Underwater Image Enhancement for Improved Visual Perception

PEDDIPALLY YUHITHA VANDANA-20MIS1138
M. MOUNIKA-20MIS1154
THANDLE YESWANTH RAO-20MIS1141

***Abstract*—In this letter, we present a conditional generative adversarial network-based model for real-time underwater image enhancement. To supervise the adversarial training, we formulate an objective function that evaluates the perceptual image quality based on its global content, color, local texture, and style information. We also present EUVP, a large-scale dataset of a paired and an unpaired collection of underwater images (of 'poor' and 'good' quality) that are captured using seven different cameras over various visibility conditions during oceanic explorations and human-robot collaborative experiments. In addition, we perform several qualitative and quantitative evaluations which suggest that the proposed model can learn to enhance underwater image quality from both paired and unpaired training. More importantly, the enhanced images provide improved performances of standard models for underwater object detection, human pose estimation, and saliency prediction. These results validate that it is suitable for real-time preprocessing in the autonomy pipeline by visually-guided underwater robots.**

I. Introduction

VISUALLY-GUIDED AUVs (Autonomous Underwater Vehicles) and ROVs (Remotely Operated Vehicles) are widely used in important applications such as the monitoring of marine species migration and coral reefs , inspection of submarine cables and wreckage , underwater scene analy-sis, seabed mapping, human-robot collaboration , and more.One major operational challenge for these underwater robotsis that despite using high-end cameras, visual sensing is often greatly affected by poor visibility, light refraction, absorption, and scattering. These optical artifacts trigger non-linear distortions in the captured images, which severely affect the performance of vision-based tasks such as tracking, detection and classification, segmentation, and visual servoing. Fast and accurate image enhancement techniques can alleviate these problems by restoring the perceptual and statistical qualities of the distorted images in real-time.

As light propagation differs underwater (than in the atmosphere), a unique set of non-linear image distortions occur which are propelled by a variety of factors. For instance, underwater images tend to have a dominating green or blue hue because red wavelengths get absorbed in deep water (as light travels further). Such wavelength dependant attenuation , scattering, and other optical properties of the waterbodies cause irregular non-linear distortions ,whichresult in low-contrast, often blurred, and color-degraded im- ages. Some of these aspects can be modeled and well esti- mated by physics-based solutions, particularly for dehazing and color correction .However, information such as the scene depth and optical water-quality measures are not always available in many robotic applications. Besides, these models are often computationally too demanding for real-timedeployments.

A practical alternative is to approximate the underlying solution by learning-based methods, which demonstrated remarkable success in recent years. Several models based on deep Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) provide state-of-the-art performance in learning to enhance perceptual image quality from a large collection of paired or unpaired data. For underwater imagery,in particular, a number of GAN-based models and CNN-based residual models report inspiring progress for automatic color enhancement, dehazing, and contrast adjustment. However, there is significant room for improvement as learning perceptual enhancement for underwater imagery is a more challenging ill-posed problem (than terrestrial imagery). Additionally, due to the high costs and difficulties associated with acquiring large-scale underwater data, most learning-based models use small-scale and often only synthetically generated images that fail to capture a wide range of natural variability. Moreover, designing robust yet efficient image enhancement models and investigating their applicability for improving real- time underwater visual perception have not been explored in the literature in depth.

We attempt to address these challenges by designing a fast underwater image enhancement model and analyzing its feasibility for real-time applications. We formulate the problem as an image-to-image translation problem by assuming there exists a non-linear mapping between the distorted (input) and enhanced (output) images. Then, we design a conditional GAN-based model to learn this mapping by adversarial training on a large-scale dataset named EUVP (Enhancement of Underwater

Visual Perception). From the perspective of its design, implementation, and experimental validation, we make the following contributions in this letter:

a) We present a fully-convolutional conditional GAN-based model for real-time underwater image enhancement, which we refer to as FUnIE-GAN. We formulate a multi- modal objective function to train the model by evaluating the perceptual quality of an image based on its global content, color, local texture, and style information.
b) Additionally, we present the EUVP dataset, a paired and an unpaired collection of 20K underwater images (of poor and good quality) that can be used for *one-way* and *two-way* adversarial training.
c) Furthermore, we present qualitative and quantitative performance evaluations compared to state-of-the-art models. The results suggest that FUnIE-GAN can learn to enhance perceptual image quality from both paired and unpaired training. More importantly, the enhanced images significantly boost the performance of several underwater visual perception tasks such as object detection, human pose estimation, and saliency prediction; a few sample demonstrations are highlighted in Fig. 1.

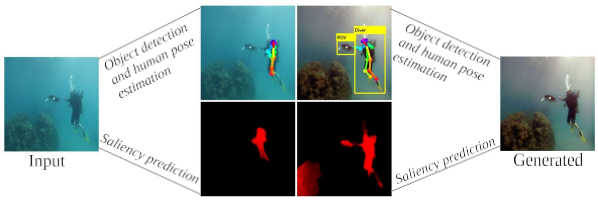
In addition to presenting the conceptual model of FUnIE-GAN, we analyze important design choices and relevant practicalities for its efficient implementation. We also conduct a user study and a thorough feasibility analysis to validate its effectiveness for improving the real-time perception performance of visually-guided underwater robots.

II. Related Work

A. Automatic Image Enhancement

Automatic image enhancement is a well-studied problem in the domains of computer vision, robotics, and signal processing. Classical approaches use hand-crafted filters to enforce local color constancy and improve contrast/lightness rendition. Additionally, prior knowledge or statistical assumptions about a scene (e.g., haze-lines, dark channel prior, etc.) are often utilized for global enhancements such as image deblurring, dehazing , etc. Over the last decade, single image enhancement has made remarkable progress due to the advent of deep learning and the availability of large-scale datasets. The contemporarydeep CNN-based models provide state-of-the-art performancefor problems such as image colorization , color/contrast adjustment, dehazing, etc. These models learn a sequence of nonlinear filters from paired training data, which providemuch better performance compared to using hand-crafted filters. Moreover, the GAN-based models have shown great success for style-transfer and image-to-image translation prob-lems . They employ a two-player min-max game wherethe 'generator' tries to fool the 'discriminator' by generating*fake* images that appear to be sampled from the *real* distri-bution. Simultaneously, the discriminator tries to get betterat discarding fake images and eventually (in equilibrium) the generator learns to model the underlying distribution. Although

(a) Perceptual enhancement of underwater images.



(b) Improved performance for underwater object detection [7], human body-pose estimation [8], and saliency prediction [9].

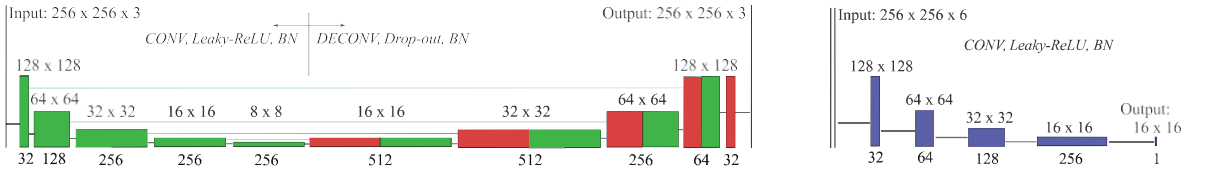
Fig. 1. Demonstration of underwater image enhancement using our proposed model and its practical feasibility.

such adversarial training can be unstable, several tricks and choices of loss functions are proposed in the literature to mitigate that. For instance, Wasserstein GAN improves the training stability by using the earth-mover distance to measure the distance between the data distribution and the model distribution. Energy-based GANs also improve training stability by modeling the discriminator as an energy function, whereas the Least-Squared GAN addresses the vanishing gradients problem by adopting a least-square loss function for the discriminator. On the other hand, conditional GANs allow constraining the generator to produce samples that follow a pattern or belong to a specific class, which is particularly useful to learn a pixel-to-pixel (Pix2Pix) mapping [16] between an arbitrary input domain (e.g., distorted images) and the desired output domain (e.g., enhanced images).

A major limitation of the above-mentioned models is that they require paired training data, which may not be available or can be difficult to acquire for many practical applications. The two-way GANs (e.g., CycleGAN , DualGAN , etc.) solve this problem by using a 'cycle-consistency loss' that allows learning the mutual mappings between two domains from unpaired data. Such models have been effectively used for unpaired learning of perceptual image enhancement as well. Furthermore, Ignatov *et al.* showed that additional loss-terms for preserving the high-level feature-based content improve the quality of image enhancement using GANs.

B. Improving Underwater Visual Perception

Traditional physics-based methods use the atmospheric de- hazing model to estimate the *transmission* and *ambient* light in a scene to recover true pixel intensities. Another class of methods design a series of bilateral and trilateral filters to reduce noise and improve global contrast . In recent work, Akkaynak *et al.* proposed a revised imaging model



(a) Generator: five encoder-decoder pairs with mirrored skip-connections (inspired by the success of U-Net [30]; however, it is a much simpler model).

(b) Discriminator: a Markovian PatchGAN [16] with four layers and a patch-size of 16×16.

Fig. 2. Network architecture of the proposed model: FUnIE-GAN.

that accounts for the unique distortions pertaining to underwater light propagation; this contributes to a more accurate color reconstruction and overall a better approximation to the ill-posed underwater image enhancement problem. Nevertheless, these methods require scene depth (or multiple images) and optical waterbody measurements as prior.

On the other hand, several single image enhancement models based on deep adversarial and residual learn- ing have reported inspiring results of late. However, they typically use only synthetically distorted images for paired training, which often limit their generalization performance. The extent of large-scale unpaired training on naturally distorted underwater images have not been explored in the literature. Moreover, most existing models fail to ensure fast inference on single-board robotic platforms, which limits their applicability for improving real-time visual perception. We attempt to address these aspects in this letter.

III. Proposed Model and Dataset

A. FUnIE-GAN Architecture

Given a source domain X (of distorted images) and desired domain Y (of enhanced images), our goal is to learn a mapping $G : X \rightarrow Y$ in order to perform automatic image enhancement. We adopt a conditional GAN-based model where the generator tries to learn this mapping by evolving with an adversarial discriminator through an iterative min-max game. As illustrated in Fig. 2, we design a generator network by following the principles of U-Net. It is an encoder-decoder network (e_1-e_5, d_1-d_5) with connections between the mirrored layers, i.e., between (e_1, d_5), (e_2, d_4), (e_3, d_2), and (e_4, d_4). Specifically, the outputs of each encoders are concatenated to the respective mirrored decoders. This idea of *skip-connections* in the generator network is shown to be very effectivefor image-to-image translation and image quality enhancement problems. In FUnIE-GAN, however, we employ a much simpler model with fewer parameters in order to achieve fast inference. The input to the network is set to $256 \times 256 \times 3$ and the encoder (e_1-e_5) learns only 256 feature-maps of size 8×8. The decoder (d_1-d_5) utilizes these feature-maps and inputs from the skip-connections to learn to generate a $256 \times 256 \times 3$ (enhanced) image as output. The network is fully-convolutional as no fully-connected layers are used. Additionally, 2D convolutions with 4×4 filters are applied at each layer, which is then followed by a Leaky-ReLU non-linearity and Batch Normalization (BN) . The feature-map sizes in each layer and other model parameters are annotated in Fig. 2(a).

For the discriminator, we employ a Markovian Patch-GAN [16] architecture that assumes the independence of pixels beyond the patch-size, i.e., only discriminates based on the patch-level information. This assumption is important to effectively capture high-frequency features such as local texture and style. In addition, this configuration is computationally more efficient as it requires fewer parameters compared to discriminating globally at the image level. As shown in Fig. 2(b), four convolutional layers are used to transform a $256 \times 256 \times 6$ input (real and generated image) to a $16 \times 16 \times 1$ output that represents the averaged *validity* responses of the discriminator. At each layer, 3×3 convolutional filters are used with a stride of 2; then the non-linearity and BN are applied the same way as the generator.

B. Objective Function Formulation

A standard conditional GAN-based model learns a mapping $G : \{X, Z\} \rightarrow Y$, where X (Y) represents the source (desired) domain, and Z denotes random noise. The conditional adversarial loss function [29] is expressed as:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{X,Y}\,[\log D(Y)] + \mathbb{E}_{X,Y}\,[\log(1 - D(X, G(X, Z)))]$$

Here, the generator G tries to minimize $\mathcal{L}_{cGAN}$ while the discriminator D tries to maximize it. In FUnIE-GAN, we associate three additional aspects, i.e., global similarity, image content, and local texture and style information in the objective to quantify perceptual image quality.

- **Global similarity**: existing methods have shown that adding an L_1 (L_2) loss to the objective function enables G to learn to sample from a globally similar space in an L_1 (L_2) sense [16], [18]. Since the L_1 loss is less prone to introduce blurring, we add the following loss term in the objective:

$$\mathcal{L}_1(G) = \mathbb{E}_{X,Y,Z}\,\| Y - G(X, Z) \|_1$$

- **Image content**: we add a *content loss* term in the objective in order to encourage G to generate enhanced image that has similar content (i.e., feature representation) as the target (i.e., real) image. Being inspired by [14], [36], we define the image content function $\Phi(.)$ as the high-level features extracted by the `block5_conv2` layer of a pre-trained VGG-19 network. Then, we formulate the content loss as follows:

$$\mathcal{L}_{con}(G) = \mathbb{E}_{X,Y,Z}\,\| \Phi(Y) - \Phi(G(X, Z)) \|_2$$

- **Local texture and style**: as mentioned, Markovian PatchGANs are effective in capturing high-frequency information pertaining to the local texture and style [16]. Hence, we rely on D to enforce the local texture and style consistency in adversarial fashion.

1) Paired Training: For paired training, we formulate an objective function that guides G to learn to improve the perceptual image quality so that the generated image is close to the respective ground truth in terms of its global appearance and high-level feature representation. On the other hand, D will discard a generated image that has locally inconsistent texture and style. Specifically, we use the following objective function for paired training:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda_1 \mathcal{L}_1(G) + \lambda_c \mathcal{L}_{con}(G)$$

Here, $\lambda_1 = 0.7$ and $\lambda_c = 0.3$ are scaling factors that we empirically tuned as hyper-parameters.

2) Unpaired Training: For unpaired training, we do not enforce the global similarity and content loss constraints as the pairwise ground truth is not available. Instead, the objective is to learn both the forward mapping $G_F : \{X, Z\} \rightarrow Y$ and the reconstruction $G_R : \{Y, Z\} \rightarrow X$ simultaneously by maintaining cycle-consistency. As suggested by Zhu *et al.*], we formulate the cycle-consistency loss as follows:

$$\mathcal{L}_{cyc}(G_F, G_R) = \mathbb{E}_{X,Y,Z}\left[\|X - G_R(G_F(X, Z))\|_1\right] + \mathbb{E}_{X,Y,Z}\left[\|Y - G_F(G_R(Y, Z))\|_1\right] \quad (4)$$

Therefore, our objective for the unpaired training is:

$$G_F^*, G_R^* = \arg\min_{G_F, G_R} \max_{D_Y, D_X} \mathcal{L}_{cGAN}(G_F, D_Y) + \mathcal{L}_{cGAN}(G_R, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G_F, G_R)$$

Here, D_Y (D_X) is the discriminator associated with the generator G_F (G_R), and the scaling factor $\lambda_{cyc} = 0.1$ is an empirically tuned hyper-parameter. We do not enforce additional global similarity loss-term because the $\mathcal{L}_{cyc}$ computes analogous reconstruction loss for each domain in L_1 space.

C. EUVP Dataset

The EUVP dataset contains a large collection of paired and unpaired underwater images of poor and good perceptual quality. We used seven different cameras, which include multiple GoPros [37], Aqua AUV's uEye cameras [38], low-light USB cameras [39], and Trident ROV's HD camera [40], to capture images for the dataset. The data was collected during oceanic explorations and human-robot cooperative experiments in different locations under various visibility conditions. Additionally, images extracted from a few publicly available YouTube videos are included in the dataset. The images are carefully selected to accommodate a wide range of natural variability (e.g., scenes, waterbody types, lighting conditions, etc.) in the data.

The unpaired data is prepared, i.e., good and poor quality images are separated based on visual inspection by six human participants. They inspected several image properties (e.g., color, contrast, and sharpness) and considered whether the scene is visually interpretable, i.e., foreground/objects are identifiable. Hence, the unpaired training endorses the modeling of human perceptual preferences of underwater image quality. On the other hand, the paired data is prepared by following a procedure suggested in [6]. Specifically, a CycleGAN-based model is trained on our unpaired data to learn the domain transformation between the good and poor quality images. Subsequently, the good quality images are distorted by the learned model to generate respective pairs; we also augment a set of underwater images from the ImageNet dataset and from Flickr.

There are over 12K paired and 8K unpaired instances in the EUVP dataset; a few samples are provided in Fig. 3. It is to be noted that our focus is to facilitate *perceptual image enhancement* for boosting robotic scene understanding, not to model the underwater optical degradation process for *image restoration*, which requires scene depth and waterbody properties.

IV. Experimental Results

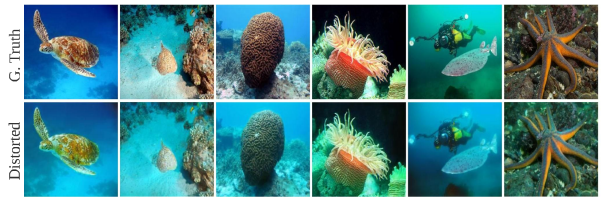
We use TensorFlow libraries [42] to implement the FUnIE-GAN model. It is trained separately on 11K paired and 7.5K unpaired instances; the rest are used for respective validation and testing. Four NVIDIA GeForce GTX 1080 graphics cards are used for training; both models are trained for 60K-70K iterations with a batch-size of 8. We now present the experimental evaluations based on a qualitative analysis, standard quantitative metrics, and a user study.

A. Qualitative Evaluations

We first qualitatively analyze the enhanced color and sharpness of the FUnIE-GAN-generated images compared to their respective ground truths. As Fig. 4a shows, the true color, and

sharpness is mostly recovered in the enhanced images. Additionally, as shown in Fig. 4b, the greenish hue in underwater images are rectified and the global contrast is enhanced. These are the primary characteristics of an effective underwater image enhancer. We further demonstrate the contributions of each loss- terms of FUnIE-GAN: global similarity loss ($_1$), and image content loss ($_{con}$), for learning the enhancement. We observe that the $_1$ term helps to generate sharper images, while the $_{con}$ term contributes to furnishing finer texture details (see Fig. 4c). Moreover, we found slightly better numeric stability for $_{con}$ with the `block5_conv2` layer of VGG-19 compared to its last feature extraction layer (`block5_conv4`).

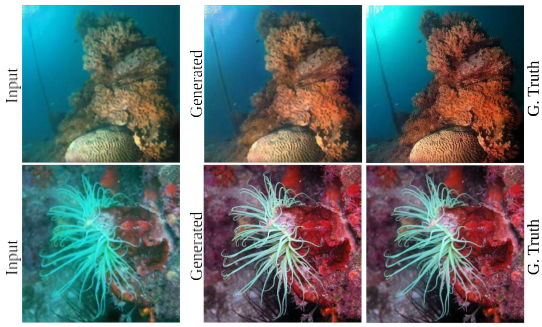
Next, we conduct a qualitative comparison of perceptual im- age enhancement by FUnIE-GAN with several state-of-the-art models. We consider five learning-based models: (i) underwater GAN with gradient penalty (UGAN-P), (ii) Pix2Pix ,
(iii) least-squared GAN (LS-GAN [28]), (iv) GAN with residual blocks in the generator (Res-GAN), and (v) Wasserstein GAN with residual blocks in the generator (Res-WGAN). These models are implemented with 8 encoder-decoder pairs (or 16 residual blocks) in the generator network and 5 convolutional layers in the discriminator. They are trained on the paired EUVP dataset using the same setup as the FUnIE-GAN. Addition- ally, we consider CycleGAN as a baseline for comparing the performance of FUnIE-GAN with unpaired training (i.e.,



(a) Paired instances: ground truth images and their respective distorted pairs are shown on the top and bottom row, respectively.

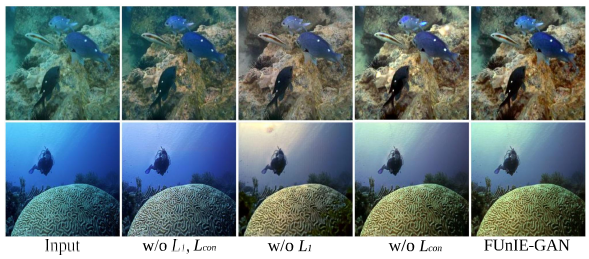
(b) Unpaired instances: good and poor quality images are shown on the top and bottom row (in no particular order), respectively.

Fig. 3. A few sample images from the EUVP dataset are shown.



(a) True color and sharpness is restored in the enhanced image.

(b) The greenish hue is rectified and global contrast is enhanced.



(c) Ablation experiment: learning enhancement without (w/o) $\mathcal{L}_1$ and $\mathcal{L}_{con}$, w/o $\mathcal{L}_1$, and w/o $\mathcal{L}_{con}$ loss-terms in FUnIE-GAN.

Fig. 4. Demonstration of improved image attributes by FUnIE-GAN in terms of color, sharpness, and contrast.

FUnIE-GAN-UP). We also include two physics-based models in the comparison: Multi-band fusion-based enhancement (Mbad-EN, and haze-line-aware color restoration (Uw-HL). A common test set with 1K images (of 256×256 resolution) are used for the qualitative evaluation; it also includes 72 images with known waterbody types . A few sample comparisons are illustrated in Fig. 5.

As demonstrated in Fig. 5, Res-GAN, Res-WGAN, and Mbad-EN often suffer from over-saturation, while LS-GAN generally fails to rectify the greenish hue in images. UGAN-P, Pix2Pix, and Uw-HL perform reasonably well and their enhanced images are comparable to that of FUnIE-GAN; however, UGAN-P often over-saturates bright objects in the scene while Pix2Pix fails to enhance global brightness in some cases. On the other hand, we observe that achieving color consistency and hue rectification are relatively more challenging through unpaired learning. This is mostly because of the lack of reference color or texture information in the loss function. Nevertheless, FUnIE-GAN-UP still outperforms CycleGAN in general. Overall, FUnIE-GAN performs as well and often better without using scene depth or prior waterbody information as the physics-based models, and despite having a much simpler network architecture compared to the existing learning-based models.

B. *Quantitative Evaluation*

We consider two standard metrics named Peak Signal-to-Noise Ratio (PSNR) and Structural Similar- ity (SSIM) in order to quantitatively compare FUnIE-GAN- enhanced images with their respective ground truths. The PSNR approximates the reconstruction quality of a generated image $\mathbf{x}$ compared to its ground truth $\mathbf{y}$ based on their Mean Squared Error (MSE) as follows:

$$PSNR(\mathbf{x}, \mathbf{y}) = 10 \log_{10} \quad 255^2 / MSE(\mathbf{x}, \mathbf{y})$$

On the other hand, the SSIM [45] compares the image patches based on three properties: luminance, contrast, and structure. It is defined as follows:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + c_1}{\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 +} \quad \frac{2\sigma_{\mathbf{xy}} + c_2}{\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 +}$$

In Eq. 6, $\mu_{\mathbf{x}}$ ($\mu_{\mathbf{y}}$) denotes the mean, and σ^2 (σ^2) is the variance of $\mathbf{x}$ ($\mathbf{y}$); whereas $\sigma_{\mathbf{xy}}$ denotes the cross-correlation between $\mathbf{x}$ and $\mathbf{y}$. Additionally, $c_1 = (255 \times 0.01)^2$ and $c_2 = (255 \times 0.03)^2$ are constants that ensure numeric stability.

In Table I, we provide the averaged PSNR and SSIM values over 1K test images for FUnIE-GAN and compare the results with the same models used in the qualitative evaluation. The results indicate that FUnIE-GAN performs best on both PSNR and SSIM metrics. We conduct a similar analysis for Underwater Image Quality Measure (UIQM), which quantifies underwater image colorfulness, sharpness, and contrast. We present the results in Table II, which indicates that although FUnIE-GAN-UP performs better than CycleGAN, its UIQM

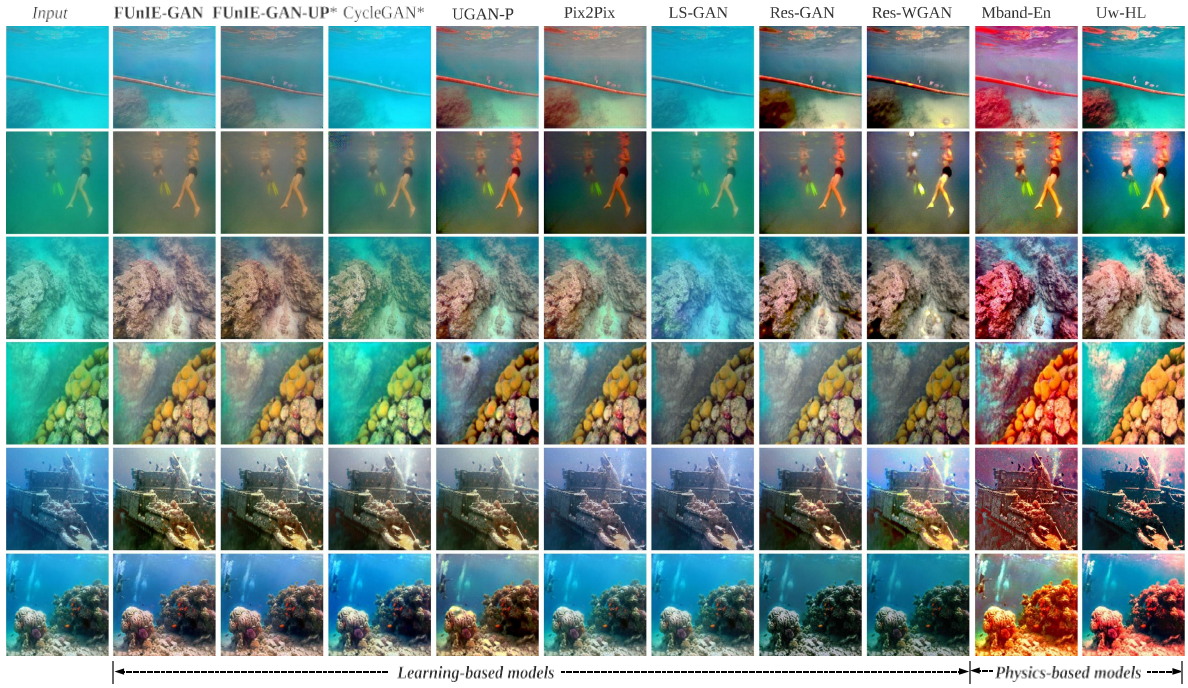


Fig. 5. Qualitative performance comparison of FUnIE-GAN and FUnIE-GAN-UP with learning-based methods: CycleGAN [17], UGAN-P [6], Pix2Pix [16], LS-GAN [28], Res-GAN [43], and Res-WGAN [26]; the super-scripted asterisk (*) denotes unpaired training. Two physics-based models: Mband-EN [32] and Uw-HL [13], are also included in the comparison. (Best viewed at 400% zoom).

TABLE I
QUANTITATIVE COMPARISON FOR AVERAGE PSNR AND SSIM VALUES ON 1K PAIRED TEST IMAGES OF EUVP DATASET

Model	$PSNR(G(\mathbf{x}), \mathbf{y})$ Input: 17.27 ± 2.88	$SSIM(G(\mathbf{x}), \mathbf{y})$ Input: 0.62 ± 0.075
Uw-HL	18.85 ± 1.76	0.7722 ± 0.066
Mband-EN	12.11 ± 2.55	0.4565 ± 0.097
Res-WGAN	16.46 ± 1.80	0.5762 ± 0.014
Res-GAN	14.75 ± 2.22	0.4685 ± 0.122
LS-GAN	17.83 ± 2.88	0.6725 ± 0.062
Pix2Pix	20.27 ± 2.66	0.7081 ± 0.069
UGAN-P	19.59 ± 2.54	0.6685 ± 0.075
CycleGAN	17.14 ± 2.65	0.6400 ± 0.080
FUnIE-GAN-UP	21.36 ± 2.17	0.8164 ± 0.046
FUnIE-GAN	21.92 ± 1.07	0.8876 ± 0.068

TABLE II
QUANTITATIVE COMPARISON FOR AVERAGE UIQM VALUES ON 1K PAIRED AND 2K UNPAIRED TEST IMAGES OF EUVP DATASET

Model	**Paired data** Input: 2.20 ± 0.69 G. Truth: 2.91 ± 0.65	**Unpaired data** Input: 2.29 ± 0.62 G. Truth: N/A
Uw-HL	2.62 ± 0.35	2.75 ± 0.32
Mband-EN	2.28 ± 0.87	2.34 ± 0.45
Res-WGAN	2.55 ± 0.64	2.46 ± 0.67
Res-GAN	2.62 ± 0.89	2.28 ± 0.34
LS-GAN	2.37 ± 0.78	2.59 ± 0.52
Pix2Pix	2.65 ± 0.55	2.76 ± 0.39
UGAN-P	2.72 ± 0.75	2.77 ± 0.34
CycleGAN	2.44 ± 0.71	2.62 ± 0.67
FUnIE-GAN-UP	2.56 ± 0.63	2.81 ± 0.65
FUnIE-GAN	2.78 ± 0.43	2.98 ± 0.51

values on the the paired dataset are relatively poor. Interestingly, the models trained on paired data, particularly FUnIE-GAN, UGAN-P, and Pix2Pix, produce better results. We postulate that

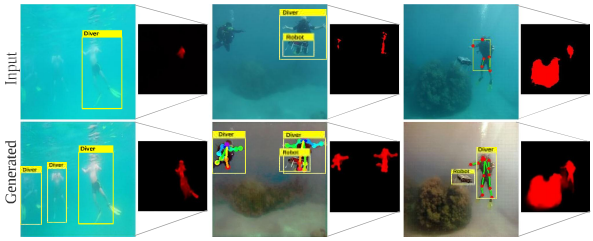
TABLE III
RANK-n ACCURACY (n = 1, 2, 3) FOR THE TOP FOUR MODELS BASED ON 312 RESPONSES PROVIDED BY 78 INDIVIDUALS

Model	Rank-1 (%)	Rank-2 (%)	Rank-3 (%)
FUnIE-GAN	24.50	68.50	88.60
FUnIE-GAN-UP	18.67	48.25	76.18
UGAN-P	21.25	65.75	80.50
Pix2Pix	11.88	45.15	72.45

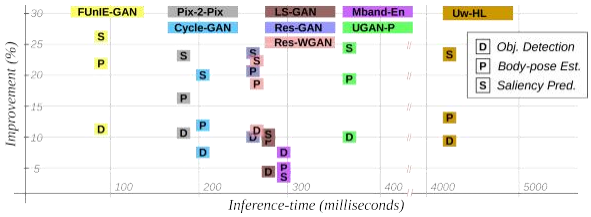
the global similarity loss in FUnIE-GAN and Pix2Pix, or the gradient-penalty term in UGAN-P contribute to this enhancement, as they all add L_1 terms in the adversarial objective. Our ablation experiments of FUnIE-GAN (see Fig. 4c) reveal that the L_1 loss-term contributes to 4.58% improvements in UIQM, while L_{con} contributes 1.07%. Moreover, without both L_1 and L_{con} loss-terms, the average UIQM values drop by 17.6%; we observe similar statistics for PSNR and SSIM as well.

C. User Study

We also conduct a user study to add human preferences to our qualitative performance analysis. The participants are shown different sets of 9 images (one for each learning-based models) and asked to rank top 3 best quality images. A total of 78 individuals participated in the study and a total of 312 responses are recorded. Table III compares the average rank-1, rank-2, and rank-3 accuracy of the top 4 categories. The average rank-3 accuracy of the original images is recorded to be 6.67, which suggests that the users clearly preferred enhanced images over the original ones. Moreover, the results indicate that the users prefer the images enhanced by FUnIE-GAN, UGAN-P,



(a) A few snapshots showing qualitative improvement on FUnIE-GAN-generated images; a detailed demonstration can be found at: https://youtu.be/1ewcXQ-jgB4.



(b) Improvement versus inference-time comparison with the state-of-the-art models; FUnIE-GAN offers over 10 FPS speed (on common platform: Intel Core-i5 3.6GHz CPU); note that the run-times are evaluated on 256×256 image patches for all the models.

Fig. 6. Improved performance for object detection, saliency prediction, and human body-pose estimation on enhanced images.

and Pix2Pix compared to the other models; these statistics are consistent with our qualitative and quantitative analysis.

D. Improved Visual Perception

As demonstrated in Fig. 6a, we conduct further experiments to quantitatively interpret the effectiveness of FUnIE-GAN-enhanced images for underwater visual perception over a variety of test cases. We analyze the performance of standard deep visual models for underwater object detection , 2D human body-pose estimation , and visual attention-based saliency prediction; although results vary depending on the image qualities of a particular test set, on an average, we observe 11-14%, 22-28%, and 26-28% improvements, respectively. We also evaluate other state-of-the-art models on the same test sets; as Fig. 6b suggests, images enhanced by UGAN-P, Res-GAN, Res-WGAN, Uw-HL, and Pix2Pix also achieve considerable performance improvements. However, these models offer significantly slower inference-rates than FUnIE-GAN, most of which are not suitable for real-time deployment in robotic platforms.

FUnIE-GAN's memory requirement is 17 MB and it operates at a rate of 25.4 FPS (frames per second) on a single-board computer (NVIDIA Jetson TX2), 148.5 FPS on a graphics card (NVIDIA GTX 1080), and 7.9 FPS on a robot CPU (Intel Core-i3 6100 U). These computational aspects are ideal for it to be used as an image processing pipeline by visually-guided underwater robots in real-time applications.

E. Limitations and Failure Cases

We observe a couple of challenging cases for FUnIE-GAN, which are depicted by a few examples in Fig. 7. First, FUnIE-GAN is not very effective for enhancing severely degraded and

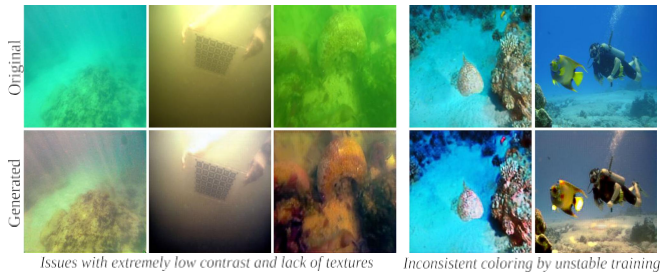


Fig. 7. Extremely low-contrast and texture-less images are generally challenging for FUnIE-GAN, whereas FUnIE-GAN-UP often suffers from inconsistent coloring due to training instability.

texture-less images. The generated images in such cases are often over-saturated by noise amplification. Although the hue rectification is generally correct, the color and texture recovery remains poor. Secondly, FUnIE-GAN-UP is prone to training instability. Our investigations suggest that the discriminator often becomes too good too early, causing a *diminishing gradient* effect that halts the generator's learning. As shown in Fig. 7 (right), the generated images in such cases lack color consistency and accurate texture details. This is a fairly common issue in unpaired training of GANs , and requires meticulous hyper-parameter tuning.

FUnIE-GAN balances a trade-off between robustness and efficiency which limits its performance to a certain degree. More powerful deep models (i.e., denser architectures with more parameters) can be adopted for non-real-time applications; moreover, the input/output layers can be modified with additional bottleneck layers for learning enhancement at higher resolution than 256×256. On the other hand, FUnIE-GAN does not guarantee the recovery of true pixel intensities as it is designed for perceptual image quality enhancement. If scene depth and optical waterbody properties are available, underwater light propagation and image formation characteristics can be incorporated into the optimization for more accurate image restoration.

V. Conclusion

We present a simple yet efficient conditional GAN-based model for underwater image enhancement. The proposed model formulates a perceptual loss function by evaluating image quality based on its global color, content, local texture, and style information. We also present a large-scale dataset containing a paired and an unpaired collection of underwater images for supervised training. We perform extensive qualitative and quantitative evaluations, and conduct a user study which show that the proposed model performs as well and often better compared to the state-of-the-art models, in addition to ensuring much faster inference time. Moreover, we demonstrate its effectiveness in improving underwater object detection, saliency prediction, and human body-pose estimation performances. In the future, we plan to investigate its feasibility in other underwater human-robot cooperative applications, marine trash identification, etc. We seek to improve its color consistency and stability for unpaired training as well.

REFERENCES

Shkurti *et al.*, "Multi-domain monitoring of marine environments using a heterogeneous robot team," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 1747–1753.

B. Bingham *et al.*, "Robotic tools for deep water archaeology: Surveying an ancient shipwreck with an autonomous underwater vehicle," *J. Field Robot.*, vol. 27, no. 6, pp. 702–717, 2010.

M. J. Islam, M. Ho, and J. Sattar, "Understanding human motion and gestures for underwater human-robot collaboration," *J. Field Robot.*, pp. 1–23, 2018.