-ICCV2021/`.

The videos cover challenging indoor and outdoor scenes with varied text styles and geometries, different types of camera and object motion, and lighting effects. Each video is a montage with the **original clip on the left, STRIVE output in the middle and SRNet framewise output on the right**.

2. Comparisons with Pix2Pix

In the main paper we compare STRIVE, SRNet and pix2pix on quantitative metrics in Table 1. Here we include qualitative results. Fig. 1 compares STRIVE and pix2pix outputs on reference frames for 4 videos. Note that on the reference frame, STRIVE and SRNet produce identical outputs, hence are not duplicated in this figure. As seen in these examples, pix2pix does not perform well. Our explanation is that pix2pix is a general-purpose style transfer method. Although trained on text ROIs from our video datasets with temporal consistency loss over multiple frames, its architecture is not specialized for text replacement. Therefore it is unable to compete with techniques like SRNet which is explicitly designed to transfer foreground text style to new content while preserving the background. Given that pix2pix is a weak baseline, we do not include it in the video demonstrations.

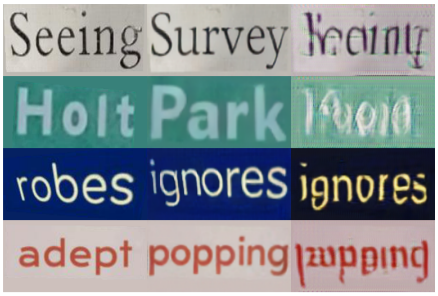

Figure 1: From left to right: Input reference frame, STRIVE, Pix2Pix. Top two are real videos, and bottom two are synthetic videos.

*Corresponding author: bg.vijay.k@gmail.com

Filename	Dataset	Source-to-Target Text	Key Observations
video1	Synthtext	Depts to Envoy	Demonstrates importance of disentangling geometry from style. SRNet fails to replace text with strong perspective distortion.
video2	Synthtext	Cincinnati to Wisconsin	Demonstrates importance of temporal consistency constraints in STRIVE that produce temporally smooth output. SRNET output exhibits jitter, although the style transfer is of high quality.
video3	Robotext	SECT to LOSE	STRIVE preserves pose and is more temporally stable than SRNet.
video4	RealWorld (ICDAR)	Monde to World	Performance in presence of walking motion. STRIVE exhibits superior temporal stability compared with SRNet.
video5	RealWorld (ICDAR)	Oranges to Apples	Performance in presence of walking motion. STRIVE is temporally smooth, while SRNET output exhibits jitter, deforms the character "a" near the end, and exhibits undesirable color shifts at the text ROI boundary.
video6	RealWorld	Ball to Prom	STRIVE is robust to complex lighting changes. Note SRNet failures towards end of clip.
video7	RealWorld	CLEAN to CASH	STRIVE effectively predicts text blur via BPN and mimics changes in text sharpness due to depth defocus. SRNet is unable to successfully transfer text distorted by blur.
video8	RealWorld	Seeing to Survey	STRIVE and SRNet both effectively track changes in text appearance due to depth defocus.
video9	RealWorld	Holt to Park	Challenging noisy video with vehicle motion and atmospheric distortion. STRIVE maintains temporal consistency and character integrity, while SRNet deforms the character "a" from frame to frame.
video10	RealWorld	COFFEE to ROBERT	STRIVE preserves geometry, while SRNet fails under strong perspective in the final segment.
video11	RealWorld	Palm to Cold	Challenging due to small text and vehicle motion. STRIVE output is considerably smoother than SRNet.
video12	RealWorld	OSCAR to VIOLA	STRIVE is robust to rapid nonlinear camera motion.
video13	RealWorld	THURLBY to FAIRPOT	SRNet exhibits considerably more jitter than STRIVE

Table 1: **Explanation of Video Demonstrations**