Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training

Faisal Mahmood, Richard Chen, and Nicholas J. Durr

***Abstract*— To realize the full potential of deep learning for medical imaging, large annotated datasets are required for training. Such datasets are difficult to acquire due to privacy issues, lack of experts available for annotation, underrepresentation of rare conditions, and poor standardization. The lack of annotated data has been addressed in conventional vision applications using synthetic images refined via unsupervised adversarial training to look like real images. However, this approach is difficult to extend to general medical imaging because of the complex and diverse set of features found in real human tissues. We propose a novel framework that uses a reverse flow, where adversarial training is used to make real medical images more like synthetic images, and clinically-relevant features are preserved via self-regularization. These domain-adapted synthetic-like images can then be accurately interpreted by networks trained on large datasets of synthetic medical images. We implement this approach on the notoriously difficult task of depth-estimation from monocular endoscopy which has a variety of applications in colonoscopy, robotic surgery, and invasive endoscopic procedures. We train a depth estimator on a large data set of synthetic images generated using an accurate forward model of an endoscope and an anatomically-realistic colon. Our analysis demonstrates that the structural similarity of endoscopy depth estimation in a real pig colon predicted from a network trained solely on synthetic data improved by 78.7% by using reverse domain adaptation.**

***Index Terms*— Convolutional neural networks, synthetic data, adversarial training, GANs, domain adaptation, medical imaging, endoscopy.**

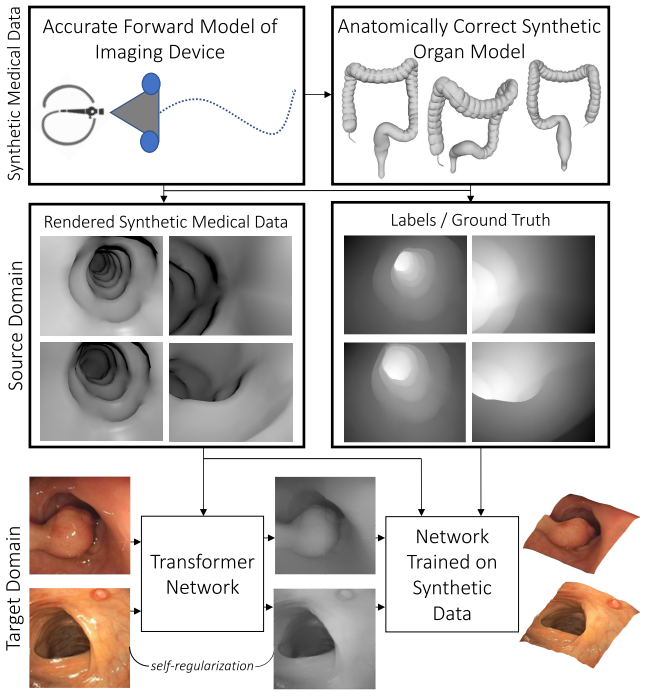

Fig. 1. Unsupervised reverse domain adaptation for endoscopy images. We use an accurate forward model of an endoscope and an anatomically correct colon model to generate synthetic endoscopy images with ground truth depth. This large synthetic dataset can be used to train a deep network for depth estimation. An adversarial network transforms input endoscopy images to a synthetic-like representation while preserving clinically relevant features via self-regularization. These synthetic-like images can be directly used for depth estimation from the network trained on synthetic images.

I. Introduction

DEEP Learning offers great promise for the reconstruction and interpretation of medical images [1]–[6]. Countless applications in clinical diagnostics, disease screening, interventional planning, and therapeutic surveillance rely on the subjective interpretation of medical images from health-care providers. This approach is costly, time-intensive, and has well-known accuracy and precision limitations—all of which could be mitigated by objective, automatic image analysis [7].

For conventional images, deep learning has achieved remarkable performance for a variety of computer vision tasks, typically by utilizing large sets of real-world images for training, such as ImageNet [8], COCO [9] and Pascal VOC [10]. Unfortunately, the potential benefits of deep learning have yet to transfer to the most critical needs in medical imaging because of the limited availability of annotated datasets.

Manuscript received February 27, 2018; revised May 17, 2018; accepted May 23, 2018. Date of publication June 1, 2018; date of current version November 29, 2018. This work was supported by Johns Hopkins University. *(Corresponding author: Faisal Mahmood.)*

F. Mahmood and N. J. Durr are with the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: faisalm@jhu.edu; ndurr@jhu.edu).

R. Chen is with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: rchen40@jhu.edu).

Images are best viewed in color on the electronic version of this document.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2018.2842767

Despite the compelling need for such datasets, there are practical concerns that impede their development, including the cost, time, expertise, privacy, and regulatory issues associated with medical data collection, annotation, and dissemination.

The obstacles associated with developing a large dataset of real images can be circumvented by generating synthetic images [11]–[13]. Considerable effort has been devoted to adapting models generated with synthetic data as the source domain to real data as the target domain [14]. Advances in adversarial training have sparked interest in making synthetic data look more realistic via unsupervised adversarial training (SimGAN) [15]. In the medical imaging domain, there has been recent success in generating realistic synthetic data for the relatively constrained problem of 2D retinal imaging using standard GANs [16], [17]. In more complex applications, it is challenging to generate an appropriate span of synthetic medical images for training, because few models exist that accurately simulate the anatomical complexity and diversity found in healthy to pathologic tissues. Moreover, the forward models for medical imaging devices are more complex than those used in many conventional vision applications. Consequently, models trained on synthetic medical data may fail to generalize to real medical images, where accurate interpretation may be critically important.

Cross-patient network adaptability is a well-known challenge to learning-based medical imaging methods. Often a network trained on data from one patient fails to generalize to other patients. This is commonly observed for optical imaging methods, such as endoscopy, which capture a broad span of both low- and high-frequency texture details of the patient. Some of these details are patient specific and do not hold any diagnostic information that can be generalized across patients. Training from features that are patient specific can limit the generalizability of a network. This complication makes it difficult for methods like SimGAN [15] to work both accurately and generally because the span of realistic images produced will be similar to the real images used for training.

In this work, we propose to reverse the flow of traditional adversarial training-based domain adaptation. Instead of changing synthetic images to appear realistic [15], we transform real images to look more synthetic (Fig. 1). *We train an adversarial transformation network that transforms real medical images to a synthetic-like representation while preserving clinically relevant information.* In summary, we can train solely on synthetic medical data as the source domain and transform real data in the target domain to a more synthetic interpretation, thus bridging the gap between the source and target domains in a reverse manner. Visually it appears that this gap can also be reduced by using simple filtering and noise removal methods. However, we observed that such methods can also remove or smooth out intensity information which is a major cue for many prediction tasks.

To transform real images to a synthetic-like representation, we train a transformer with an adversarial loss similar to GANs [18] and SimGAN [15]. However, unlike SimGAN that trains for inducing realism to synthetic data, we train for a synthetic-like representation of real data. With the roles of synthetic and real data reversed, the overall transformer architecture is similar to a standard GAN and is composed of a transformer network that tries to fool a discriminator network into thinking that the transformed medical image is synthetic. In addition to removing patient specific details from the data, the synthetic image should preserve enough information within the data that it could be used for the task at hand. To preserve this information a fully connected network is used and the adversarial loss is complemented with a self-regularization term which constrains the amount of deviation from the real image.

Contributions: In this work, we propose an adversarial training-based reverse domain adaptation method which uses unlabeled synthetic data to transform real data to a synthetic-like representation while maintaining clinically relevant diagnostic features via self-regularization.

- **Synthetic Medical Images with Ground Truth:** We generate a large dataset of perfectly-annotated synthetic endoscopy images from an endoscope forward model and an anatomically correct colon model.
- **Reverse Domain Adaptation:** We train a transformer network via adversarial training composed of a generator which generates a synthetic-like representation of real endoscopy images. The loss function of the generator contains a discriminator to classify the endoscopy images as real or synthetic and a self-regularization term that penalizes large deviations from the real image.
- **Qualitative and Quantitative Study:** We validate our domain adaptation approach by using synthetically generated endoscopy data to train a monocular endoscopy depth estimation network and quantitatively testing it with real endoscopy data from: a) Colon Phantom b) Real Pig Colon and qualitatively testing it with real human endoscopy data.

II. Related Work

A. Navigating Limited Medical Imaging Data

Improving the performance of deep learning methods with limited data is an active research area. Standard data augmentation has been used for medical imaging for the past few years [1], [2], [19], [20]. Ronneberger *et al.* [21] demonstrated success with using elastic augmentation with U-Net architectures for medical image segmentation. Payer *et al.* [22] have demonstrated that incorporating application-specific *a priori* information can train better deep networks. There is a growing interest in transferring knowledge from networks trained for conventional vision to the medical imaging domain [7], [23]–[26].

B. Generative Adversarial Networks

The GAN framework was first presented by Goodfellow *et al.* [18] and was based on the idea of training two networks, a generator and a discriminator simultaneously with competing losses. While the generator learns to generate realistic data from a random vector, the discriminator classifies the generated image as real or fake and gives feedback to the generator. Once the training reaches equilibrium the generator

is able to fool the discriminator every time it generates a new image. Initially GANs were applied to the MINST dataset [18] but recently the framework has been refined and used for a variety of applications [27]. Models with adversarial losses have been used for synthesis of 3D shapes, image-to-image translation, generating radiation patterns etc. Recently, Zhu *et al.* [23] proposed iGAN which enables interactive image manipulation on a natural image manifold. Shrivistava *et al.* [15] have proposed an unsupervised method for refining synthetic images to look more realistic using a modified adversarial training framework.

C. Adversarial Training for Biomedical Imaging

Adversarial training has been applied to a variety of medical imaging tasks. GANs have been used for noise reduction in low-dose CT in [28] and for MR to CT prediction in [29]. Osokin *et al.* [30] use GANs for synthesizing biological cells imaged by fluorescence microscopy. Costa *et al.* [17] and Guibas *et al.* [16] synthesize retinal images using adversarial training. BenTaieb *et al.* propose using GANs for stain transfer in histopathology image analysis. GANs have also been used for segmentation [31], [32], detection [33], reconstruction [34], [35] and classification [36].

All current adversarial training-based image synthesis methods attempt to generate realistic images from a random noise vector or refine synthetic images to create more realistic images. Our method, by contrast, transforms real images to a synthetic-like representation allowing the desired network to be trained only on synthetic images.

D. Endoscopy Depth Estimation

We implemented reverse domain adaptation on the problem of depth estimation in endoscopy. Recent advances in machine learning have led to significant progress on detection [4], [37], [38], segmentation [39] and classification [40] of lesions in endoscopy images. Depth information has been shown to improve performance of these tasks [41], but depth measurement remains challenging through an endoscope because of unpredictable movement, small working distances, limited endoscope size, non-uniform tissue texture, and the deformable nature of the tissue being imaged. Monocular depth is also a common goal for a variety of applications in robotic and laparoscopic surgery [42]–[45]. An elastic video interpolation approach for 3D reconstruction of the digestive wall was presented in [46]. Other approaches for depth estimation have focused on geometric modeling of the organ being imaged [47], have trained on shallow networks from a single patient [48] or use data that does not follow the inverse square law of intensity fall-off [49]. Photometric stereo endoscopy captures the 3D topography of the mucosa [50] but is inherently qualitative due to the unknown working distances from each object point to the endoscope. Learning-based approaches to depth estimation have been restricted by the lack of endoscopy images with ground truth depth [48], [51]. Moreover, optical endoscopy data has patient specific details, such as vascular patterns, that do not generalize across patients. In contrast to previous approaches, our reverse domain adaptation pipeline enables the use of models trained entirely on synthetic endoscopy data with no patient specific details but a rich variation of diagnostic representations and endoscope parameters. This is accomplished with adversarial training to remove patient specific details from real test images while maintaining diagnostic details, enabling a transformation to a synthetic-like representation.

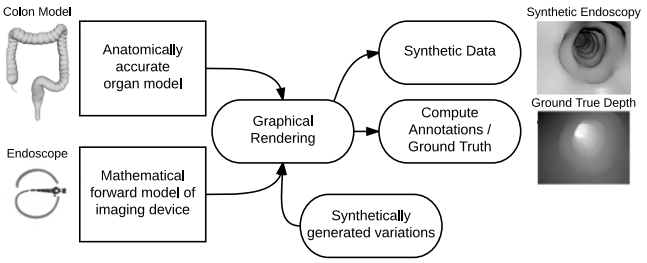

Fig. 2. General framework for generating synthetic medical imaging data with endoscopy as an example.

III. Generating Synthetic Medical Data

Despite the widespread use of synthetic data for training deep networks for real world images [52]–[55], its use for medical imaging applications has been relatively limited. Recent work on generating synthetic retinal [17] and histopathology [56] images has demonstrated promise. Unlike conventional real-world images that may contain a constrained span of object diversity, medical images capture information of biological tissues which contain unique patient-specific texture that is difficult to model. We therefore propose a framework where we generate a large dataset of medical images with this patient-specific detail removed so that a network can be trained on universal diagnostic features. In general, this synthetic data can be generated by (Fig. 2):

1) Developing an accurate forward model for the medical imaging device.
2) Generating an anatomically accurate model of the organ being imaged.
3) Rendering images from a variety of positions, angles and parameters.

Typically, forward models for medical imaging devices are more complicated as compared to typical cameras, and anatomically accurate models need to represent a high degree of variation and rare conditions to cater for a diverse set of patients [57].

Synthetic Endoscopy Data With Ground Truth Depth: For the purpose of demonstration of our proposed methods we focus on the task of depth estimation from monocular endoscopy images. This is a notoriously difficult problem because of the lack of clinical images with available ground truth data, since it is difficult to include a depth sensor on an endoscope. We generate synthetic data to overcome this issue. We develop a forward model of an endoscope with a wide-angle monocular camera and two to three light sources that exhibit realistic inverse square law intensity of fall-off.

We use a synthetically generated and anatomically accurate colon model and image it using the virtual endoscope placed at a variety of angles and varying conditions to mimic the movement of an actual endoscope. We also generate pixel-wise ground truth depth for each rendered image. We finally create a dataset with 200, 000 images with ground truth depth. Although this large dataset of images is able to train efficient deep networks these networks are not effectively generalizable to real world images.

IV. Proposed: Reverse Domain Adaptation

A. Transformer Loss

Formally, the goal of our proposed reverse domain adaptation method is to use a set of synthetic images $\boldsymbol{g}_i \in \mathcal{G}$ to learn a transformer $\mathcal{T}_{\gamma_t}(\boldsymbol{x})$ that can transform real images $\boldsymbol{x}$ to a synthetic-like representation $\boldsymbol{x}''$. The transformer should be able to fool a discriminator $\mathcal{D}_{\gamma_d}$ where γ_t and γ_d are the learning parameters. There are three key requirements for this setup:

- The transformer output should only remove the patient specific details in the image, while preserving diagnostic features.
- The adversarial training should not introduce artifacts in the transformed image.
- The adversarial training should be stable and should not suffer from mode collapse.

The transformer loss function can be defined as,

$$\mathcal{L}_{\mathcal{T}}(\gamma_t) = \sum_i \psi(\boldsymbol{x}_i, \mathcal{G}; \gamma_t) + \lambda\phi(\boldsymbol{x}_i; \gamma_t), \tag{1}$$

where, $\boldsymbol{x}_i$ is the i^{th} real image. The first term of the loss function, ψ transforms the real image to a synthetic-like representation and the second term, ϕ penalizes large variations to preserve specific properties of the real image. λ controls the amount of self-regularization enforced by ϕ.

B. Discriminator Loss

In order to transform a real image to its synthetic-like counterpart, the gap between the representations of the real and synthetic image needs to be minimized. An ideal transformer should be able to produce an indistinguishable synthetic representation of a real image every time, which is possible if a discriminator is embedded within the transformer's loss function (Fig. 3). As explained in [15], [18], and [27], a discriminator is a classifier that classifies the output of another network as real or fake. However, unlike [15], in our case the role of the discriminator is reversed—instead of enforcing the transformer to produce more realistic images the role of the discriminator is to enforce the transformer to produce synthetic images. The discriminator loss can be defined as follows,

$$\mathcal{L}_{\mathcal{D}}(\gamma_d) = -\sum_i \log(\mathcal{D}_{\gamma_d}(x_i'')) - \sum_j \log(1 - \mathcal{D}_{\gamma_d}(g_j)). \tag{2}$$

This is essentially a two class classification problem with cross-entropy error where the first term represents the probability of the input being a synthetic image and the second term represents the probability of the input image being synthetic-like representation of a real image. The discriminator works on a patch level rather than the entire image to prevent artifacts.

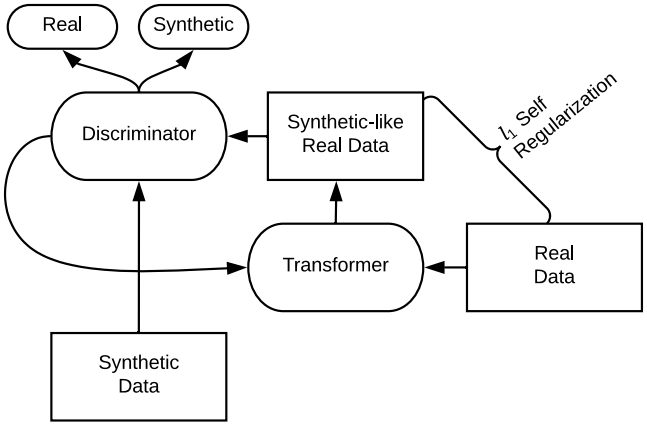

Fig. 3. A top-level flow of our proposed adversarial training setup. Real data is transformed into a synthetic-like representation using a transformer network that minimizes an adversarial loss term and a self-regularization term. The discriminator acts as a classifier to identify the image as real or synthetic and gives feedback to the transformer via the adversarial loss. The total loss essentially defines how well the discriminator is tricked into believing that the transformed real image is synthetic and how close the transformed image is to the real image.

To train our network, we randomly sample mini-batches of synthetic images and images transformed to a synthetic-like representation by the transformer. Instead of using individual outputs of the transformer we use randomly sampled, buffered outputs and a set of randomly sampled synthetic images. This increases the stability of the adversarial training since the lack of memory can diverge the adversarial training and introduce artifacts [15]. At each step the discriminator trains using this mini-batch and parameters γ_d are updated using stochastic gradient decent (SGD). The transformer loss is then updated with the trained discriminator, the ψ term in Eq. 1 can be defined as,

$$\psi(\boldsymbol{x}_i, \mathcal{G}; \gamma_t) = -\log(1 - \mathcal{D}_{\gamma_d}(\mathcal{T}_{\gamma_t}(\boldsymbol{x}))). \tag{3}$$

As the training shuffling between the transformer and discriminator reaches equilibrium the transformer is able to fool the discriminator every time. The loss in Eq. 3 forces the discriminator to fail to classify transformed images as synthetic-like real.

C. Self-Regularization

As mentioned earlier, a key requirement for the transformer is that it should only remove patient specific data and should preserve diagnostic features. For the proof-of-concept proposed in this work, we utilize a simple, per-pixel loss term between the real image and the synthetic-like real representation of the image to penalize the transformed image from deviating significantly from the real image. The self regularization term ϕ can be defined as,

$$\phi(\boldsymbol{x}_i; \gamma_t) = || \Phi(\mathcal{T}_{\gamma_t}(\boldsymbol{x})) - \Phi(\boldsymbol{x}) ||_1, \tag{4}$$

where Φ represents the feature transform and $|| \cdot ||_1$ represents the ℓ_1 norm.

Algorithm 1 Adversarial training of a Transformer $x'' = \mathcal{T}_\gamma(x)$

INPUT: Synthetic Data: $\boldsymbol{g}_i \in \mathcal{G}$, Real Data: $\boldsymbol{x}_i \in \mathcal{X}$, Transformer Updates/step: n_t, Discriminator Updates/step: n_d, Maximum number of steps: S

for $s = 1, 2, 3 \ldots S$ **do**
  for $n_t = 1, 2, 3 \ldots N_t$ **do**
    Sample a mini-batch $\{x_1, x_2, \ldots x_k\}$ of k real images.
    Update the transformer network parameters γ_t by taking an SGD step:
  $\nabla_{\gamma_t} \frac{1}{k} \sum_i \psi(\boldsymbol{x}_i, \mathcal{G}; \gamma_t) + \lambda \phi(\boldsymbol{x}_i; \gamma_t)$
  end for
  for $n_d = 1, 2, 3 \ldots N_d$ **do**
    Sample a mini-batchs of k synthetic images $\{g_1, g_2, \ldots g_k\}$ and transformed real images $\{x_1, x_2, \ldots x_k\}$.
    $\boldsymbol{x}_i'' \leftarrow \mathcal{T}_\gamma(\boldsymbol{x}_i)$
    Update the discriminator network parameters γ_d by taking an SGD step:
  $-\nabla_{\gamma_d} \frac{1}{k} \sum_i \log(\mathcal{D}_{\gamma_d}(\boldsymbol{x}'')) - \sum_j \log(1 - \mathcal{D}_{\gamma_d}(g_j))$
  end for
end for

OUTPUT: Trained Transformer Model $\mathcal{T}_\gamma(x)$

The overall transformer loss term can be rewritten as,

$$\mathcal{L}_\mathcal{T}(\gamma_t) = -\sum_i \log(1 - \mathcal{D}_{\gamma_d}(\mathcal{T}_{\gamma_t}(\boldsymbol{x_i}))) + \lambda \mid\mid \Phi(\mathcal{T}_{\gamma_t}(\boldsymbol{x_i})) - \Phi(\boldsymbol{x_i}) \mid\mid_1, \quad (5)$$

In summary, the total loss measures how well the discriminator is tricked into believing that the transformed real image is synthetic, and how close the transformed image is to the real image. If the discriminator is trained for the entire image it can introduce artifacts because it can over-emphasize on features from the current discriminator. In order to prevent these artifacts, rather than using a global discriminator we train the discriminator patch-wise. This overall training process has been explained in detail in Algorithm 1 and the network architectures are described in the implementation section and in the supplement.

V. Depth Estimation From Monocular Endoscopy Images

The previous sections have talked about generating synthetic medical data and adversarial training to bring real images within the domain of the synthetic data via a reverse domain adaptation pipeline. In order to evaluate the effectiveness of our proposed reverse domain adaptation pipeline, we train a network from synthetically generated endoscopy data (Fig. 2) and demonstrate that it can be adapted to three different target domains. By demonstrating that the distribution of the target domain can be brought closer to the source domain via adversarial training, we show that our depth estimation paradigm is domain independent. Once the synthetic data with ground truth depths is generated, we use the CNN-CRF based joint training framework described in [57] and [58].

Assuming $\boldsymbol{g} \in \mathbb{R}^{n \times m}$ is a synthetic image which has been segmented into p super-pixels and $\boldsymbol{y} = [y_1, y_2, \ldots, y_p] \in \mathbb{R}$ is the ground truth depth vector for each super-pixel and $\boldsymbol{h}$ represented the predicted depth vector. For this case, the conditional probability distribution of the synthetic data can be defined as,

$$Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{exp(E(\boldsymbol{y}, \boldsymbol{x}))}{\int_{-\infty}^{\infty} exp(E(\boldsymbol{y}, \boldsymbol{x}))d\boldsymbol{y}}. \quad (6)$$

where, E is the energy function. To estimate the depth of an incoming image we need to solve, $\widehat{\boldsymbol{y}} = \text{argmax}_y \, Pr(\boldsymbol{y}|\boldsymbol{x})$.

Let ξ and η be unary and pairwise potentials over nodes $\mathcal{N}$ and edges $\mathcal{S}$ of $\boldsymbol{x}$, then the energy function can be written as,

$$E(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i \in \mathcal{N}} \xi(y_i, \boldsymbol{x}; \boldsymbol{\theta}) + \sum_{(i,j) \in \mathcal{S}} \eta(y_i, y_j, \boldsymbol{x}; \boldsymbol{\beta}), \quad (7)$$

where, the unary potential estimates the depth from a single superpixel and the pairwise part smartly smooths the depths between superpixels. The goal is to learn the unary and pairwise part in a unified joint deep neural network framework. The unary part takes a single image superpixel patch as an input and feeds it to a fully convolutional CNN which outputs a depth of that superpixel. The unary potential can be defined as,

$$\xi(y_i, \boldsymbol{x}; \boldsymbol{\theta}) = -(y_i - h_i(\boldsymbol{\theta}))^2 \quad (8)$$

where h_i is the regressed depth of superpixel and θ represents network parameters [57], [58].

The pairwise function smooths neighboring superpixels learned from similarity metrics. $\boldsymbol{\beta}$ are the pairwise network parameters and $\boldsymbol{S}$ is the similarity matrix where $S_{i,j}^k$ is a similarity metric between the i^{th} and j^{th} superpixel. We use intensity difference and grayscale histogram as pairwise similarities. The pairwise potential can then be defined as,

$$\eta(y_i, y_j; \boldsymbol{\beta}) = -\frac{1}{2} \sum_{k=1}^{K} \beta_k S_{i,j}^k (y_i - y_j)^2. \quad (9)$$

The overall energy function can be written as,

$$E = -\sum_{i \in \mathcal{N}} (y_i - h_i(\boldsymbol{\theta}))^2 - \frac{1}{2} \sum_{(i,j) \in \mathcal{S}} \sum_{k=1}^{K} \beta_k S_{i,j}^k (y_i - y_j)^2. \quad (10)$$

For training the negative log likelihood of the probability density function is minimized with respect to the two learning parameters. Two weight decay terms are added to the objective function to penalize heavily weighted vectors $(\lambda_\theta, \lambda_\beta)$. Assuming N is the number of images in the training data,

$$\min_{\theta, \beta \geq 0} -\sum_{1}^{N} \log Pr(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\beta}) + \frac{\lambda_\theta}{2} \|\theta\|_2^2 + \frac{\lambda_\beta}{2} \|\beta\|_2^2. \quad (11)$$

The optimization problem is solved using stochastic gradient decent-based back propagation. More details about this paradigm are given in the supplement with this paper.

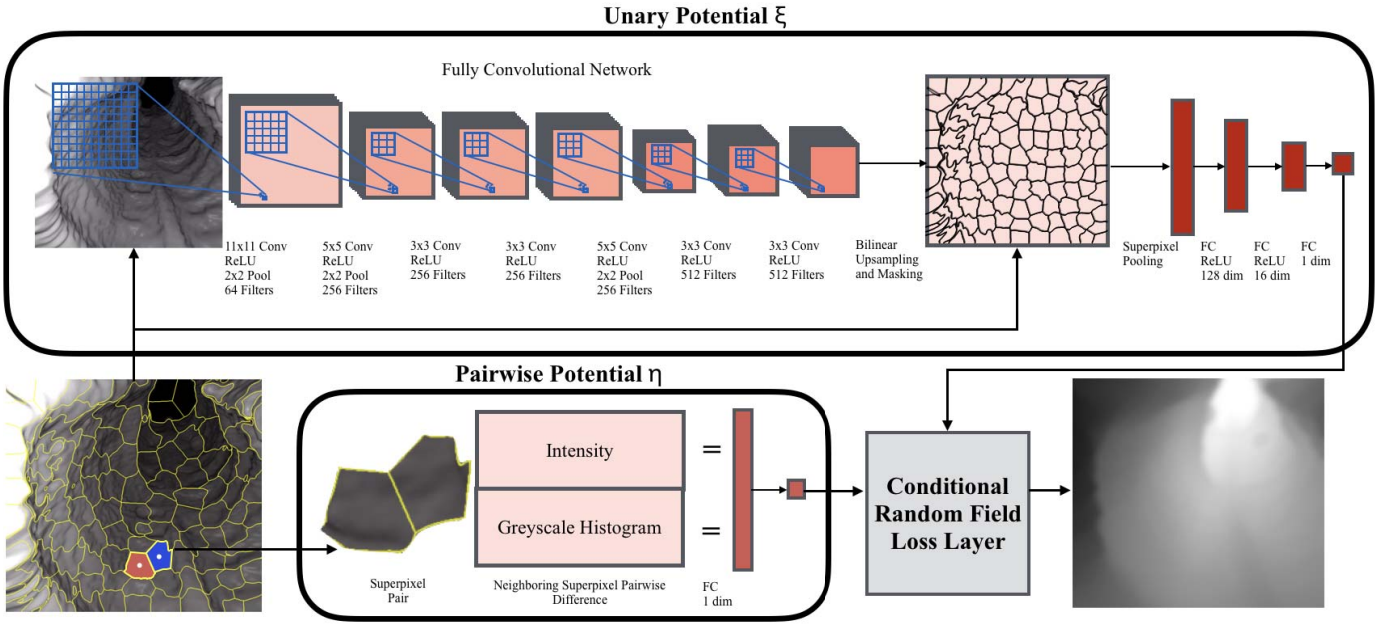

Fig. 4. The transformed synthetic-like real image is fed into a CNN-CRF jointly trained network. The unary part of the CNN-CRF setup is composed of a fully convolution network (FCN) that produces convolution maps which are retraced back to superpixels in a pooling layer which gives feature for each superpixel, followed by 3 fully connected layers. In the pairwise part, similarities of neighboring superpixels are calculated and fed into a fully connected layer.

VI. Experiments

A. Evaluation Datasets

We use three kinds of datasets in our quantitative and qualitative study of the proposed methods. Since there are no publicly available endoscopy datasets with ground true depth, we generate two kinds of datasets for quantitative evaluation:

1) Virtual endoscopy images with ground truth depth from a colon phantom modeled from a human cadaver.
2) Optical endoscopy images of a real pig colon with ground truth depth from registered CT.

We also use publicly-available human colonoscopy images to qualitatively assess if intuitive depth maps can be generated from real endoscopy videos.

1) Colon Phantom Data: The colon phantom data is generated from a CT-reconstructed model of a colon phantom molded from a real colon.[1] A virtual endoscope is used to render images from a variety of endoscopy images with corresponding ground truth from the CT-reconstructed model. 2,160 images are generated via this procedure and are used for evaluation (Fig. 5).

2) Real Pig Colon Data: Real optical endoscopy images were recorded from a pig colon fixed to a scaffold (Fig. 5). A 3D model of the scaffold was then reconstructed from CT measurements, and ground truth depth was generated for each real endoscopy image by registering virtual endoscopy views from the CT and optical endoscopy views from an endoscope (Fig. 5, bottom). 1,400 images with corresponding ground truth depth are generated using this procedure and are used for evaluation.

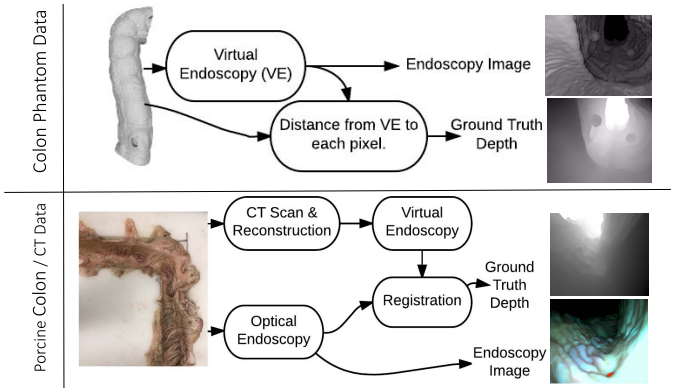

Fig. 5. The image collection and generation pipeline for colon phantom data and pig colon data. The colon phantom data is collected from a 3D rendered colon phantom using a virtual endoscope. The ground truth depth is calculated using the 3D model. The pig colon data is collected by imaging a porcine colon mounted on a scaffold using an optical endoscope and reconstructing a 3D model of the colon from CT measurements. The optical endoscopy and CT views are then registered to get the ground truth depth maps.

3) Real Endoscopy Data: We also evaluate our networks on publicly available endoscopy data[2] [59], [60]. However, these datasets do not have ground true depth and can only be used for qualitative evaluations.

B. Depth Estimation Network Trained on Synthetic Images

Implementation Details: The architecture used for training an endoscopy depth estimation network includes training the unary and a pairwise parts of a CRF in a unified framework

[1] https://www.thecgroup.com/product/colonoscopy-trainer-2003/

[2] https://polyp.grand-challenge.org/databases/

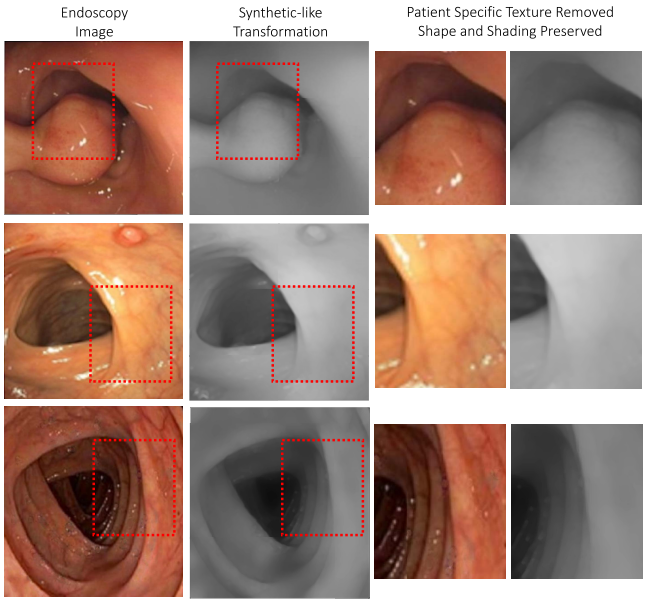

Fig. 6. Examples of real endoscopy images transformed to their synthetic-like representations. Patient-specific texture is clearly removed during the transformation.

presented in [57] and [58]. The unary part is composed of a fully convolutional network which generates convolution maps that are fed into a superpixel pooling layer followed by three fully connected layers. The pairwise part operates on a superpixel level and is composed of a single fully connected layer (Fig. 4). This setup was implemented using VLFeat MatConvNet[3] using MATLAB 2017a and CUDA 8.0. The training data was prepared by segmenting each virtual endoscopy image into 800 superpixels and corresponding ground truth depth were assigned to each superpixel. Segmenting the image into superpixels and estimating the depth of each superpixels reduces the overall resolution of the depth estimate but significantly improves training efficiency. Synthetic endoscopy data and its corresponding ground truth depth was generated according to the synthetic data generation pipeline presented in Section 3. The sequence of the generated data was randomized to prevent the network from learning too many similar features. 55% of the data was used for training, 40% for validation and 5% for testing. Training was done using K80 GPUs. Momentum was set at 0.9 as suggested in [58] and both weight decay parameters in Eq. 11 ($\lambda_\theta, \lambda_\beta$) were set to 0.0007. The learning rate was initialized at 0.00001 and decrease by 20% every 20 epochs. These parameters were tuned to achieve best results. A total of 300 epochs were run and the epochs with least $\log_{10}$ error were selected to avoid the selection of an over-fitted model. During training the synthetic data was sampled in randomized batches to prevent over-fitting. However, our experimentation demonstrated that repeating the training process produces only marginally different results.

C. Adversarial Training for Reverse Domain Adaptation

Implementation Details: Since the depth estimation network was trained solely on synthetic data all test images need to have a synthetic-like representation for the depth estimation to perform effectively. A transformer network was trained using the reverse domain adaptation paradigm presented in Section IV. The transformer and discriminator networks were implemented using tensorflow. The synthetic and real endoscopy images were down-sampled to a pixel size of 244 × 244 for computational efficiency. The real images were also converted to grayscale. The training between the transformer and the discriminator proceeds alternatively.

The transformer network was a standard residual network (ResNet) [61]. This is similar to [15], but for refining real data to be synthetic rather than the other way around. An input image of size 244 × 244 is convolved with a filter of 7 × 7 that outputs 64 feature maps that are then passed to 10 ResNet blocks followed by a 1 × 1 convolution layer resulting in one feature map. The transformer is trained with the discriminator for 200 steps and the self-regularization term for 800 steps to prevent it from deviating significantly from the actual image. The discriminator network is a standard classifier with five convolution layers, a max-pooling layer and softmax. More details about this architecture have been given in the supplement with this paper (Appendix B).

[3]http://www.vlfeat.org/matconvnet/

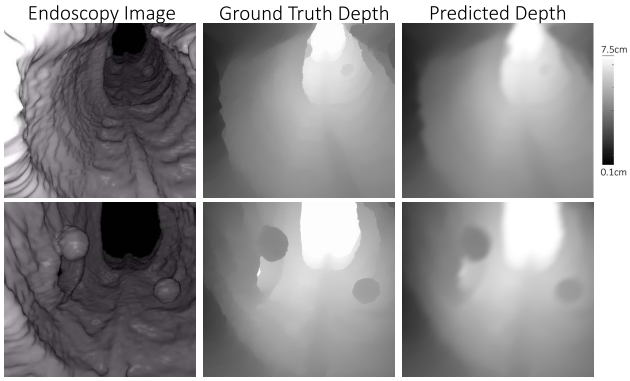

Fig. 7. Examples of rendered images, corresponding ground truth depth, and depth estimates from a colon phantom.

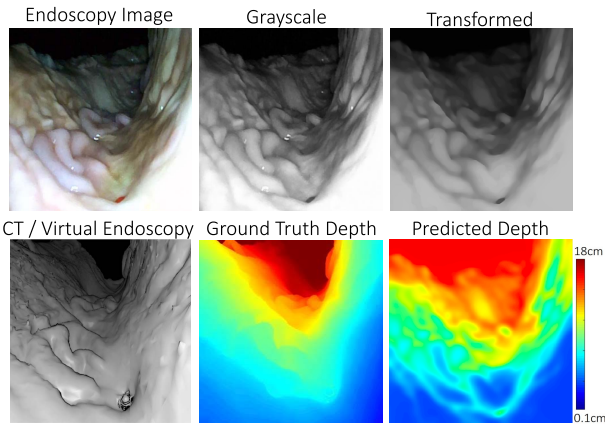

Fig. 8. Depth estimates from porcine colon data. The optical endoscopy image is converted to grayscale and transformed to its synthetic-like representation using our transformer network. The optical endoscopy view is registered to it's corresponding CT view to obtain ground truth depth.

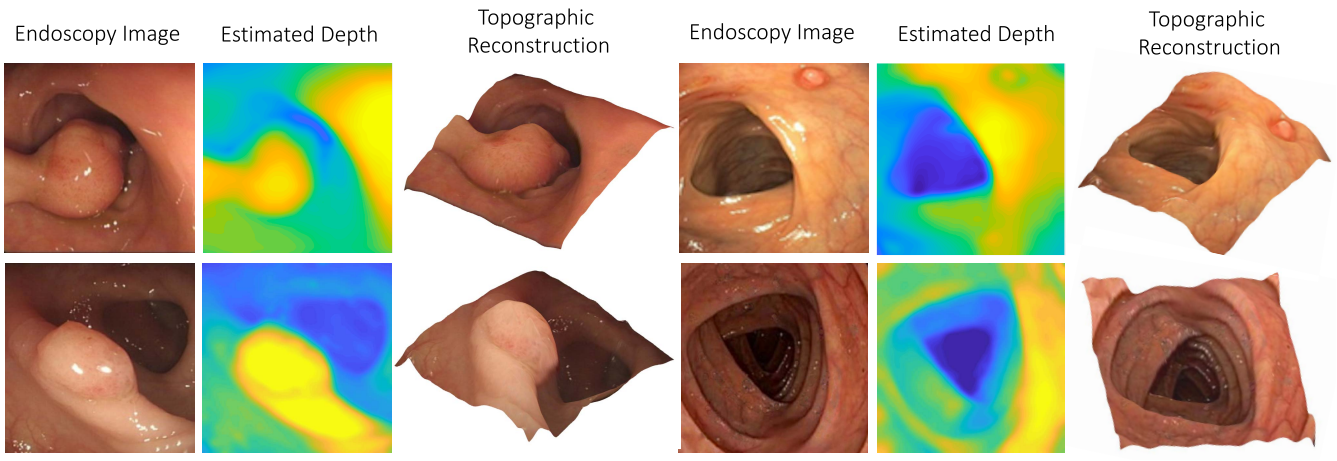

Fig. 9. Depth estimates and topographical reconstructions from monocular endoscopy images. Each endoscopy image is transformed to its synthetic-like representation as shown in Fig. 6 and is fed into our depth estimation network. The depth is then used to reconstruct the surface topography.

D. Results

1) Synthetic-Like Transformer: Fig. 6 shows examples of real endoscopy images transformed to their synthetic-like representations. It can be seen that the patient specific information has been removed and clinically-relevant features have been preserved for depth estimation and polyp identification. A close-up of the images show that the vasculature has been removed while preserving the shape information. In the next subsections we demonstrate that the depth estimation network trained on synthetic data performs significantly better with images transformed to their synthetic-like representations.

2) Depth Evaluation Metrics: We compare our depth estimates to corresponding ground truth values based on three metrics which have been used in previous endoscopy depth estimation work [49]:

- **Normalized root mean square deviation (NRMSD):** NRMSD $= \frac{\sqrt{\frac{\sum_i (h_i - y_i)^2}{n}}}{|h_{max} - h_{min}|}$, is a normalized version of the root mean square error and facilitates comparison across datasets and models with different scales. The lower the RMSD the closer the to the ground truth.
- **Pompeiu–Hausdorff distance (HD):** This metric calculates the greatest of all the distances from a point in the ground truth to the closest point in the prediction [62]. It can be calculated as $P(x, y) = \max(\boldsymbol{p}(h, y), \boldsymbol{p}(y, h))$ where $\boldsymbol{p} = \max_a \min_b \|a - b\|$ and $a \in h$ and $b \in y$. A lower HD indicates the two datasets being compared are more similar.
- **Structural Similarly Index (SSIM):** The SSIM is an assessment index calculated on the basis of contrast, luminance and structure. The SSIM in this paper is calculated according to, $SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$. Where μ and σ represent the mean and variance, σ_{xy} represents the covariance between x and y, c_1 and c_2 prevent division with a very small divisions. A window size of 8×8 was used when calculating the SSIM. This index is a decimal value between -1 and 1 with 1 indicating identical images. SSIM was used because it has shown to outperform other image comparison metrics [63].

TABLE I
A COMPARISON BETWEEN DEPTH ESTIMATED FROM RAW IMAGES AND DOMAIN ADAPTED IMAGES VIA OUR TRANSFORMER NETWORK

Test Dataset	NRMSD ↓	HD ↓	SSIM ↑
Colon Phantom	0.38	0.36	0.52
Trans. Colon Phantom (**Proposed**)	**0.23**	**0.23**	**0.77**
Real Pig Colon	0.61	0.58	0.33
Trans. Real Pig Colon (**Proposed**)	**0.32**	**0.30**	**0.59**

TABLE II
RESULTS OF OUR METHOD AS COMPARED TO THE STATE-OF-THE-ART ENDOSCOPY DEPTH ESTIMATION METHOD

Method	NRMSD ↓	HD ↓	SSIM ↑
Dictionary Learning (No Texture)[49]	0.57	0.56	0.35
Dictionary Learning (Texture)[49]	0.43	0.43	0.30
Ours (No Texture)	0.19	0.18	0.81
Ours (Pig Colon)	0.32	0.30	0.59
Ours (Colon Phantom)	0.23	0.23	0.77

3) Quantitative Results: Table I compares depth estimation results from colon phantom and real porcine colon data with and without domain transformation. It can clearly be seen that depth estimation is improved by domain transformation. As expected, the improvement in depth estimation that domain transformation provides is marginal in the colon phantom data, which has homogenous material properties, and more significant in real porcine tissue, which has natural biological variation in mucosal texture. There is a **78.78% improvement** in the SSIM for the porcine colon data and a **48.07% improvement** for the colon phantom data by transforming the input data using our proposed paradigm. Fig. 7 and 8 show representative depth estimation results for the colon phantom and porcine colon, respectively.

4) Comparative Analysis: We compared our depth estimation with reverse domain adaptation paradigm with existing results from [49] which uses dictionary learning (DiL) and trains with CT colonoscopy data. Table II shows a comparative analysis of their results compared to those from our approach. We demonstrate that our depth estimation is significantly better than their method.

5) Qualitative Results: For the purposes of demonstration, we also show that it is possible to estimate depth

from real human endoscopy data. Fig. 9 shows monocular endoscopy images, their estimated depth, and corresponding topographic reconstructions. Topographical reconstructions are reconstructed by overlaying depth on a 3D manifold. These depth estimates are qualitative and there is no corresponding ground truth depth available.

VII. Conclusions and Discussion

As the emerging tools for deep learning are increasingly applied to challenges in medical imaging, there is a growing need for large, annotated datasets that can fuel the development of these algorithms. GANs have demonstrated the potential to synthetically generate relevant datasets that are otherwise difficult to acquire and/or annotate, and are already improving the diagnostic performance of algorithms in a variety of disciplines, including radiology, histopathology, ophthalmology, and endoscopy. In general, however, few models exist that can accurately simulate the anatomical complexity and diversity found in healthy and pathologic tissues. Consequently, synthetic data driven by these models can fail to generalize to the real dataset.

In this paper, we propose a new approach to the problem of limited availability of annotated medical images. We introduced a novel unsupervised reverse domain adaptation paradigm that entails transforming real images to a synthetic-like domain, where a network trained on a large dataset of synthetically-generated data can be applied. We demonstrate that reverse domain adaptation using adversarial training with self-regularization can maintain diagnostic information (depth and tissue topography) in the transformed images. At the same time, our domain adaptation removes patient specific details from real test images so they can be used with a network trained entirely on synthetic data. This approach also addresses the issue of cross-patient network adaptability, where a network trained on one patient fails to generalize to other patients because it learns from patient specific texture or color. Removing patient specific details has no effect on diagnostic information because physicians do not look for unique patient data (like a fingerprint), but rather look for features that are known to map to a healthy or pathological domain. Our experimentation validates that if healthy and pathological features can be modeled in the synthetic images, the discriminator is not likely to be fooled if clinically-relevant features are removed by the transformer. Using real endoscopy images from a pig colon, we quantitatively validate that the relevant details needed to predict depth are preserved after transforming to a synthetic-like domain using self-regularization. We demonstrated a 78.78% and 48.07% improvement in predicting depth from synthetic-like domain-adapted images over raw images from both a real pig colon and a colon phantom respectively.

Our depth estimation paradigm is directly generalizable to gastroscopy, bronchoscopy and other endoscopic modalities. It is also generalizable to any medical imaging modality with an accurate forward model of the imaging device for synthetic data generation. This is true for several modalities such as CT, PET, MRI, and OCT. Future work will focus on using the proposed reverse domain adaptation paradigm for other medical imaging modalities and exploring the feature transform in the self-regularization term to handle more complicated cases. We will also incorporate the depth predicted by this approach into algorithms for automated polyp localization and classification.

Acknowledgment

The authors would like to thank Adam J. Sierakowski (Maryland Advanced Computing Cluster) for HPC training, and J. Webster Stayman and Steven Tilley (Johns Hopkins University) for collecting CT data.

References

[1] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.
[2] H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, Mar. 2016.
[3] J.-Y. He, X. Wu, Y.-G. Jiang, Q. Peng, and R. Jain, "Hookworm detection in wireless capsule endoscopy images with deep learning," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2379–2392, May 2018.
[4] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.
[5] Y. Yuan and M. Q.-H. Meng, "Deep learning for polyp recognition in wireless capsule endoscopy images," *Med. Phys.*, vol. 44, no. 4, pp. 1379–1389, 2017.
[6] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.
[7] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[9] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 740–755.
[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.
[11] W. Qiu and A. Yuille, "UnrealCV: Connecting computer vision to unreal engine," in *Proc. Comput. Vis.–Workshops (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 909–916.
[12] A. Shafaei, J. J. Little, and M. Schmidt. (2016). "Play and learn: Using video games to train computer vision models." [Online]. Available: https://arxiv.org/abs/1608.01745
[13] A. A. Rusu, M. Vecerik, T. Rothörl, N. Heess, R. Pascanu, and R. Hadsell. (2016). "Sim-to-real robot learning from pixels with progressive nets." [Online]. Available: https://arxiv.org/abs/1610.04286
[14] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. (2016). "Unsupervised pixel-level domain adaptation with generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1612.05424
[15] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. (2016). "Learning from Simulated and unsupervised images through adversarial training." [Online]. Available: https://arxiv.org/abs/1612.07828
[16] J. T. Guibas, T. S. Virdi, and P. S. Li. (2016). "Synthetic medical images from dual generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1709.01872
[17] P. Costa *et al.*, "End-to-end adversarial retinal image synthesis," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 781–791, Mar. 2018.
[18] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[19] A. Pezeshk, N. Petrick, W. Chen, and B. Sahiner, "Seamless lesion insertion for data augmentation in CAD training," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 1005–1015, Apr. 2017.

[20] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Munich, Germany: Springer, 2015, pp. 234–241.

[22] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using CNNs," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Athens, Greece: Springer, 2016, pp. 230–238.

[23] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 597–613.

[24] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[25] A. Van Engelen *et al.*, "Multi-center MRI carotid plaque component segmentation using feature normalization and transfer learning," *IEEE Trans. Med. Imag.*, vol. 34, no. 6, pp. 1294–1305, Jun. 2015.

[26] A. V. Opbroek, M. A. Ikram, M. W. Vernooij, and M. D. Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1018–1030, May 2015.

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[28] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017.

[29] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum, "Deep MR to CT synthesis using unpaired data," in *Proc. Int. Workshop Simulation Synth. Med. Imag.* Québec City, QC, Canada: Springer, 2017, pp. 14–23.

[30] A. Osokin, A. Chessel, R. E. C. Salas, and F. Vaggi, "GANs for biological image synthesis," in *Proc. ICCV*, Oct. 2017, pp. 2233–2242.

[31] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. ChenDeep, "Adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Québec City, QC, Canada: Springer, 2017, pp. 408–416.

[32] P. Moeskops, M. Veta, M. W. Lafarge, K. A. Eppenhof, and J. P. Pluim, "Adversarial training and dilated convolutions for brain MRI segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Québec City, QC, Canada: Springer, 2017, pp. 56–64.

[33] S. Kohl *et al.* (2017). "Adversarial networks for the detection of aggressive prostate cancer." [Online]. Available: https://arxiv.org/abs/1702.08014

[34] M. Mardani *et al.* (2017). "Deep generative adversarial networks for compressed sensing automates MRI." [Online]. Available: https://arxiv.org/abs/1706.00051

[35] G. Yang *et al.*, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[36] L. Zhang, A. Gooya, and A. F. Frangi, "Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets," in *Proc. Int. Workshop Simulation Synthesis Med. Imag.* Québec City, QC, Canada: Springer, 2017, pp. 61–68.

[37] S.-H. Bae and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. 2379–2393, Nov. 2015.

[38] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Trans. Med. Imag.*, vol. 33, no. 7, pp. 1488–1502, Jul. 2014.

[39] M. Mackiewicz, J. Berens, and M. Fisher, "Wireless capsule endoscopy color video segmentation," *IEEE Trans. Med. Imag.*, vol. 27, no. 12, pp. 1769–1781, Dec. 2008.

[40] P. Mesejo *et al.*, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2051–2063, Sep. 2016.

[41] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.* Taipei, Taiwan: Springer, 2016, pp. 213–228.

[42] Ó. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel, "Visual SLAM for handheld monocular endoscope," *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 135–146, Jan. 2014.

[43] J. D. Stefansic *et al.*, "Registration of physical space to laparoscopic image space for use in minimally invasive hepatic surgery," in *Proc. 5th IEEE EMBS Int. Summer School Biomed. Imag.*, Jun. 2002, p. 12.

[44] M. S. Nosrati *et al.*, "Simultaneous multi-structure segmentation and 3D nonrigid pose estimation in image-guided robotic surgery," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 1–12, Jan. 2016.

[45] P. Haigron *et al.*, "Depth-map-based scene analysis for active navigation in virtual angioscopy," *IEEE Trans. Med. Imag.*, vol. 23, no. 11, pp. 1380–1390, Nov. 2004.

[46] A. Karargyris and N. Bourbakis, "Three-dimensional reconstruction of the digestive wall in capsule endoscopy videos using elastic video interpolation," *IEEE Trans. Med. Imag.*, vol. 30, no. 4, pp. 957–971, Apr. 2011.

[47] D. Hong, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen, "3D reconstruction of virtual colon structures from colonoscopy images," *Comput. Med. Imag. Graph.*, vol. 38, no. 1, pp. 22–33, 2014.

[48] A. Reiter, S. Lëonard, A. Sinha, M. Ishii, R. H. Taylor, and G. D. Hager, "Endoscopic-CT: learning-based photometric reconstruction for endoscopic sinus surgery," *Proc. SPIE*, vol. 9784, p. 978418, Mar. 2016.

[49] S. Nadeem and A. Kaufman, "Computer-aided detection of polyps in optical colonoscopy images," *Proc. SPIE*, vol. 9785, p. 978525, Mar. 2016.

[50] V. Parot, D. Lim, G. González, G. Traverso, N. S. Nishioka, B. J. Vakoc, and N. J. Durr, "Photometric stereo endoscopy," *J. Biomed. Opt.*, vol. 18, no. 7, p. 076017, 2013.

[51] M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, and S. Koto, "Deep monocular 3D reconstruction for assisted navigation in bronchoscopy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 7, pp. 1089–1099, 2017.

[52] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNS trained with rendered 3D model views," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2686–2694.

[53] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2315–2324.

[54] G. Varol *et al.* (2017). "Learning from synthetic humans." [Online]. Available: https://arxiv.org/abs/1701.01370

[55] B. Planche *et al.* (2017). "Depthsynth: Real-time realistic synthetic data generation from cad models for 2.5D recognition." [Online]. Available: https://arxiv.org/abs/1702.08558

[56] L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta, and J. H. Saltz. (2017). "Unsupervised histopathology image synthesis." [Online]. Available: https://arxiv.org/abs/1712.05021

[57] F. Mahmood and N. J. Durr. (2017). "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy." [Online]. Available: https://arxiv.org/abs/1710.11216

[58] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[59] J. Bernal, F. J. Sänchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[60] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2016.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[62] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[63] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. 19th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2012, pp. 1477–1480.