Vid2Pix - A Framework for Generating High-Quality Synthetic Videos

Oda O. Nedrejord*†, Vajira Thambawita*‡, Steven A. Hicks*‡, Pål Halvorsen*‡, and Michael A. Riegler*
*SimulaMet, Norway †University of Oslo, Norway ‡Oslo Metropolitan University, Norway

***Abstract*—Data is arguably the most important resource today as it fuels the algorithms powering services we use every day. However, in fields like medicine, publicly available datasets are few, and labeling medical datasets require tedious efforts from trained specialists. Generated synthetic data can be to future successful healthcare clinical intelligence. Here, we present a GAN-based video generator demonstrating promising results.**

***Index Terms*—Deep learning, generative adversarial networks, data up-sampling, video generation.**

I. Introduction

Data-driven technology has become ingrained in all areas of modern society, and healthcare is no exception. For example, machine learning-based systems have shown tremendous results in automatic detection of gastrointestinal (GI) anomalies for colonoscopies (e.g., [1], [2]). Despite these impressive results, these methods do not generalize well [3]. This is mostly due to a lack of training data as making medical data public is difficult, i.e., due to legal restrictions, patient privacy, and a manual time-consuming, tedious labelling task for trained medical experts.

Generated "fake" synthetic data can be the key to successful clinical and business intelligence [4], [5]. Therefore, in this paper, we present our Vid2Pix system that takes existing datasets and generates synthetic videos using a generative adversarial network (GAN). As an initial use-case, we use data collected from GI colonoscopies where anomalies are often missed and overlooked. We limit our scope to polyp videos, but the presented method should generalize well to other domains as well. The realism of the generated data is evaluated by two medical doctors, and quantitative measurements. The results suggest that the generated synthetic data is sometimes indistinguishable from real data and can, in the future, be used as training data for machine learning-based algorithms.

II. The Proposed Method: Vid2Pix

Using a dataset collected from two hospitals in Norway containing $83,088$ video frames, downsized to 128×128, we developed a system that can create more data from data we already have. Specifically, we aim to generate artificial videos of colon polyps by using real videos of colon polyps. Our system can be broken down into three distinct steps:

1) Skip Frames using Dense Optical Flow (step 1): A high frame rate combined with inconsistent camera movements causes inconsistencies in the videos. To address this problem, we first process the videos by using dense optical flow. Since the movement direction is not critical to solve our problem, we only consider the magnitude of the motion to decide whether to keep or to skip a frame using a threshold of 20% above the average magnitude between each continuous frame in a video. We create each video with a fixed length of 8 frames. If the difference in frame numbers are larger than 10 frames, we create a new video to avoid large jumps in the videos. With the method, we managed to optimize the dataset by removing duplicate frames and large jumps between frames.

2) Future Frame Generation with Vid2Pix (step 2): Our proposed architecture is a conditional GAN [6] that uses a generator and discriminator based on Pix2Pix [7]. Pix2Pix was developed to translate an *image* in one domain to an image in another domain. However, we are trying to learn past *image sequences* (videos) in one domain to generate future sequences in the same domain. Thus, our Vid2Pix system is a generative model that predicts a future frame conditioned on the past frames in a sequence.

We first add an additional dimension in order to use the temporal dimension to generate realistic motions. The additional dimension leads to a replacement of 2D with 3D convolutions and deconvolutions. The 3D convolutions extract features from the temporal dimension as well as the spatial dimension. Instead of using 2D convolutions as Pix2Pix does for down- and upsampling, we use 3D convolutions for both operations (to ensure support features). We use the additional dimension to input temporal information by stacking frames through that dimension. The height, width, and channels dimensions are used to input spatial features of input frames. The discriminator outputs a downsampled feature map from either a concatenation of the input sequence and a generated image or from a concatenation of the input sequence and the ground truth. The discriminator is a PatchGAN [7] .

3) Pipeline of predicting frame sequence (step 3): In order to generate a video in Vid2Pix, we need to iterate over the model with shifted input several times. Figure 1 shows how we generate a video from generated images. The Vid2pix model generates one image at a time as depicted in Figure 2.

III. Quality Evaluation of Generated Videos

As an initial step in evaluating our GAN-based system, we generated videos consisting of four frames and calculated a dense optical flow visualization between each frame (Figure 3). An **initial inspection** suggests that they look realistic, and the dense optical flow proves that the model also learned to capture the correct movement of the videos.

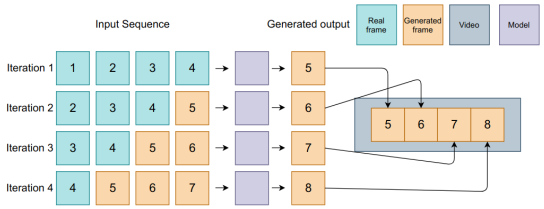


Fig. 1: Frame pipeline to create a video from Vid2Pix.

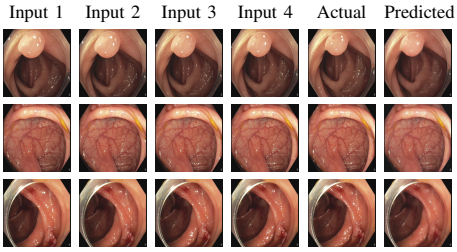
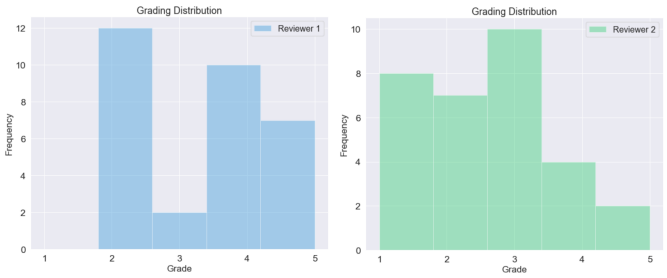


Fig. 2: Videos with four input frames representing the stacked input to the model, the ground truth and the predicted output.

Fig. 3: Generated four frames and the corresponding dense optical flow between consecutive frames. Hue values represent the direction and the amount of motion.

Furthermore, to assess the realism of the generated content, we conducted a **subjective human assessment** to evaluate the generated videos. We recruited two medical doctors with endoscopy data experience, i.e., a medical doctor with two years of experience, and a gastroenterologist with extensive experience. The assessment is divided into two sessions, where we for each provided detailed information. The *classification* session involves classifying videos into two classes, either *real* or *fake*, where ten were artificial, and ten were real. Our results show that a total of six real videos were miss-classified as fake, and six videos were miss-classified as real when they were fake. Moreover, in a *grading* assessment, the reviewers assessed 31 fake videos. For each video, they were asked to give scores from one to five where one is least real and five is most real. Figures 4a and 4b show the grading results. The average grade from the first reviewer (the junior doctor) is 3.4, and the average grade from the second reviewer (the senior doctor) is 2.8. Overall, the doctors found many examples where it was hard to differentiate between real and fake as the shapes and colors appeared realistic, but the differences indicate that there is room for improvement, especially since the participants found some examples of strange motions and tissues.

Finally, we assessed the system using objective **similarity measures**. Using the generated frame and the corresponding ground truth on all generated videos we calculated the mean square error (MSE), peak signal to noise ratio (PSNR) and structural similarity (SSIM) values. Using 626 videos, we achieved respective PSNR, MSE, and SSIM averages



(a) Reviewer 1 (b) Reviewer 2

Fig. 4: Subjective grading distributions of 31 generated videos.

of 72.1301, 0.0050, and 0.8011 for Pix2Pix, and 73.3718, 0.0042, and 0.8409 for Vid2Pix. From these numbers, we observe that when we modify the original Pix2Pix model by using our intermediate experiments, such as predicting four frames at once, the SSIM and PSNR values first decrease, and MSE increases. Finally, SSIM shows good values for our last model modification, where we changed the model to predict one image instead of a sequence of images and reduced the discriminator complexity. We conclude that reducing the discriminator complexity and changing the output dimension has a positive effect on the quality of generated output.

IV. Conclusion

We have developed a conditional GAN to generate "fake" synthetic future frames using real videos as input. The key parts of the model were the 3D convolutional and deconvolutional layers creating realistic-looking spatio-temporal features. Moreover, to improve quality, we implemented a dense optical flow-based preprocessing framework, which could filter away stationary frames of a video. From our quantitative measurements, the MSE, PSNR, and SSIM metrics show that the Vid2Pix model outperforms the Pix2Pix model for artificial video generation. We also found that experienced medical doctors struggle to differentiate between real and synthetic videos, which indicates that synthetic videos look real. Still, there is a large room for improvement, and we currently work on model enhancements and trying different use-cases.

References

[1] G. Urban *et al.*, "Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy," *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078.e8, Oct. 2018.

[2] M. Riegler *et al.*, "EIR - Efficient computer aided diagnosis framework for gastrointestinal endoscopies," in *Proc. of CBMI*, Jun. 2016, pp. 1–6.

[3] V. Thambawita *et al.*, "An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification," *ACM Trans. Comput. Healthcare*, vol. 1, no. 3, Jun. 2020.

[4] A. Rebhan, "How - and why - health organizations are using synthetic health care data," Advisory Board, Nov. 2019. [Online]. Available: https://www.advisory.com/research/health-care-it-advisor/it-forefront/2019/11/synthetic-health-data

[5] B. Siwicki, "Is synthetic data the key to healthcare clinical and business intelligence?" Healthcare IT News, Feb. 2020. [Online]. Available: https://www.healthcareitnews.com/news/synthetic-data-key-healthcare-clinical-and-business-intelligence

[6] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

[7] P. Isola *et al.*, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proc. of CVPR*, Jul. 2017, pp. 5967–5976.