CLTS-GAN: Color-Lighting-Texture-Specular Reflection Augmentation for Colonoscopy

Shawn Mathew¹*, Saad Nadeem²*, and Arie Kaufman¹

¹Department of Computer Science, Stony Brook University
²Department of Medical Physics, Memorial Sloan Kettering Cancer Center

Abstract. Automated analysis of optical colonoscopy (OC) video frames (to assist endoscopists during OC) is challenging due to variations in color, lighting, texture, and specular reflections. Previous methods either remove some of these variations via preprocessing (making pipelines cumbersome) or add diverse training data with annotations (but expensive and time-consuming). We present CLTS-GAN, a new deep learning model that gives fine control over color, lighting, texture, and specular reflection synthesis for OC video frames. We show that adding these colonoscopy-specific augmentations to the training data can improve state-of-the-art polyp detection/segmentation methods as well as drive next generation of OC simulators for training medical students. The code and pre-trained models for CLTS-GAN are available on Computational Endoscopy Platform GitHub (https://github.com/nadeemlab/CEP).

Keywords: Colonoscopy · Augmentation · Polyp Detection.

1 Introduction

Colorectal cancer is the fourth deadliest cancer. Polyps, anomalous protrusions on the colon wall, are precursors of colon cancer and are often screened and removed using optical colonoscopy (OC). During OC, variations in color, texture, lighting, specular reflections, and fluid motion make polyp detection by a gastroenterologist or an automated method challenging. Previous methods deal with these variations either by removing specular reflections [13,14], removing color/texture [15], and correcting lighting [23] in the preprocessing steps (making pipelines cumbersome) or by adding more diverse training data with expert annotations (but expensive and time-consuming). If the automated methods can be made invariant to color, lighting, texture, and specular reflections without adding any preprocessing overhead or additional annotations, then these methods can act as effective second readers to gastroenterologists, improving the overall polyp detection accuracy and potentially reducing the procedure time (end-to-end colon wall inspection from rectum to cecum and back).

We present a new deep learning model, CLTS-GAN, that provides fine-grained control over creation of colonoscopy-specific color, lighting, texture, and specular reflections. CLTS-GAN is a new deep learning model that gives fine control over color, lighting, texture, and specular reflection synthesis for OC video frames. We show that adding these colonoscopy-specific augmentations to the training data can improve state-of-the-art polyp detection/segmentation methods as well as drive next generation of OC simulators for training medical students. The code and pre-trained models for CLTS-GAN are available on Computational Endoscopy Platform GitHub (https://github.com/nadeemlab/CEP).

Keywords: Colonoscopy · Augmentation · Polyp Detection.

1 Introduction

Colorectal cancer is the fourth deadliest cancer. Polyps, anomalous protrusions on the colon wall, are precursors of colon cancer and are often screened and removed using optical colonoscopy (OC). During OC, variations in color, texture, lighting, specular reflections, and fluid motion make polyp detection by a gastroenterologist or an automated method challenging. Previous methods deal with these variations either by removing specular reflections [13,14], removing color/texture [15], and correcting lighting [23] in the preprocessing steps (making pipelines cumbersome) or by adding more diverse training data with expert annotations (but expensive and time-consuming). If the automated methods can be made invariant to color, lighting, texture, and specular reflections without adding any preprocessing overhead or additional annotations, then these methods can act as effective second readers to gastroenterologists, improving the overall polyp detection accuracy and potentially reducing the procedure time (end-to-end colon wall inspection from rectum to cecum and back).

We present a new deep learning model, CLTS-GAN, that provides fine-grained control over creation of colonoscopy-specific color, lighting, texture, and specular reflections.
specular reflection augmentations. Specifically, we use a one-to-many image-to-image translation model with Adaptive Instance Normalization (AdaIN) and noise input (StyleGAN [12]) to create these augmentations. Color and lighting augmentations are performed by injecting 1D vectors (sampled from a uniform distribution) using AdaIn, while texture and specular reflection augmentations are incorporated by directly adding 2D matrices (sampled from a uniform distribution) to the latent features. The color and lighting vectors can be extracted from one OC image and used to modify the color and lighting of another OC image. We show that these colonoscopy-specific augmentations to the training data can improve accuracy of the state-of-the-art deep learning polyp detection methods as well as drive next generation OC simulators for teaching medical students [7]. The contributions of this work are as follows:

1. CLTS-GAN, an unsupervised one-to-many image-to-image translation model
2. A novel texture loss to encourage a larger variety in texture and specular generation for OC images
3. A method for augmenting colonoscopy frames that produces state-of-the-art results for polyp detection
4. Latent space analysis to make CLTS-GAN more interpretable for generating color, lighting, texture, and specular reflection

2 Related Works

The image-to-image translation task aims to translate an image from one domain to another. Certain applications have access to ground truth information providing supervision for models like pix2pix [11]. Zhu et al. developed CycleGAN, an image-to-image translation model without needing ground truth correspondence [5]. This is done using a cycle consistency loss that drives other unsupervised domain translation models. Examples include MUNIT [9] and Augmented CycleGAN [11] which additionally incorporated noise to learn a many-to-many domain translation. This many-to-many mapping lacks control over specific image attributes. XDCycleGAN [17] and FoldIt [16] model one-to-many image-to-image translation, however their networks functionally learn a one-to-one mapping.

Generating realistic OC from CT scans has been used for OC simulators. VRCaps uses a rendering approach to simulate a camera inside organs captured in CT scans [10]. For the colon, a simple texture is mapped on a mesh where OC artifacts (e.g., specular reflections, fish-eye lense distortion) are added. However, it cannot produce complex textures and colors normally found in OC. OfGAN uses image-to-image translation with optical flow to transform colon simulator images to OC [22]. It uses synthetic colonoscopy frames embedded with texture and specular reflection, which improve the realism of generated images. The texture and specular mapping in the synthetic frames, however, restrict additional texture and specular generation. Rivoir et al. use neural textures to create realistic and temporally consistent textures [19]. They require a full 3D mesh to embed the neural textures making it difficult to augment annotated real videos.
Fig. 1. (a) shows user specified noise being used in $F$. $z_{ts}$ is a set of 2D matrices that goes through convolutional layers and is added to latent features throughout the network. $z_{cl}$ is a 1D vector that goes through fully connected layers and is distributed to AdaIn layers. Both $z_{ts}$ and $z_{cl}$ are sampled from a uniform distribution and can be sampled until the user is satisfied with the result. (b) depicts the forward cycle where OC passes through $G$, predicting its noise vectors and VC. These are then passed into $F$ to reconstruct the image. $F$ produces another OC image using different noise vectors where $L_{adv}$ is applied. (c) depicts the backwards cycle where a VC image with different $Z_{ts}$ is passed into $F$. The resulting two OC images have $L_{text}$ applied. One OC image is used for reconstruction via $G$ where $L_{cyc}$ is applied.

3 Data

10 OC videos and 10 abdominal CT scans for virtual colonoscopy (VC) were obtained at Stony Brook University Hospital. The OC videos were rescaled to 256x256 and cropped to remove borders. Since the colon is deformable and CT scans capture a single time point, there is no ground truth correspondence between OC and VC. The VC data uses triangulated meshes from abdominal CT scans similar to [18]. Flythroughs were generated using Blender with two lights on both sides of the camera to replicate a colonoscope. Additionally, the inverse square fall-off property was applied to accurately simulate lighting conditions in OC. A total of 3000 VC and OC frames were extracted. 1500 were used for training while 900 and 600 were used for validation and testing.

4 Methods

CLTS-GAN is composed of two generators and three discriminators. One generator, $G$, uses OC to predict VC with two corresponding noise parameters. The
first parameter, $z_{ts}$, is a number of matrices that represent texture and specular reflection information. The second parameter, $z_{cl}$, is a 1D vector that contains color and lighting information. The second generator, $F$, uses $z_{ts}$ and $z_{cl}$ to transform a VC image into a realistic OC image. Figure 1a shows how the noise values are used in $F$. $z_{cl}$ is incorporated using AdaIn layers, which globally affects the latent features. $z_{ts}$ is directly added to latent features offering localized information. The complete objective function for the network is defined as:

$$L_{obj} = \lambda_{adv}L_{adv} + \lambda_{cyc}L_{cyc} + \lambda_{t}L_{t} + \lambda_{idt}L_{idt} \quad (1)$$

Cycle consistency is used in many image-to-image translation models and ensures features from the input are present in the output when transformed. The cycle consistency loss used for OC is shown in Figure 1b and defined as:

$$L_{OC}^{cyc}(G,F,A) = E_{x \sim p(A)} \| x - F(G_{im}(x), G_{cl}(x), G_{ts}(x)) \|_1 \quad (2)$$

where $x \sim p(A)$ represents a data distribution and $G_{im}$, $G_{cl}$ and $G_{ts}$ represents $G$’s output. Since $G$ has additional outputs, the cycle consistency loss should incorporate these extra vectors as seen in Figure 1c.

$$L_{VC}^{cyc}(G,F,A,Z) = E_{x \sim p(A), z \sim p(Z)} \| x - G_{im}(F(x, z_{cl}, z_{ts})) \|_1 + \| z_{cl} - G_{cl}(F(x, z_{cl}, z_{ts})) \|_1 + \| z_{ts} - G_{ts}(F(x, z_{cl}, z_{ts})) \|_1 \quad (3)$$

The cycle consistency component of the objective loss function is defined as:

$$L_{cyc} = L_{cyc}^{OC}(G,F,OC) + L_{cyc}^{VC}(G,F,VC,Z) \quad (4)$$

Each generator has a discriminator, $D$, which adds an adversarial loss so the output resembles the output domain. The adversarial loss for each GAN is:

$$L_{GAN}(G,D,A,B) = E_{y \sim p(B)} \left[ \log(D(y)) \right] + E_{x \sim p(A)} \left[ \log(1 - D(G(x))) \right], \quad (5)$$

$G$ has noise vectors in its output so an additional discriminator is required. Rather than distinguishing noise values, a discriminator is applied to recreated images since our concern lies with the imaging rather than the noise. The discriminator compares images produced by random noise vectors and vectors produced by $F$. This adversarial loss is shown in Figure 1d and is defined as:

$$L_{GAN}^{rec}(G,F,D,A) = E_{x \sim p(A)} \left[ \log(D(F(G_{im}(x), G_{cl}(x), G_{ts}(x)))) \right] + E_{x \sim p(A), z \sim p(Z)} \left[ \log(1 - D(F(G_{im}(x), z_{cl}, z_{ts}))) \right], \quad (6)$$

The adversarial portion of the objective loss is as follows:

$$L_{adv} = L_{GAN}(G,D_G,OC,VC) + L_{GAN}(F,D_F,VC,OC) + L_{GAN}^{rec}(G,F,D_{rec},OC) \quad (7)$$
Fig. 2. To understand how $z_{cl}$ and $z_{ts}$ affect the output, $z_{cl}$ and $z_{ts}$ are individually linearly interpolated. The top half shows interpolation between $z_{cl}$ values, while $z_{ts}$ is fixed. The colon-specific color and lighting gradually changes with $z_{cl}$. The bottom half shows $z_{cl}$ fixed, while $z_{ts}$ is interpolated. The specular reflection shapes and texture gradually change. The last row also shows fecal matter changing between images.

During training, $F$ may ignore $z_{ts}$. To encourage using noise input, $L_t$ is added to penalize the network when different noise inputs have similar results. The function penalizing the network is defined as:

$$L_{text}(I_1, I_2) = \begin{cases} \alpha - \|I_1 - I_2\|_1 & \text{if } \alpha > \|I_1 - I_2\|_1 \\ 0 & \text{else} \end{cases}$$

where $I$ is an image and $\alpha$ they differ. $F$ is applied to two different images, and the OC images are compared using $L_t$ as seen in Figure 1c and defined as:

$$L_t = \mathbb{E}_{x \sim p(VC), z \sim p(Z)} L_{text}(F(x, z_{cl}, z_{ts}^1), F(x, z_{cl}, z_{ts}^2))$$
Fig. 3. Showing the $z_{cl}$ vector being extracted from various reference images (top most) and applied to target images (left most) to transfer its colon-specific color and lighting.

Lastly, an identity loss is added for stability. An image should be unchanged if the input is from the output domain. It is only applied to $G$ to encourage texture and specular reflection generation. The identity loss is defined as:

$$L_{idt}(G, A) = E_{x \sim p(A)} \| x - G_{im}(x) \|_1$$

The identity portion of the objective loss is defined as $L_{idt} = L_{idt}(G, VC)$. The generators are ResNets [8] with 9 blocks that use 23MB. CLTS-GAN uses PatchGAN discriminators [11], each using 3MB. The network was trained for 200 epochs on an Nvidia RTX 6000 GPU with the following parameters: $\lambda_{adv} = 1$, $\lambda_T = 10$, $\lambda_{text} = 20$, $\lambda_{idt} = 1$, and $\alpha = .1$. Inference time is .04 seconds.

CLTS-GAN controls the output using $z_{ts}$ and $z_{cl}$. For VC, if two $z_{cl}$ values are selected with a fixed $z_{ts}$ they can be linearly interpolated and passed into $F$ creating gradual changes in the colon-specific color and lighting as seen in Figure 2. The strength of the specular reflections change with $z_{cl}$ since the lighting is being altered. Similarly, $z_{ts}$ can be linearly interpolated to provide gradual changes in texture and specular reflection as well as fecal matter. Here the shape of the specular reflections and texture fade in and out. Since changes in $z_{ts}$ and $z_{cl}$ do not lead to sporadic changes, they can be used in more meaningful ways.
Fig. 4. Depicting our model using various $z_{cl}$ and $z_{ts}$ values to generate realistic OC images. The left most image is the input image for CLTS-GAN followed by the output OC images. We show results on VC, VRCaps [10] data, and OfGAN [22] synthetic input. Additional results can be found in Figure 1 of the supplementary material.

Figure 3 shows the transfer of colon-specific color and lighting information from one OC image to another. $G$ extracts the $z_{cl}$ vector from the reference and the VC and $z_{ts}$ from the target. When these values are input to $F$ it transfers the color and lighting from the reference to the target. $z_{ts}$ remains fixed since it is intended for generating realistic textures and specular for VC instead of altering geometry dependant texture and specular of OC.

5 Results and Discussion

Figure 4 shows qualitative results for CLTS-GAN’s realistic OC generation using VC images and data from VRCaps [10] and OfGAN [22]. The input was passed to $F$ with $z_{ts}$ and $z_{cl}$ randomly sampled from a uniform distribution to show a large variety in colon-specific color, lighting, texture and specular reflection.
More results can be found in the supplementary material. $z_{ts}$ and $z_{cl}$ can be individually changed to control the texture and specular reflection separately from the color and lighting as shown in Figures 2 and 3 of the supplementary material.

To show quantitative evaluation of CLTS-GAN, PraNet [6], a state-of-the-art polyp segmentation model, is trained with and without augmentation. PraNet uses CVC Clinic DB [2] and HyperKvasir [4] for training. The images were augmented with colon-specific color and lighting, while polyp specific textures and speculars were preserved. Random $z_{cl}$ values are applied to training images by extracting the VC and $z_{ts}$ using $G$ and passing the three values to $F$. Examples are shown in Figure 5. PraNet was trained having each image augmented 0, 1, and 3 times. When there was no augmentation or one augmentation the network was trained for 20 epochs. To avoid overfitting on the shapes of the polyps, the network was trained for 10 epochs when augmented 3 times. Testing results are shown in Table 1. Data augmentation from CLTS-GAN improves the DICE, IoU, and MAE scores for various testing datasets. For the CVC-T dataset, using only one augmentation appeared to have marginal improvement over using 3.

In this work we present CLTS-GAN, a one-to-many image-to-image translation model for dataset augmentation and OC synthesis with control over color, lighting, texture, and specular reflections. $z_{ts}$ and $z_{cl}$ control these attributes, but can be further disentangled. High intensity specular reflections can be extracted with a loss and stored in a separate parameter. CLTS-GAN does not contain temporal components. Adding multiple frames as input can get the network to use the texture and specular information in a temporally consistent manner. Moreover, in the future, we will also explore the utility of CLTS-GAN augmentations in depth inference [15,17] and folds detection [16]. We hypothesize that the full gamut of color-lighting-texture-specular augmentations can be used in these scenarios to improve performance.
Table 1. PraNet results with and without dataset augmentation. Colon-specific color and lighting augmentation was applied to avoid altering polyp specific textures. Results for 1 and 3 additional images are shown in the second and third rows. Both show improvement over PraNet without augmentation. PraNet with 1 augmentation is better for CVC-T which indicates the network may have overfit on the shapes of polyps.

|               | CVC-Colon DB [3] | ETIS [20] | CVC-T [21] |
|---------------|------------------|-----------|------------|
| PraNet w/out Aug | Dice↑ 0.712 IoU↑ 0.640 MAE↓ 0.043 | Dice↑ 0.628 IoU↑ 0.567 MAE↓ 0.031 | Dice↑ 0.871 IoU↑ 0.797 MAE↓ 0.10 |
| PraNet w/ 1 Aug | Dice↑ 0.750 IoU↑ 0.671 MAE↓ 0.037 | Dice↑ 0.704 IoU↑ 0.626 MAE↓ 0.019 | 0.893 0.824 0.007 |
| PraNet w/ 3 Aug | **0.781** **0.697** **0.030** | **0.710** **0.639** **0.027** | 0.884 0.815 0.010 |

Acknowledgements

This project was supported by MSK Cancer Center Support Grant/Core Grant (P30 CA008748) and NSF grants CNS1650499, OAC1919752, and ICER1940302.

References

1. Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented cyclegan: Learning many-to-many mappings from unpaired data. arXiv preprint arXiv:1802.10151 (2018)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrell, G., Gil, D., Rodríguez, C., Vilarino, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics 43, 99–111 (2015)
3. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition 45(9), 3166–3182 (2012)
4. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., de Lange, T.: Hyper-kvasir: A comprehensive multi-class image and video dataset for gastrointestinal endoscopy (Dec 2019). https://doi.org/10.31219/osf.io/mkzcq
5. Chu, C., Zhmoginov, A., Sandler, M.: Cyclegan, a master of steganography. arXiv preprint arXiv:1712.02950 (2017)
6. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 263–273 (2020)
7. Fazlollahi, A.M., Bakhtiar, M., Alsayegh, A., Yilmaz, R., Winkler-Schwert, A., Mirchi, N., Langleben, I., Ledwos, N., Sabbagh, A.J., Bajunaid, K., et al.: Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: A randomized clinical trial. JAMA Network Open 5(2), e2149008-e2149008 (2022)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
9. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. Proceedings of the European Conference on Computer Vision (ECCV) pp. 172–189 (2018)
10. ˙Incetan, K., Celik, I.O., Obeid, A., Gokceler, G.I., Ozyoruk, K.B., Almalioglu, Y., Chen, R.J., Mahmood, F., Gilbert, H., Durr, N.J., Turan, M.: VR-Caps: A virtual environment for capsule endoscopy. Medical Image Analysis p. 101990 (2021)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 1125–1134 (2017)
12. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4401–4410 (2019)
13. Ma, R., Wang, R., Pizer, S., Rosenman, J., McGill, S.K., Frahm, J.M.: Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions. International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 573–582 (2019)
14. Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S.K., Rosenman, J., Frahm, J.M.: Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. Medical Image Analysis 72, 102100 (2021)
15. Mahmood, F., Chen, R., Durr, N.J.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Transactions on Medical Imaging 37(12), 2572–2581 (2018)
16. Mathew, S., Nadeem, S., Kaufman, A.: Foldit: Haustral folds detection and segmentation in colonoscopy videos. International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 221–230 (2021)
17. Mathew, S., Nadeem, S., Kumari, S., Kaufman, A.: Augmenting colonoscopy using extended and directional cyclegan for lossy image translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4696–4705 (2020)
18. Nadeem, S., Kaufman, A.: Computer-aided detection of polyps in optical colonoscopy images. SPIE Medical Imaging 9785, 978525 (2016)
19. Rivoir, D., Pfeiffer, M., Docea, R., Kolbinger, F., Riediger, C., Weitz, J., Speidel, S.: Long-term temporally consistent unpaired video translation from simulated surgical 3D data. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 3343–3353 (2021)
20. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery 9(2), 283–293 (2014)
21. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparra, G., López, A.M., Romero, A., Drozdal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of Healthcare Engineering 2017 (2017)
22. Xu, J., Anwar, S., Barnes, N., Grimpen, F., Salvado, O., Anderson, S., Armin, M.A.: Ofgan: Realistic rendition of synthetic colonoscopy videos. International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 732–741 (2020)
23. Zhang, Y., Wang, S., Ma, R., McGill, S.K., Rosenman, J.G., Pizer, S.M.: Lighting enhancement aids reconstruction of colonoscopic surfaces. International Conference on Information Processing in Medical Imaging pp. 559–570 (2021)
CLTS-GAN: Color-Lighting-Texture-Specular Reflection Augmentation for Colonoscopy (Supplement)

Shawn Mathew\textsuperscript{1*}, Saad Nadeem\textsuperscript{2*}, and Arie Kaufman\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, Stony Brook University
\textsuperscript{2}Department of Medical Physics, Memorial Sloan Kettering Cancer Center

\textbf{Fig. 1.} This figure depicts our model with a variety of texture and color output for a number of VC input images.

\textsuperscript{*}Equal contribution
Fig. 2. This figure depicts our model’s ability to generate images with different color and lighting. Each image in a row has a different colon-specific color and lighting vector as input. Notice how the specular reflections alter based on the color and lighting vector, but the locations of the reflections remain the same.

Fig. 3. This figure depicts our model’s ability to generate images with different texture and specular reflections using $z_{ts}$. The leftmost image is CLTS-GAN’s input followed by its output with a fixed $z_{cl}$ and various $z_{ts}$ vectors. Blue boxes are used to highlight some of the differences.