Uncertainty-Aware Regularization for Image-to-Image Translation

Anuja Vats, Ivar Farup, Marius Pedersen, Kiran Raja
Department of Computer Science, NTNU, Gjøvik, Norway

Abstract

The importance of quantifying uncertainty in deep networks has become paramount for reliable real-world applications. In this paper, we propose a method to improve uncertainty estimation in medical Image-to-Image (I2I) translation. Our model integrates aleatoric uncertainty and employs Uncertainty-Aware Regularization (UAR) inspired by simple priors to refine uncertainty estimates and enhance reconstruction quality. We show that by leveraging simple priors on parameters, our approach captures more robust uncertainty maps, effectively refining them to indicate precisely where the network encounters difficulties, while being less affected by noise. Our experiments demonstrate that UAR not only improves translation performance, but also provides better uncertainty estimations, particularly in the presence of noise and artifacts. We validate our approach using two medical imaging datasets, showcasing its effectiveness in maintaining high confidence in familiar regions while accurately identifying areas of uncertainty in novel/ambiguous scenarios.

1. Introduction

The significance of quantifying the uncertainty embedded in the learning process of deep neural networks has become increasingly important for identifying the blind spots [20] and biases [13] in models before their application in the real world. Despite the quest for learning from datasets that are larger, more diverse and representative, it has become evident that not all data points will adhere to the assumed distribution, leaving room for potential inaccuracies and biases in model predictions. Model uncertainty as a measure can serve to be a very useful tool for identifying the limitations to our model predictions and accounting for instances where real data may lie outside the learning distribution. Considering critical application domains such as healthcare, military, criminal justice or automated driving for deep learning models, model performance, albeit high in isolation, is being considered grossly insufficient for their adoption in practice [26]. The inability to isolate scenarios where a model isn't confident about its decision and the causes underlying that poses a significant barrier to trust and reliability in these safety-critical domains.

This article discusses uncertainty in the context of medical I2I translation. I2I is the problem of transforming an image from one domain into a corresponding image in another domain while maintaining semantic consistency and information preservation during translation. In traditional endoscopy, narrowband imaging (NBI) is used alongside standard imaging for enhanced visualization of abnormalities. However, if a region is not captured with NBI during the procedure, the enhanced information is unavailable post-procedure, limiting its use in retrospective diagnosis. Models that translate standard images to NBI are therefore critical, offering a valuable tool for post-hoc analysis when NBI was not captured. While similar functionality would be highly beneficial in capsule endoscopy, constraints like device size and battery limit real-time acquisition. Virtual chromo-endoscopy (e.g., FICE) can be applied post-capture to enhance abnormalities [25], making I2I translation particularly relevant in endoscopic imaging. Despite advancements, deep learning methods for I2I translation often produce outputs with inherent uncertainties, particularly in ambiguous or unseen scenarios. Moreover, attempts at generalization exacerbates this uncertainty, as variations in datasets or slight shifts in capture modalities can rapidly escalate uncertainty levels. Since, medical image acquisition is often prone to noise and modality-specific artifacts, it is paramount to faithfully quantify and convey model uncertainty to ascertain the extent of generalization achievable. Delineating the model's confidence levels and identifying domain gaps where it struggles, allows to effectively discern where and how to apply the model.

Model uncertainty is broadly composed to two types, the *epistemic* or uncertainty regarding the model parameters and *aleatoric* resulting from noise inherent in the data [9, 17]. The epistemic uncertainty assumes a prior distribution over model parameters and often approximated as the variance in predictions from multiple forward passes through the network with different dropout masks applied for example. The aleatoric uncertainty, on the other hand, assumes a distribution on the models outputs and is approximated using Maximum a Posteriori (MAP) estimation [21].

It has been shown that incorporating aleatoric uncertainty during learning can provide useful guidance for learning, especially in high-data regimes [17].

In this work, we aim to develop an end-to-end model for I2I translation that incorporates aleatoric uncertainty. Our primary goal is to demonstrate that imposing regularization constraints on the assumed prior distribution can improve estimation of aleatoric uncertainty during the translation process. Additionally, we show that this regularization not only provides more robust uncertainty maps but also improves the overall reconstruction quality (Table 5) , affirming that uncertainty-estimation effectively serves as guidance for improved translation [17]. This approach offers an advantage over previous multi-stage architectures [29] by eliminating the need for sequential uncertainty estimates between models. Instead, we aim to incorporate a cost-effective regularization term directly into the optimization, facilitating concurrent and mutually beneficial refinement of uncertainty and image translation within a single model.

Our main contributions are (a) a simple and model-agnostic Uncertainty-Aware Regularization (UAR), and (b) a new paired dataset for I2I translation from RGB to FICE in capsule endoscopy. Despite its simplicity, UAR not only improved translation performance (Sections 4 and 5) but allows a more faithful estimation of data-driven uncertainty in the face of commonly encountered noise-corruptions (Section 4). Finally, UAR shows improved uncertainty prediction in the presence of unforeseen structures/artifacts (Section 4.1). To understand the effects of various design choices, we conduct ablation experiments in Section 5.

2. Related Work

Medical image-to-image translation has seen significant advancements through the use of generative adversarial networks (GANs) and its variants. Typical application in medical I2I include modality translation [5, 32], image synthesis [35], segmentation [22] and super-resolution [12]. Modality translation using CycleGAN [36] has been particularly influential, enabling unsupervised translation by employing cycle consistency losses to ensure that translated images can be mapped back to the original modality. Similarly, conditional GANs [15] have allowed generation preconditioned on inputs such as anatomical labels [4], modality [8] or priors useful to generation [3]. Another class of models includes diffusion models [14, 16] that utilize parameterized Markov chains to iteratively refine data, optimizing the lower variational bound on the likelihood function [18, 10]. In WCE, image translation has been most commonly applied for image super-resolution [2, 27]. Uncertainty quantification in medical I2I has been relatively less explored. In [23] authors argue the usefulness of uncertainty estimation in MR to CT translation for detecting synthesis failures. They use traditional formulations where epistemic uncertainty is estimated by sampling from a variational distribution using dropout, and the aleatoric component is derived from the variance of the predicted distribution. Authors in [6] and [7] utilize variations of test-time augmentation for estimating uncertainty. Ayhan et al. [6] generate augmented examples for each test case to approximate the predictive distribution, whereas Baltruschat et al. leverage predictions from multiple 2D slicing planes instead of augmentations for the same goal. Our work is most closely related to [29, 30, 28] that model predictive distributions using generalized Gaussian distributions. However, unlike [29], which employs multiple sequential GANs to iteratively reduce aleatoric uncertainty, we introduce a lightweight regularization term that achieves this within a single model. As a result, our uncertainty estimates can differentiate between familiar versus newer or significantly larger sources of uncertainty, overcoming the drawback of previous methods that treat all uncertainty sources equally.

One of the primary challenges in medical I2I translation problems is the inherent ambiguity associated with image capturing mechanisms and its effect on a model's performance. Consider the case of WCE where images are often captured using low-resolution cameras under myriad distortions [33, 1], requiring significant post-processing before they are suitable for diagnosis. Noise and compression artifacts encountered during transmission further degrade the quality [11]. The cumulative impact of these factors can manifest subtly as deviations from the anticipated model performance, potentially leading to misdiagnoses. As discussed prior, one approach to mitigating this is to quantify the uncertainty associated with model predictions. Measuring the uncertainty allows detecting unaccounted shifts that can be addressed proactively. Despite relevance, uncertainty quantification and refinement is relatively nascent in I2I translation problems.

3. Methodology

We introduce both the conventional and probabilistic formulations of paired I2I translation, highlighting their limitations. Subsequently, we present the proposed UAR for improving uncertainty estimation and guidance.

3.1. I2I Translation Formulation

Consider a collection of input images from a domain $\mathcal{A}$ denoted as $\mathcal{X}_\mathcal{A} := \{x_1^a, x_2^a, ..., x_n^a\}$, and another set of paired images originating from a domain $\mathcal{B}$, expressed as $\mathcal{X}_\mathcal{B} := \{x_1^b, x_2^b, ..., x_n^b\}$. The dataset $\mathcal{D}$ comprises pairs (x_i^a, x_i^b) drawn from the respective domains $\mathcal{A}$ and $\mathcal{B}$. The objective is to learn the underlying conditional distribution $\mathcal{P}_{\mathcal{B}|\mathcal{A}}$ facilitating the translation of images from $\mathcal{A} \rightarrow \mathcal{B}$.

As shown in Fig.1, this can typically be achieved by minimizing the point estimate for per-pixel residual at jk, $\delta_{jk} = ||\hat{x}_{jk}^b - x_{jk}^b||^2$ between the reconstructed and ground-

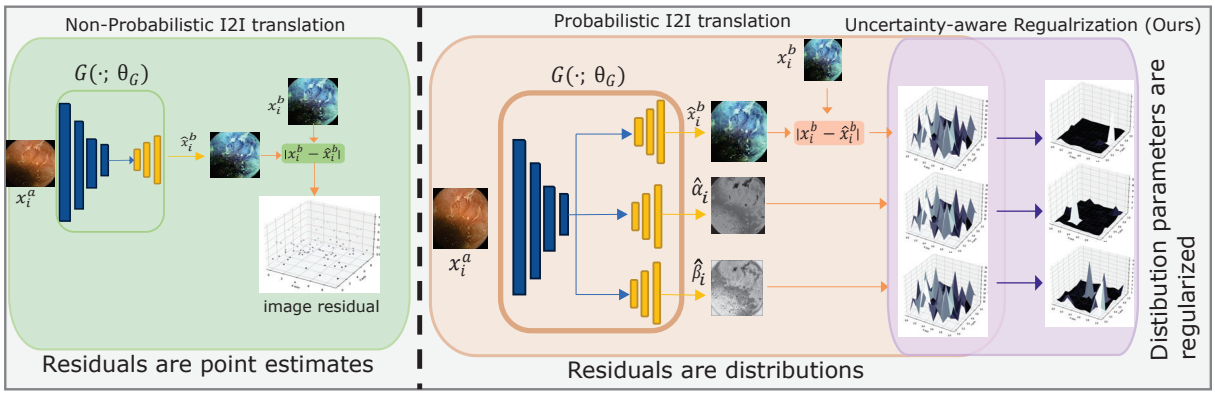

Figure 1. Non-probabilistic image translation (green) optimizes point-estimates for the residuals between the predicted and the target images. On the contrary, the probabilistic approach (orange) models the residuals with a distribution, allowing the variance of errors to change spatially. Our method takes this a step further (purple) by regularizing the predicted variances (or distribution parameters) to achieve more precise uncertainty estimation (The discriminator is omitted from the diagram for simplicity).

truth image from domain $\mathcal{B}$. However, in pixel reconstruction tasks, the solution space is often multimodal, meaning that multiple outputs can yield acceptable solutions. Thus, relying solely on point-wise estimation fails to adequately represent the distribution $\mathcal{P}_{\mathcal{B}|\mathcal{A}}$ over the output space, as well as estimate the uncertainty associated with the reconstruction process. The probabilistic remedy for this is to relax the constraint on the residual by modeling it as a distribution instead of a point estimate, the optimal parameters of which are learned from the data, thus allowing an estimation of the uncertainty. As an example, consider a deep learning model $\mathcal{F}(\mathcal{D};\theta)$ parametrized by θ to be trained for translating images from domain $\mathcal{A} \rightarrow \mathcal{B}$. While one conceivable distribution for δ might be an isotropic standard Gaussian, presuming a fixed variance not only imposes an assumption of independence and identical distribution (i.i.d.) on the residuals, which can be easily compromised by slightly out-of-distribution samples [17, 29], but also eliminates the ability to model heteroscedasticity in predictions. Alternatively, the distribution over δ can be heteroscedastic Gaussian [29] with zero mean and spatially varying-learnable standard deviation σ_{jk} as in Eq. 1,

$$\hat{x}_{jk} = x_{jk} + \delta_{jk}, \quad \delta_{jk} \sim \mathcal{N}(0, \sigma_{jk}^2); \quad \hat{x}_{jk} \sim \mathcal{N}(x_{jk}, \sigma_{jk}^2) \tag{1}$$

The parameters of the network $\mathcal{F}(\mathcal{D};\theta)$ can be optimized by maximizing the likelihood given by:

$$\begin{aligned} \mathcal{L}(\mathcal{D};\theta) &:= \prod_{i=1}^{n} \mathcal{P}_{\mathcal{B}|\mathcal{A}}(x_i^b; \{\hat{x}_i^b, \hat{\sigma}_i\}) \\ \theta^* &:= \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\mathcal{D};\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} \mathcal{P}_{\mathcal{B}|\mathcal{A}}(x_i^b; \{\hat{x}_i^b, \hat{\sigma}_i\}) \\ \theta^* &= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} e^{-\frac{|\hat{x}_i^b - x_i^b|^2}{2\hat{\sigma}_i^2}} \end{aligned} \tag{2}$$

where we omit spatial indices jk for simplicity. The negative log likelihood is,

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \frac{|\hat{x}_i^b - x_i^b|^2}{2\hat{\sigma}_i^2} + \frac{\log(\hat{\sigma}_i^2)}{2} \right\}. \tag{3}$$

Assuming that the residuals follow a normal distribution simplifies uncertainty estimation, as the per-pixel variance σ_i^2 itself is the aleatoric uncertainty in prediction. This formulation for modeling aleatoric uncertainty can be improved by assuming a more lenient Generalized Normal Distribution (GND) with zero mean over the residuals [29, 28]. The parameters governing the shape (β) and scale (α) of the predicted distribution not only accommodate the heteroscedastic variations in residuals, but also enable heavier-tails, which are beneficial for handling outliers.

$$\delta_{jk} \sim GND(\delta; 0, \alpha_{jk}, \beta_{jk}) \tag{4}$$

As before, the likelihood can be written as:

$$\begin{aligned} \mathcal{L}(\mathcal{D};\theta) &:= \prod_{i=1}^{n} \mathcal{P}_{\mathcal{B}|\mathcal{A}}(x_i^b; \{\hat{x}_i^b, \hat{\alpha}_i, \hat{\beta}_i\}) \\ \theta^* &:= \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\mathcal{D};\theta) \\ \theta^* &= \arg\max_{\theta} \prod_{i=1}^{n} \frac{\hat{\beta}_i}{2\hat{\alpha}_i \Gamma(\frac{1}{\hat{\beta}_i})} e^{-\left(\frac{|\hat{x}_i^b - x_i^b|}{\hat{\alpha}_i}\right)^{\hat{\beta}_i}} \end{aligned} \tag{5}$$

Therefore, the negative likelihood is,

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \left(\frac{|\hat{x}_i^b - x_i^b|}{\hat{\alpha}_i} \right)^{\hat{\beta}_i} - \log \frac{\hat{\beta}_i}{\hat{\alpha}_i} + \log\, \Gamma(\frac{1}{\hat{\beta}_i}) \right\} \tag{6}$$

We refer to this loss as the negative likelihood loss L_{nll}, in next sections. The aleatoric uncertainty for $\hat{x}_i^b$ can be

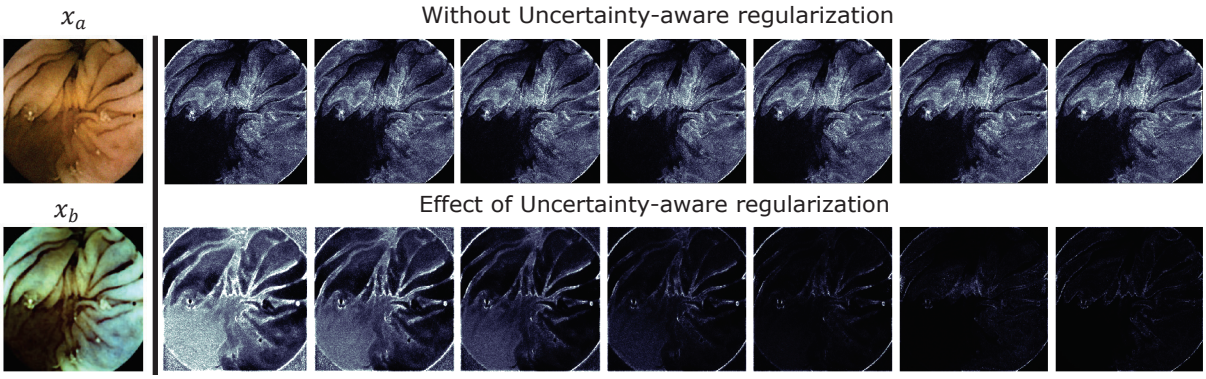

Figure 2. The figure shows the shape β parameter and predicted aleatoric uncertainty of an image at different epochs during the training. Without regularization (first row), the variances in the predictions (uncertainty) remain relatively the same throughout training. In contrast, with regularization (second row), the predicted uncertainty gets progressively less noisy and more semantically refined over the course of the training.

written down as the variance of this distribution, given by $\frac{\hat{\alpha}_i^2\,\Gamma(3/\hat{\beta}_i)}{\Gamma(1/\hat{\beta}_i)}$. The generalized normal distribution proves highly effective in encapsulating uncertainties arising from shifts due to noise and changes in modality, which often manifest as outliers within datasets. The loss in Eq.6 consists of a fidelity term along with general constraints on the shape and scale of the residual distribution to prevent divergence to infinity. But, given that I2I translation can be characterized by a lack of a unique stable solution, incorporating explicit constraints on the parameters of the residual distribution into the objective function, typically in the form of a penalty benefits to progressively refine uncertainty estimation. This is discussed in the next section.

3.2. Uncertainty-Aware Regularization

We operate under the benign assumption that for good reconstructions pixel-residuals exhibit piece-wise continuity similar to images, implying that since adjacent pixels within one image region show minimal discrepancies their residuals should also be similar, unless influenced by noise. Therefore, large residuals can come from pixels of two types, the pixels that the network actually finds hard to reconstruct to due to lack of knowledge or data drifts, and, the spurious pixels that might not strictly correspond to difficulty in reconstruction but end up having high values. To illustrate this better, we simulate this effect by injecting a small amount of noise in an image (Fig. 3), such that the corruption is visually imperceptible in the image and predict the uncertainty. As expected, the aleatoric uncertainty map is adversely affected, even with comparable reconstructions. Although sensitivity to noise in input data is generally advantageous, excessive sensitivity within the anticipated noise spectrum can result in unreliable and inaccurate uncertainty predictions. We propose to suppress this spurious component for a more accurate estimation of uncertainty by penalizing large differences in the predicted residual distributions for neighboring pixels.

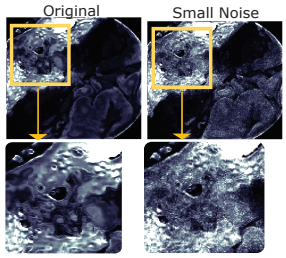

Figure 3. Uncertainty estimate is sensitive to small changes in input and network parameters, resulting in potentially noisy maps.

This prior assumption can be incorporated by adding a penalty/regularization term that discourages significant deviations between the predicted residual distributions of adjacent pixels, inline with the expectation that neighboring pixels in an image are likely to have similar residuals, unless there is noise or an edge. Further, while enforcing the above constraint, it is crucial to prevent accidentally suppressing those deviations that occur as a result of the network's incapacity/lack of knowledge to reconstruct an input. This is the interesting case when the network is unsure how to reconstruct the output for one or more regions. Thus, we propose to impose a total-variation based penalty on the estimated shape parameter $\hat{\beta}$ during the learning process, to smooth out noise in the estimated parameters across neighboring pixels while preserving regions of true uncertainty.

For a predicted $\hat{\beta}_i$ image corresponding to input x_i^b, the total variation is shown in Eq. 7.

$$TV(\beta_i) = \int_v |\nabla \hat{\beta}_i(v)| dv \tag{7}$$

The approximation in two dimension yields,

$$R_{\beta_i} = \sum_{jk} \sqrt{(\hat{\beta}_{ij+1k} - \hat{\beta}_{ijk})^2 + (\hat{\beta}_{ijk+1} - \hat{\beta}_{ijk})^2} \quad (8)$$

To prevent derivative singularity, a small positive constant $\epsilon = 10^{-7}$ is added, introducing some blurring,

$$R_{\beta_i} = \sum_{jk} \sqrt{\epsilon^2 + (\hat{\beta}_{ij+1k} - \hat{\beta}_{ijk})^2 + (\hat{\beta}_{ijk+1} - \hat{\beta}_{ijk})^2} \quad (9)$$

Fig.2 shows the impact of Eq.9 on uncertainty-estimation over the course of training, highlighting in contrast to its absence. Without regularization, as expected although the parameters α and β are predicted, they are not subsequently optimized, even as the image reconstruction continues to improve (due to pixel-wise and GAN loss terms in the objective, equations 10-12). On the other hand, in the regularized variant, residual errors continue to reduce the in tandem with optimizing for parameters α and β, resulting in cleaner and visually more interpretable uncertainty maps. One notable effect of the UAR penalty on the map is the accentuation of edges around uncertain structures, owing to the edge-preserving nature of total-variation. This facilitates visual interpretation of uncertainty maps, as they correspond to image structures and their uncertainty levels.

Model

Our model is composed of a single conditional GAN consisting of a generator $G(\cdot;\theta_G)$, and a discriminator $D(\cdot;\theta_D)$. The discriminator follows the commonly used patch architecture [15, 18] while the generator is based on U-Net [24]. Like [29], the generator outputs spatially varying α and β, along with the output image from domain $\mathcal{B}$. In addition to the negative log likelihood loss (Eq. 6) and chosen variation-based regularization (Eq. 8), the generator is trained using the adversarial loss L_{adv} defined as a mean-squared error between the discriminators predictions for the generated image (Eq. 10) against the label vectors of ones. This formulation is an alternative to the commonly used cross-entropy loss.

$$L_{adv} = \frac{1}{n} \sum_i \text{MSE}(\hat{x}_i^b, 1) \quad (10)$$

Finally, an additional L1-fidelity term, $L_1 = |x_i^b - \hat{x}_i^b|$ is added to enforce pixel-level reconstruction fidelity between the image and its reconstruction. Thus, the total loss for the generator is given by

$$L_G = w_{L_1} L_1 + w_{adv} L_{adv} + w_{nll} L_{nll} + \lambda R_{\beta_i} \quad (11)$$

where w_{L_1}, w_{adv}, w_{nll} and λ are the respective weights for each term. The discriminator is trained using the above-mentioned mean-squared error, with target vectors one for real images and zeros for the generated images.

$$L_D = \frac{1}{2} \left[\frac{1}{n} \sum_i \text{MSE}(\hat{x}_i^b, 0) + \frac{1}{n} \sum_i \text{MSE}(x_i^b, 1) \right] \quad (12)$$

3.3. Training Details and Evaluation Metrics

We test UAR on two datasets, a new WCE dataset and a public colonoscopy CPC-paired dataset [19]. From the WCE dataset, 5,000 images were utilized for training, and 5,000 for validation. All results are reported on a test-set of another 5,000 image pairs, which is further divided into three subsets for comprehensive evaluation. The training and validation images are sourced from WCE videos of seven patients, while the test images are obtained from three new patients, potentially containing new or different abnormalities. The hyperparameters optimized on the WCE dataset were also effective for the CPC-paired dataset. Consequently, the CPC-paired dataset was split into training and testing subsets (80:20), with results reported on the test set.

Both the discriminator and generator utilize the Adam optimizer with an initial learning rate of 10^{-4}, following a cosine annealing schedule for learning rate adjustment. The outputs of the discriminator are passed through an average pooling layer before applying the MSE loss (equations 10 and 12). We found that results improved when the variation-based regularization (Eq. 8) was activated a little later in the initial learning phase, giving the network a chance to predict unregularized values for α and β. Thus, the total variation regularization is activated around epoch 5. Through experimentation, we found that a value of 10^{-12} for the regularization weight, λ in Eq. 11 yielded satisfactory results, with room for further optimization and performance improvements (more details in Section 5). Other weights in Eq.11 are $w_{L_1} = 1$, $w_{adv} = 10^{-3}$ and $w_{nll} = 10^{-4}$. All models were trained with an image size of 490×490 and a batch size of 4, using twin-titan RTX GPUs with 48 GB of RAM, achieving a processing speed of approximately 20 images per second. The UAR term can be integrated with minimal computational overhead, as it involves only element-wise operations on the grayscale maps of β resulting in execution speeds comparable to those of L1 loss.

To assess the quality of the generated images, we report the results on four metrics, namely Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [31], Relative Root Mean Squared Error (RRMSE) and Learned Perceptual Image Patch Similarity (LPIPS) [34]. While, SSIM, PSNR and RRMSE are more common, we use LPIPS additionally as it has shown to correlate better with human visual perception [34] over pixel-wise metrics.

3.4. Dataset

In this work, we introduce a new paired image-to-image translation dataset for capsule endoscopy. The dataset fa-

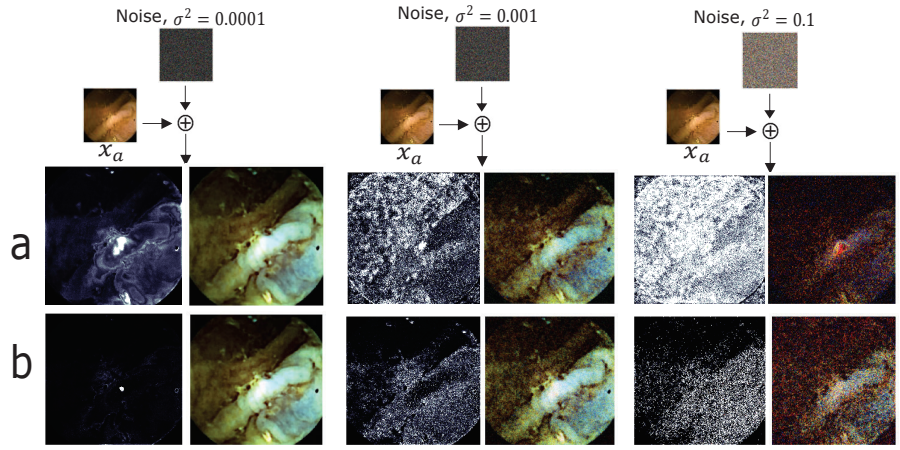

Figure 4. Impact of noise on uncertainty prediction and image reconstruction. The figure illustrates the impact of varying levels of Gaussian noise on image reconstruction and predicted aleatoric uncertainty. Row (a) depicts the non-regularized version, whereas row (b) shows the regularized variant. In the presence of noise, the non-regularized version is more significantly affected, with uncertainty maps rapidly diverging from the original regions of uncertainty. In contrast, the regularized version demonstrates greater robustness to noise.

cilitates the translation between original WCE images and their corresponding Flexible Spectral Imaging Color Enhancement (FICE) mode images, and vice versa.

The images were collected during capsule endoscopy trials conducted on patients at Innlandet Hospital, Norway. The trials involved the use of a capsule endoscope to capture images of the gastrointestinal tract in established patients, which were later converted to their corresponding FICE versions using a WCE diagnostic software called rapid reader. The dataset comprises over 15,000 pairs of carefully curated WCE images, which can be employed for paired as well as unpaired image-translation methods. The dataset is available at `https://doi.org/10.18710/BSXNA1`.

4. Results

This section presents the qualitative and quantitative evaluation of our method on the two datasets. The approach is tested on three types of commonly occurring noises: Gaussian, Uniform and Impulse (also called salt and pepper noise) at different levels. The baseline corresponds to I2I-translation without regularization as in [29].

Fig.4 shows the qualitative effect of increasing levels of noise on the predicted uncertainty and reconstruction. Comparing the reconstructed image, it is seen that the regularized variant results in a more visually coherent reconstruction even at high noise levels, as compared to the non-regularized method.

Fig.5 shows, the residual errors and the uncertainty maps derived from the two methods, under the impact of noise. As seen in columns 4 and 7 (σ^2), UAR generates less noisy uncertainty maps, consistent with the distinctive features within the images, while reducing the residual errors (columns 3 and 6 ($||x - \hat{x}||^2$)).

Given the consistent capture modality and similar patient population, the primary source of noise in the data is the added noise itself. If the predicted uncertainties accurately reflect this, they should be low for familiar image structures and higher in noise-affected regions. Uncertainty maps generated using UAR adhere to this expectation. Conversely, in the absence of regularization, the uncertainties are uniformly high, obscuring relative differences in uncertainty and hindering interpretability. Further, we analyze the quantitative impact of UAR on reconstruction quality in Table 1 and 2. It is seen that the effect of UAR is overall positive on image reconstruction with equivalent or better SSIM and PSNR values, across different noise types and levels. The regularization also consistently improves the LPIPS and RRMSE metrics, across both datasets.

4.1. Impact of Artifacts

Additionally, we evaluate the performance of uncertainty estimation by systematically introducing more pronounced artifacts into the image. Fig.6 illustrates images with circular artifacts. The UAR variant prominently displays high uncertainty across the entire artifact region, with comparatively lower uncertainties in other areas of the image. In contrast, the baseline method fails to differentiate the network's confidence between these two regions effectively.

Fig.7 replaces the circular artifact with a ring artifact to examine behaviors near the artifact boundaries. Here again, the baseline method significantly underestimates the uncertainty associated with the artifact, whereas UAR accurately delineates uncertainty regions with precise boundaries (notice last row in Fig.7).

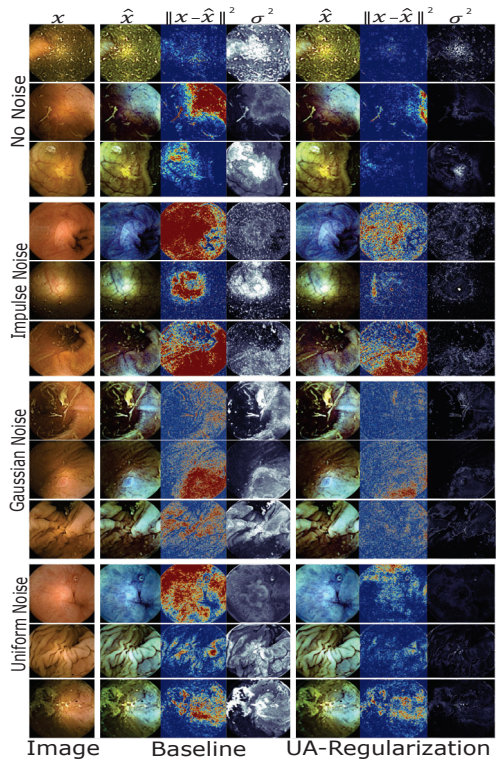

Figure 5. Qualitative Comparison. As can be seen from columns 3 and 6, while the regions of residual errors are consistent between the regularized and non-regularized variant, UAR shows consistently low residual errors. Correspondingly, the uncertainty maps are less-noisy and more structurally coherent.

		Approach	SSIM ↑	PSNR ↑	LPIPS ↓	RRMSE ↓
Gaussian	-	Baseline	0.925	28.714	0.128	0.174
		UAR (Ours)	**0.931**	**29.34**	**0.126**	**0.148**
	$\mathcal{N}(0, 0.001)$	Baseline	0.639	26.610	0.275	0.212
		UAR (Ours)	**0.650**	**27.04**	**0.261**	**0.204**
	$\mathcal{N}(0, 0.01)$	Baseline	**0.310**	**22.240**	0.452	0.399
		UAR (Ours)	0.309	22.023	**0.431**	**0.372**
Uniform	$\mathcal{U}(0, 0.1)$	Baseline	0.680	**27.007**	0.287	0.212
		UAR (Ours)	**0.695**	26.421	**0.268**	**0.208**
	$\mathcal{U}(0, 0.01)$	Baseline	0.912	28.14	0.135	0.173
		UAR (Ours)	**0.928**	**29.38**	**0.134**	**0.161**
Impulse	$\mathcal{I}(0.005)$	Baseline	**0.735**	26.63	0.380	0.226
		UAR (Ours)	0.724	**26.91**	**0.371**	**0.216**
	$\mathcal{I}(0.01)$	Baseline	**0.601**	25.23	0.442	0.268
		UAR (Ours)	0.5847	25.26	**0.440**	**0.256**

Table 1. Impact of uncertainty-guidance on reconstruction quality. UAR consistently achieves lower LPIPS and RRMSE values across various types and levels of noise, with comparable or superior SSIM and PSNR metrics.

Model	SSIM↑	PSNR↑	LPIPS↓	RRMSE↓
Baseline	0.891	35.358	0.410	0.291
UAR (Ours)	**0.925**	**38.297**	**0.289**	**0.221**

Table 2. Impact of uncertainty-guidance on reconstruction quality on CPC-dataset [19]. Further results in the supplementary.

5. Ablation

Ablation I: Other variation-based losses. Given that the primary goal of the regularization term is to attenuate spurious variances between nearby pixels, other types of variation-based losses are also conceivable. We experiment with two other variations, and analyze their effect on the uncertainty estimation. We hypothesize that, at a minimum, imposing similar penalties should not negatively affect the reconstruction quality for more faithful uncertainty maps.

One such penalty could be simply to penalize the squared L2-norm of gradients of the β map. This modifies Eq.9 so that it is differentiable and avoids singularity. However, it comes at the cost of reduced invariance to sharp features, in other words it introduces slight smoothing in the uncertainty map. This variant is referred to as UAR$_{L2}$.

$$R_{\beta_i} = \sum_{jk} \left((\hat{\beta}_{ij+1k} - \hat{\beta}_{ijk})^2 + (\hat{\beta}_{ijk+1} - \hat{\beta}_{ijk})^2 \right) \quad (13)$$

Next, we test the regularization of the anisotropic variant of total variation. This is the L1-norm on the gradients of β_i given in its discrete form by,

$$R_{\beta_i} = \sum_{jk} \left(\sqrt{(\hat{\beta}_{ij+1k} - \hat{\beta}_{ijk})^2} + \sqrt{(\hat{\beta}_{ijk+1} - \hat{\beta}_{ijk})^2} \right) \quad (14)$$

We compare the effects of these different penalty formulations on the reconstruction quality (Table 3) (as well as qualitatively on the generated uncertainty maps in supplementary). Imposing these constraints does not negatively impact the reconstruction quality, as seen in Table 3. As expected, The uncertainty maps for UAR$_{\text{L2}}$ are smoother compared to UAR and UAR$_{\text{Aniso}}$. This is because, while the TV variant prioritizes preserving edges around the uncertain structures, the edges in UAR$_{\text{L2}}$ have been smoothed out, though the regions of uncertainty remain consistent. The choice of the best variant may depend on the application's demands or the user preference.

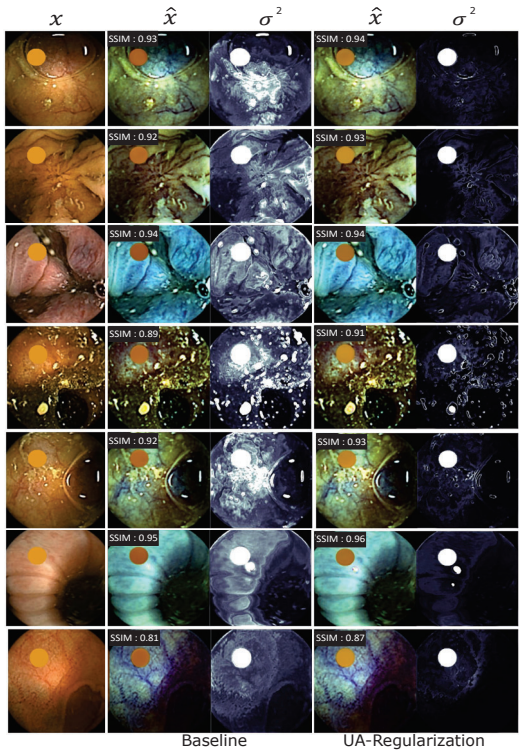

Figure 6. UAR distinctly identifies regions affected by the artificially introduced artifact as having high uncertainty, contrasting with relatively lower uncertainty in unaffected areas. In contrast, the baseline approach shows similar uncertainty levels across different regions, failing to differentiate between previously unseen and unseen image structures.

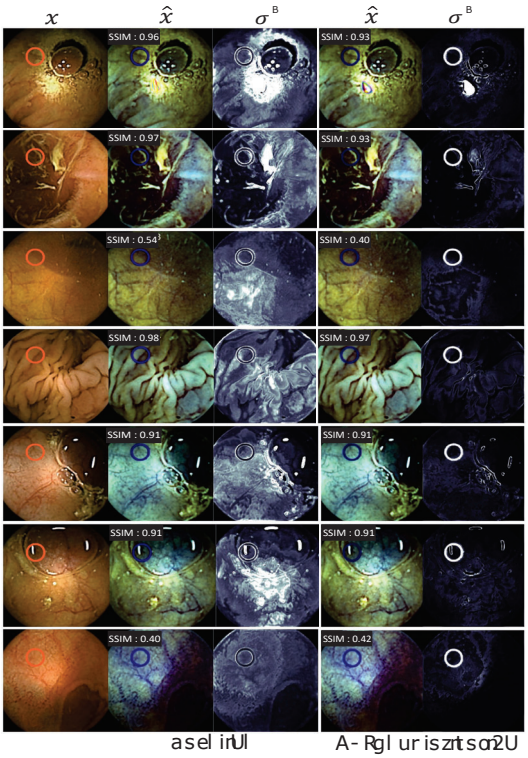

Figure 7. Synthetic artifact and its effect on uncertainty. The uncertainty estimation provided by UAR faithfully reflects the presence of the injected artifact. In the last row, where the artifact is subtle, it induces low uncertainty except around edges in the Baseline method. However, UAR accurately identifies it as an unseen region with high uncertainty.

Overall, each of the regularization results in significantly less noisy maps than those without regularization. Given that the testing dataset is similar to the training dataset, low uncertainties are expected, except in the presence of unseen artifacts, as shown in Figures 6 and 7. For qualitative comparison, please refer to the supplementary material.

Model	SSIM↑	PSNR↑	LPIPS↓	RRMSE↓
Baseline	0.925	28.714	0.128	0.174
UAR_{L2}	0.922	27.283	**0.126**	0.149
$\text{UAR}_{\text{Aniso}}$	0.927	**29.825**	0.133	0.215
UAR	**0.931**	29.340	**0.126**	**0.148**

Table 3. Effect of different variation-based penalties on WCE data.

Ablation II : λ. We conducted experiments with three values, $10^{-12}, 10^{-7}$ and 10^{-4}, to capture behaviors across a wide range. For a high value of $\lambda = 10^{-4}$, the regularization effect on β is excessively strong. This causes the predicted β values to become too similar, suppressing any disparities. Conversely, $\lambda = 10^{-7}$ strikes a balance, offering effective regularization without excessively homogenizing the β values. In contrast, using $\lambda = 10^{-12}$ as employed in this study reflects a cautious approach, yielding satisfactory results. We anticipate that the optimal value for λ to be within the range $[10^{-7}, 10^{-12}]$ (details in supplementary).

6. Conclusion and Limitations

In this work, we presented an end-to-end model for I2I translation that integrates an uncertainty-aware regularization. UAR aims at ensuring that the model's confidence levels are clearly delineated and easily interpretable while improving the overall reconstruction quality, thereby facilitating better decision-making in safety-critical applications. Through systematic evaluation and ablation studies, we demonstrated that our approach maintains high fidelity in familiar regions while accurately identifying and quantifying uncertainty in novel situations. This paper employs a basic conditional GAN for I2I translation, but more advanced architectures and improved reconstruction losses could enhance translation quality. Since UAR is model-agnostic, it can be seamlessly integrated with these improvements. Additionally, we plan to involve more clinical experts to assess the quality of uncertainty maps, complementing the current qualitative and quantitative evaluations in the future.

References

[1] Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James E East, Xin Lu, and Jens Rittscher. A deep learning framework for quality assessment and restoration in video endoscopy. *Medical image analysis*, 68:101900, 2021.

[2] Yasin Almalioglu, Kutsev Bengisu Ozyoruk, Abdulkadir Gokce, Kagan Incetan, Guliz Irem Gokceler, Muhammed Ali Simsek, Kivanc Ararat, Richard J Chen, Nicholas J Durr, Faisal Mahmood, et al. Endol2h: deep super-resolution for capsule endoscopy. *IEEE Transactions on Medical Imaging*, 39(12):4297–4309, 2020.

[3] Hamed Amini Amirkolaee and Hamid Amini Amirkolaee. Medical image translation using an edge-guided generative adversarial network with global-to-local feature fusion. *Journal of Biomedical Research*, 36(6):409, 2022.

[4] Sina Amirrajab, Yasmina Al Khalil, Cristian Lorenz, Jürgen Weese, Josien Pluim, and Marcel Breeuwer. Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks. *Computerized Medical Imaging and Graphics*, 101:102123, 2022.

[5] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

[6] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

[7] Ivo M Baltruschat, Parvaneh Janbakhshi, Melanie Dohmen, and Matthias Lenga. Uncertainty estimation in contrast-enhanced mr image translation with multi-axis fusion. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[8] Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10):2375–2388, 2019.

[9] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[10] Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan. Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *International conference on medical image computing and computer-assisted intervention*, pages 339–349. Springer, 2023.

[11] Pål Anders Floor, Ivar Farup, Marius Pedersen, and Øistein Hovde. Error reduction through post processing for wireless capsule endoscope video. *EURASIP Journal on Image and Video Processing*, 2020:1–15, 2020.

[12] Yuchong Gu, Zitao Zeng, Haibin Chen, Jun Wei, Yaqin Zhang, Binghui Chen, Yingqin Li, Yujuan Qin, Qing Xie, Zhuoren Jiang, et al. Medsrgan: medical images super-resolution using generative adversarial networks. *Multimedia Tools and Applications*, 79:21815–21840, 2020.

[13] Jessica Guynn. Google photos labeled black people'gorillas'. *USA today*, 1, 2015.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[16] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846, 2023.

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[18] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 702–716. Springer, 2016.

[19] Weijie Ma, Ye Zhu, Ruimao Zhang, Jie Yang, Yiwen Hu, Zhen Li, and Li Xiang. Toward clinically assisted colorectal polyp recognition via structured cross-modal representation consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 141–150. Springer, 2022.

[20] NHTSA. Pe 16-007 technical report, u.s. department of transportation, national highway traffic safety administration. `https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF`, Jan 2017. [Accessed 22-Feb-2023].

[21] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

[22] Moritz Platscher, Jonathan Zopes, and Christian Federau. Image translation for medical image generation: Ischemic stroke lesion segmentation. *Biomedical Signal Processing and Control*, 72:103283, 2022.

[23] Jacob C Reinhold, Yufan He, Shizhong Han, Yunqiang Chen, Dashan Gao, Junghoon Lee, Jerry L Prince, and Aaron Carass. Validating uncertainty in medical image translation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 95–98. IEEE, 2020.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[25] Yasushi Sato, Tamotsu Sagawa, Masahiro Hirakawa, Hiroyuki Ohnuma, Takahiro Osuga, Yutaka Okagawa, Fumito Tamura, Hiroto Horiguchi, Kohichi Takada, Tsuyoshi Hayashi, et al. Clinical utility of capsule endoscopy with

flexible spectral imaging color enhancement for diagnosis of small bowel lesions. *Endoscopy international open*, 2(02):E80–E87, 2014.

[26] David Schneeberger, Karl Stöger, and Andreas Holzinger. The european legal framework for medical ai. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 209–226. Springer, 2020.

[27] Mehmet Turan. A generative adversarial network based super-resolution approach for capsule endoscopy images. *Medicine Science*, 10(3):1002–1007, 2021.

[28] Uddeshya Upadhyay, Yanbei Chen, and Zeynep Akata. Robustness via uncertainty-aware cycle consistency. *Advances in neural information processing systems*, 34:28261–28273, 2021.

[29] Uddeshya Upadhyay, Yanbei Chen, Tobias Hepp, Sergios Gatidis, and Zeynep Akata. Uncertainty-guided progressive gans for medical image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 614–624. Springer, 2021.

[30] Uddeshya Upadhyay, Viswanath P Sudarshan, and Suyash P Awate. Uncertainty-aware gan with adaptive loss for robust mri image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3255–3264, 2021.

[31] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[32] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I-Chao Chang, and Yan Xu. Mri cross-modality image-to-image translation. *Scientific reports*, 10(1):3753, 2020.

[33] Diana E Yung, John N Plevris, Romain Leenhardt, Xavier Dray, Anastasios Koulaouzidis, and ESGE Small Bowel Research Working Group. Poor quality of small bowel capsule endoscopy images has a significant negative effect in the diagnosis of small bowel malignancy. *Clinical and Experimental Gastroenterology*, pages 475–484, 2020.

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[35] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251, 2018.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.