Low Light Video Enhancement using Synthetic Data Produced with an Intermediate Domain Mapping

Danai Triantafyllidou, Sean Moran, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh

Huawei Noah's Ark Lab

1 Supplementary material

We provide further material and details that supplement our main paper, notably by including qualitative results on dynamic videos. We report additional results as follows:

- Qualitative results on dynamic test videos (section 1.1)
- Quantitative results on dynamic video using the Temporal warping error (section 1.2)
- Experimental study on real-data training quantity and ratio (section 1.3)

1.1 Qualitative results on Dynamic Test Videos

Our supp. material package includes example video results. Using the DVR test data [1], we provide sets of example output video files (test-time inferences) where each video set explores a different ratio of synthetic and real training data (see section 1.3 for further detail). Within each data ratio sub-directory, we provide a video per test scene (`M0002, M0014, M00015, M0016`). Each video corresponds to a comparison between two models; (1) training using only real data *vs*. (2) training using both synthetic and real data.

Large qualitative temporal stability improvements can be observed when the model is trained with both synthetic and real data *vs*. training using real data only. For each video; we note that the the split screen comparative result shows (1) training a model using only real data and (2) training on both synthetic and real data on **left hand** and **right hand** side of the split screen, respectively.

1.2 Quantitative results on Dynamic Videos

In this section, we provide quantitative results on dynamic videos using our synthetic data generation approach and compare it to the SID Motion model for variable amounts of real training data. Temporal warping error provides a metric for quantifying the temporal stability of dynamic video frames. Lower error indicates higher temporal stability. Given a pair of video frames V_t, V_{t+1} at times t and t+1, the temporal warping error E_{warp} is defined as in Equation 1:

$$E_{warp}(V_t, V_{t+1}) = \frac{1}{\sum_{i=1}^{N} M_t^i} \sum_{i=1}^{N} M_t^i \|V_t^i - \hat{V}_{t+1}^i\|_2^2 \tag{1}$$

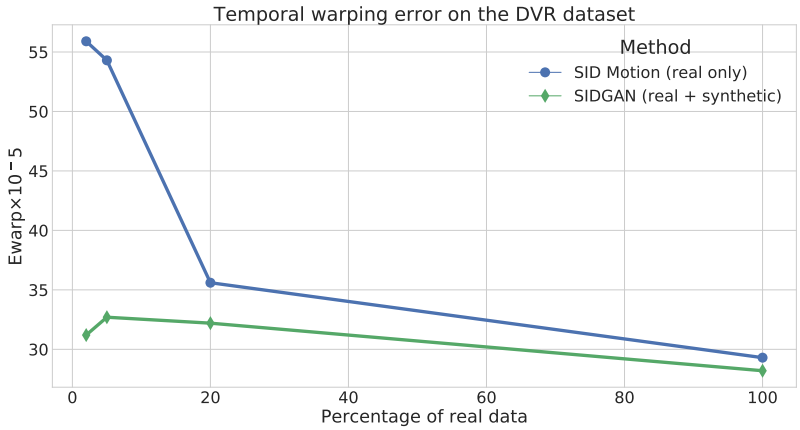

Fig. 1: Temporal warping error [2], E_{warp}, averaged over the DVR [1] *dynamic* video test set for the SID Motion RAW-to-RGB model. The model is trained on both increasing fractions of only real data (blue) and the trained on increasing fractions of real and synthetic data (green).

where $\hat{V}_{t+1}$ is the optical flow warped frame V_{t+1}, $M_t^i \in \{0, 1\}$ is a non-occlusion mask for pixel i, estimated using [3]. For a video consisting of T frames the average warping error is defined in Equation 2:

$$E_{warp}(V) = \frac{1}{T-1} \sum_{t=1}^{T-1} E_{warp}(V_t, V_{t+1}). \tag{2}$$

In Figure 1 we observe that the SID Motion forward RAW-to-RGB model has lower temporal warping error (improved temporal stability) when trained with a mixture of real and synthetic data generated by our SIDGAN, as compared to the same model trained purely with real data. This lends support to our hypothesis; *the addition of synthetic dynamic video data leads to enhanced temporal stability.* Furthermore, as the amount of real training data is reduced, the difference in temporal stability become both quantitatively larger and visually more obvious (*c.f.* accompanying video results). For example, the SID Motion model trained with 2% real data plus synthetic data has a temporal warping error ∼45% lower than the model trained using only 2% real data (31.2×10^{-5} *vs.* 55.9×10^{-5} respectively).

In summary, particularly in cases where real data collection budgets are limited, it is highly beneficial to augment the training dataset with synthetically generated dynamic videos for our target task. We provide evidence that this affords significant improvement for the temporal stability of the RAW-to-RGB video mapping task.

1.3 Real, Synthetic training data ratio

We investigate the effect of varying the quantity of real training data whilst also considering a fixed quantity of synthetically generated training data. We evaluate our performance in comparison with the full (exclusively) real data test performance (100%) by alternatively using a fraction the real data available and supplementing this with our synthetically generated data from SIDGAN. In summary we explore a mixture of real and synthetic training data with various real data ratios. We randomly re-sample seven subsets of the original training data (DVR dataset). Subsets comprise of $2\%, 5\%, 10\%, 20\%, 40\%, 60\%$ and 80% of the full dataset respectively. Table 1 summarizes the number of training samples corresponding to each subset.

We adopt the architecture described in [1] (Unet with 16 residual layers) and compare the performance of our model in two training scenarios. Our baseline experiment aims to investigate the effect, on model performance, of reducing available real data (no synthetic data augmentation). The second set of experiments involve training the same model on a collection of 9366 synthetic videos and then fine-tuning the model using available real data. For fair comparison, both models are trained using the same hyper-parameters. We perform 14 different training experiments and quantitatively evaluate image quality after 1000 epochs using Peak-noise-to-signal ratio (PSNR) and Structure Similarity (SSIM).

Tables 2 and 3 provide summary statistics for PSNR and SSIM scores derived from all subsets. The set of models trained using only (fractions of) real data achieve maximum PSNR and SSSIM scores of 28.17 and 0.81 respectively. Comparatively, models trained using both synthetic and real data consistently, for every subset ratio considered, result in performance boosts; improving the max PSNR and SSIM scores to 28.44 and 0.83 respectively. We also observe that the extra data, afforded by training the model using both synthetic and real videos results lower model performance standard deviation across subset size, indicating that augmenting with SIDGAN synthetic data can slightly improve performance stability and predictability.

Figure provides a visualizations of model comparison across all data subset scenarios considering PSNR, SSIM image quality metrics. Firstly, it can be observed that training with extremely small ($\leq 5\%$) amounts of real data significantly reduces performance for both model variants. However, even when considering these extreme scenarios, using both synthetic and real data results in significant boosts in relative performance increasing PSNR from 17.70 to 22.32, from 21.35 to 23.35 and from 24.04 to 25.19 for the cases of $2\%, 5\%$ and 10%, respectively. As the fraction of real data is increased, the gap in performance between the considered models can be observed to reduce, highlighting that the

addition of synthetic data is most valuable in scenarios where we are extremely real-data hungry *e.g.* when collection of real training samples is expensive or indeed not possible. Finally, Figures 3, 5 and 4 provide comparison of model inference examples. Large visual improvements are consistently most evident in the scenarios where real-data is most scarce.

Fig. 2: Peak signal-to-noise-ratio (PSNR) versus Structure similarity (SSIM) using a fraction of the real data

Table 1: Real data ratios and corresponding number of frames for each domain

split	**real ratio**	**# videos**	**# frames C**	**# frames B**
1	2 %	3	332	3
2	5 %	6	671	6
3	10 %	13	1443	13
4	20 %	25	2780	25
5	40 %	52	5758	52
6	60 %	78	8611	78
7	80 %	103	11156	103

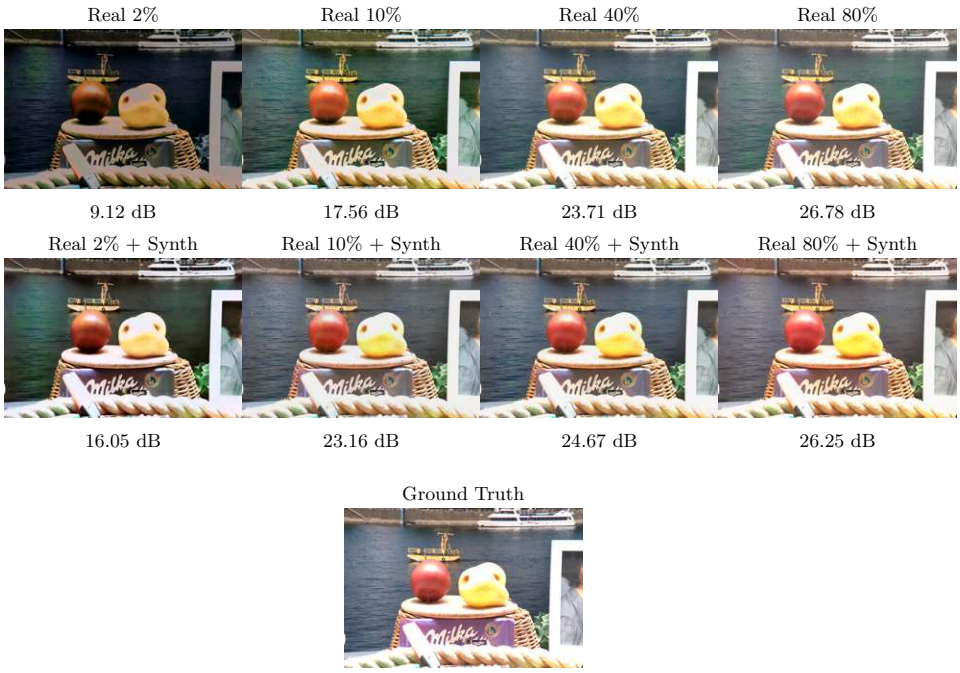

Fig. 3: Visual comparisons of test-frame inference resulting from models trained using only "Real" (top row) and "Real + Synthetic" (middle row) data subsets

Table 2: Descriptive statistics for PSNR scores across all splits at epoch 1000.

Model	**Mean**	**Std**	**Min**	**Max**	**50%**	**75%**
Real	24.530	3.795	17.707	28.173	26.148	27.115
Real + Synth	26.074	2.473	22.338	28.445	26.769	28.163

Table 3: Descriptive statistics for SSIM scores across all splits at epoch 1000.

Model	**Mean**	**Std**	**Min**	**Max**	**50%**	**75%**
Real	0.741	0.085	0.599	0.817	0.776	0.807
Real + Synth	0.790	0.043	0.720	0.831	0.804	0.824

Fig. 4: Visual comparisons of test-frame inference resulting from models trained using only "Real" (top row) and "Real + Synthetic" (middle row) data subsets

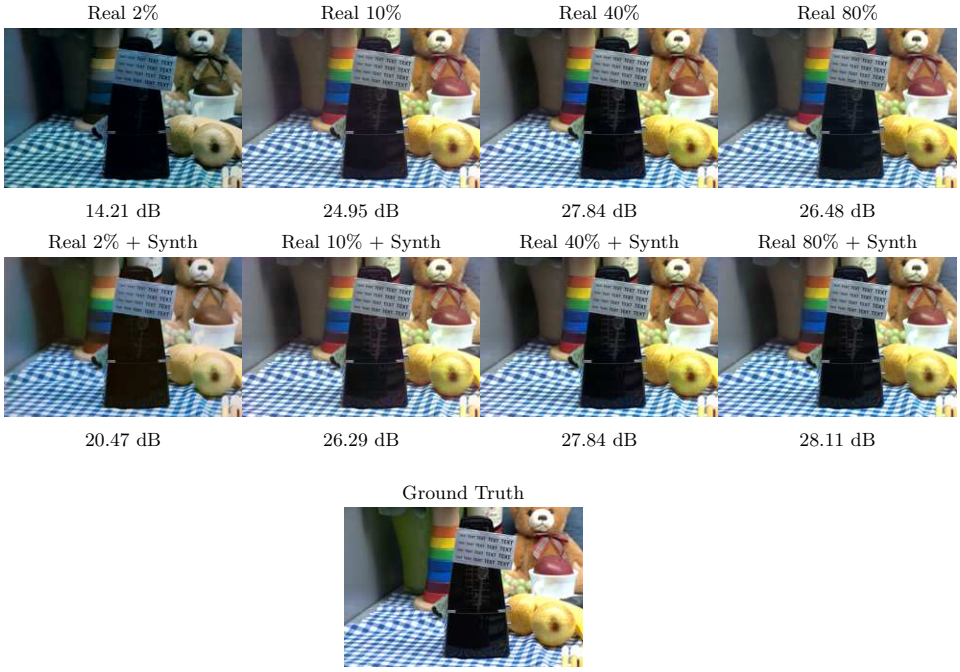

Fig. 5: Visual comparisons of test-frame inference resulting from models trained using only "Real" (top row) and "Real + Synthetic" (middle row) data subsets

References

1. Chen, C., Chen, Q., Do, M.N., Koltun, V.: Seeing motion in the dark. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
2. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: European Conference on Computer Vision (2018)
3. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos. In: German Conference on Pattern Recognition (GCPR) (2016), http://lmb.informatik.uni-freiburg.de/Publications/2016/BD16