Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception

Md Jahidul Islam*, Peigen Luo† and Junaed Sattar‡
Department of Computer Science and Engineering, Minnesota Robotics Institute,
University of Minnesota Twin Cities, MN 55455, United States.
Email: {*islam034, †luo00034, ‡junaed}@umn.edu

Abstract—In this paper, we introduce and tackle the simultaneous enhancement and super-resolution (SESR) problem for underwater robot vision and provide an efficient solution for near real-time applications. We present Deep SESR, a residual-in-residual network-based generative model that can learn to restore perceptual image qualities at $2\times$, $3\times$, or $4\times$ higher spatial resolution. We supervise its training by formulating a multi-modal objective function that addresses the chrominance-specific underwater color degradation, lack of image sharpness, and loss in high-level feature representation. It is also supervised to learn salient foreground regions in the image, which in turn guides the network to learn global contrast enhancement. We design an end-to-end training pipeline to jointly learn the saliency prediction and SESR on a shared hierarchical feature space for fast inference. Moreover, we present UFO-120, the first dataset to facilitate large-scale SESR learning; it contains over 1500 training samples and a benchmark test set of 120 samples. By thorough experimental evaluation on UFO-120 and several other standard datasets, we demonstrate that Deep SESR outperforms the existing solutions for underwater image enhancement and super-resolution. We also validate its generalization performance on several test cases that include underwater images with diverse spectral and spatial degradation levels, and also terrestrial images with unseen natural objects. Lastly, we analyze its computational feasibility for single-board deployments and demonstrate its operational benefits for visually-guided underwater robots.

Fig. 1: The proposed 'Deep SESR' model offers perceptually enhanced HR image generation and saliency prediction by a single efficient inference. The enhanced images restore color, contrast, and sharpness at higher scales (up to $4\times$) to facilitate an improved visual perception, whereas the saliency map can be further exploited for attention modeling. All figures in this paper are best viewed digitally by zoom for colors and details.

I. INTRODUCTION

Automatic generation of high resolution (HR) images from low resolution (LR) sensory measurements is a well-studied research problem in the domains of robotics and computer vision [64, 72, 34]. For visually-guided robots, in particular, this single image super-resolution (SISR) capability allows *zooming-in* regions of interests (RoIs) for detailed perception, to eventually make navigational and other operational decisions. However, if the LR images suffer from noise and optical distortions, those get amplified by SISR, resulting in uninformative RoIs. Hence, restoring perceptual and statistical image qualities is essential for robust visual perception in noisy environments (*e.g.*, underwater [9, 25]). Although large bodies of literature on perceptual image enhancement and SISR offer solutions separately for both, a unified approach is more viable for computationally constrained real-time applications.

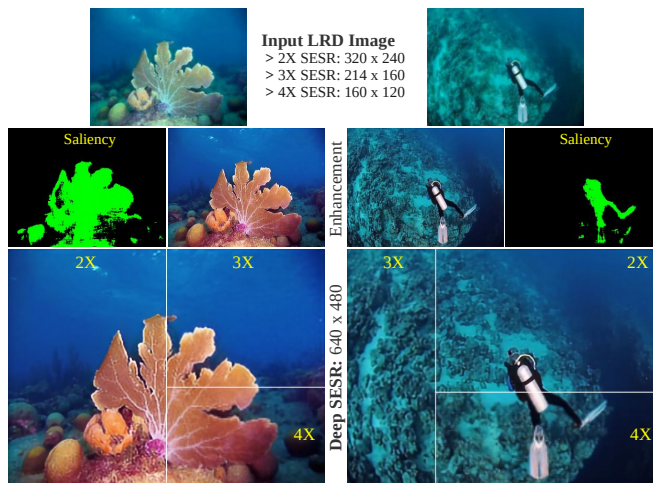To this end, we introduce **simultaneous enhancement and super-resolution (SESR)** and demonstrate its effectiveness for improved visual perception. SESR is particularly useful in the underwater domain due to its unique optical properties [2], *e.g.*, attenuation, refraction, and backscatter. These artifacts cause range-and-wavelength-dependent non-linear distortions that severely affect vision despite often using high-end cameras [35]. Specifically, the captured images exhibit various levels of hue distortion, blurriness, low contrast, and color degradation based on the waterbody types, distances of light sources, etc. Some of these aspects can be modeled and estimated by physics-based solutions, particularly for dehazing [7], color correction [10], water removal [3], etc. However, these methods are often computationally too demanding for real-time deployments. Besides, dense scene depth and optical waterbody measures are not always available in practice.

The learning-based approaches attempt to address the practicalities by approximating the underlying solution to the ill-posed problem of underwater image restoration with RGB data alone. Several existing models based on convolutional neural networks (CNNs) [48, 63] and generative adversarial networks (GANs) [35, 44, 20] provide state-of-the-art (SOTA) perfor-

mance for perceptual color enhancement, dehazing, deblurring, and contrast adjustment. Additionally, inspired by the success of deep residual networks for terrestrial SISR [72, 42, 30], several models have been proposed for underwater SISR in recent years [13, 34], which report exciting results with reasonable computational overhead. Contemporary research work [35, 34] further demonstrates that the perceptually enhanced underwater images provide significantly improved performance for widely-used object detection and human body-pose estimation tasks; moreover, detailed perception on salient image regions facilitates better scene understanding and attention modeling. However, as mentioned, separately processing visual data for these capabilities, even with the fastest available solutions, is not computationally feasible on single-board platforms.

In this paper, we present the first unified approach for SESR with an end-to-end trainable model. The proposed **Deep SESR** architecture incorporates dense residual-in-residual sub-networks to facilitate multi-scale hierarchical feature learning for SESR and saliency prediction. For supervision, we formulate a multi-modal objective function that evaluates the degree of chrominance-specific color degradation and loss in image sharpness, contrast, and high-level feature representation. As demonstrated in Fig. 1, it learns to restore perceptual image qualities at higher spatial scales (up to $4\times$); as a byproduct, it learns to identify salient foreground regions in the image. We also present the **UFO-120 dataset**, which contains over 1500 annotated samples for large-scale SESR training, and a test set with an additional 120 samples. The dataset and relevant resources are available for academic research purposes at http://irvlab.cs.umn.edu/resources/ufo-120-dataset.

Furthermore, we evaluate the perceptual enhancement and super-resolution performance of Deep SESR on UFO-120 and several other standard datasets. The results suggest that it provides superior performance over SOTA methods on respective tasks, and achieves considerably better generalization performance on unseen natural images. Finally, we specify several design choices for Deep SESR, analyze their computational aspects, and discuss the usability benefits for its robotic deployments.

II. Background

Underwater image enhancement is an active research problem that deals with correcting optical image distortions to recover true pixel intensities [3, 10]. Classical approaches use hand-crafted filters to improve local contrast and enforce color constancy. These approaches are inspired by the *Retinex theory* of human visual perception [37, 71, 23], and mainly focus on restoring background illumination and lightness rendition. Another class of physics-based approaches uses an atmospheric dehazing model to estimate true *transmission* and *ambient* light in a scene [15, 27]. Additional prior knowledge or statistical assumptions (*e.g.*, haze-lines, dark channel prior [7], etc.) are often utilized for global enhancements. Recent work by Akkaynak *et al.* [2, 3] introduces a revised image formation model that accounts for the unique characteristics of underwater light propagation; this contributes to a more accurate estimation of range-dependent attenuation and backscatter.

While accurate underwater image recovery remains a challenge, the learning-based approaches for *perceptual enhancement* have made remarkable progress in recent years. Driven by large-scale supervised training [35, 68], these approaches learn sequences of non-linear filters to approximate the underlying pixel-to-pixel mapping [36] between the *distorted* and *enhanced* image domains. The contemporary deep CNN-based generative models provide SOTA performance in learning such image-to-image translation for both terrestrial [14, 11] and underwater domains [35, 48]. Moreover, the GAN-based models attempt to improve generalization performance by employing a two-player min-max game [26], where a *generator* learns realistic enhancement while evolving with an adversarial *discriminator* toward equilibrium. Several GAN-based underwater image enhancement models have reported impressive results from both paired [20, 45] and unpaired training [35]. However, they are prone to training instability, hence require careful hyper-parameter choices and intuitive loss function adaptation to ensure convergence [5, 53].

Single image super-resolution (SISR) problem deals with automatically generating a sharp HR image from its LR measurements. Although SISR is relatively less studied in the underwater domain, a rich body of literature exists for terrestrial imagery [66]. In particular, existing deep CNN-based models [18, 42] and GAN-based models [57, 59] provide good solutions for SISR. Researchers have also exploited contemporary techniques [39, 40, 61] such as gradient clipping, dense skip connection, and sub-pixel convolution to improve SISR performance on standard datasets. Moreover, deep residual networks [42, 30] and residual-in-residual networks [64, 47] are known to be very effective for learning SISR. Such networks employ skip connections to preserve the identity mapping within *repeated blocks* of convolutional layers; this contributes to a stable training of very deep models. Zhang *et al.* [72] further demonstrated that dense skip connections within a residual block allow combining of hierarchical features from each layer, which boosts the SISR performance.

In recent years, similar ideas have been effectively applied for underwater imagery as well. For instance, Chen *et al.* [13] adopt residual-in-residual learning for underwater SISR, whereas Islam *et al.* [34] introduce a deep residual *multiplier* model that can be dynamically configured for $2\times$, $4\times$, or $8\times$ SISR. Although these models report inspiring results, they do not account for underwater image distortions, and hence rely on a secondary network for enhancement. On the contrary, traditional approaches primarily focus on enhancing underwater image reconstruction quality by de-blurring/denoising [12, 56], or descattering [50]. Hence, their applicability for end-to-end SESR is limited.

Visual attention-based saliency prediction refers to finding *interesting* foreground regions in the image space [51, 63]. The classical *stimulus-driven* approaches use features such as luminance, color, texture, and often depth information to quantify feature contrast in a scene, which are subsequently exploited for spatial saliency computation. Automatic saliency

prediction over a sequence of frames is explored extensively as well [6]. Another genre of approaches deal with *goal-driven* saliency prediction for visual question answering [67], *i.e.*, finding the image regions that are relevant to a specific query. In the underwater domain, however, existing research work mainly focuses on salient feature extraction for enhancing object detection performance [19, 52, 70]. Hence, they do not provide a general solution for attention modeling that can facilitate faster visual search or better scene understanding. Nevertheless, finding salient RoIs in distorted underwater images and generating corresponding enhanced HR patches can be extremely useful for visually-guided robots.

III. PROBLEM FORMULATION

A. Learning SESR

SESR refers to the task of generating perceptually enhanced HR images from their LR and possibly distorted (LRD) input measurements. We formulate the problem as learning a pixel-to-pixel mapping from a source domain X (of LRD images) to its target domain Y (of enhanced HR images); we represent this mapping as a generative function $G : X \rightarrow Y$. We adopt an extended formulation by considering the task of learning SESR and saliency prediction on a shared feature space. Specifically, Deep SESR learns the generative function $G : X \rightarrow S, E, Y$; here, the additional outputs S and E denote the predicted saliency map, and enhanced image (in the same resolution as the input X), respectively. Additionally, it offers up to $4\times$ SESR for the final output Y.

B. Data Preparation: The UFO-120 Dataset

We utilize several existing underwater image enhancement and super-resolution datasets to supervise the SESR learning with paired data of the form $(\{X\}, \{S, E, Y\})$. We prepare over 1500 samples for training and another 120 for testing in the UFO-120 dataset. It contains HR natural underwater images collected from oceanic explorations in multiple locations having different water types, as seen in Fig. 2a. The saliency maps are annotated by human participants, whereas standard procedures for optical/spatial image degradation [35, 46, 20] are followed to create the respective LRD samples.

Specifically, we adopt a widely used domain-transfer technique [35, 20] that deploys a CycleGAN [73]-based model trained on unpaired natural data to generate distorted images by mimicking underwater optical distortion characteristics. Subsequently, we prepare the LRD samples by Gaussian blurring (GB) and bicubic down-sampling (BD). Based on their relative order, we group the data into three sets: **i) Set-U:** GB is followed by BD, **ii) Set-F:** the order is interchanged with a 0.5 probability, and **iii) Set-O:** BD is followed by GB; we use a 7×7 kernel and a noise level of 20% for the GB.

As Fig. 2b illustrates, we use $2\times$, $3\times$, and $4\times$ BD to generate the LRD samples from the synthetically distorted HR pairs. Hence, there are nine available training combinations for SESR. Note that the UFO-120 dataset can also be used for training underwater SISR ($E{\rightarrow}Y$), image enhancement ($X{\rightarrow}E$), or saliency prediction ($E{\rightarrow}S$) models.

(a) A few sample ground truth images and corresponding saliency maps are shown on the top, and bottom row, respectively.

(b) Two particular instances are shown: the HR ground truth images are of size 640×480; their corresponding LR distorted (LRD) images are of size 320×240, 214×160, and 160×120.

Fig. 2: The UFO-120 dataset facilitates paired training of $2\times$, $3\times$, and $4\times$ SESR models; it also contains salient pixel annotations for all training samples. The combined data is used for the supervised training of Deep SESR model.

IV. DEEP SESR MODEL

As shown in Figure 3, the major components of our Deep SESR model are: residual dense blocks (RDBs), a feature extraction network (FENet), and an auxiliary attention network (AAN). These components are tied to an end-to-end architecture for the combined SESR learning.

Residual Dense Blocks (RDBs) consist of three sets of convolutional (conv) layers, each followed by Batch Normalization (BN) [32] and ReLU non-linearity [54]. As Figure 3a illustrates, the input and output of each layer is concatenated to subsequent layers. This architecture is inspired by Zhang *et al.* [72] who demonstrated that such dense skip connections facilitate an improved hierarchical feature learning. Each conv layer learns 64 filters of a given kernel size; their outputs are then fused by a 1×1 conv layer for local residual learning.

Feature Extraction Network (FENet) uses RDBs as building blocks to incorporate two-stage residual-in-residual learning. As shown in Figure 3b, on the first stage, two parallel branches use eight RDB blocks each to separately learn 3×3 and 5×5 filters in input image space; these filters are then concatenated and passed to a common branch for the second stage of learning. Four RDB blocks with 3×3 filters are used in the later stage which eventually generates 32 feature maps. Our motive for such design is to have the capacity to learn locally dense informative features while still maintaining a globally shallow architecture to ensure fast feature extraction.

Auxiliary Attention Network (AAN) learns to model visual attention in the FENet-extracted feature space. As shown in Figure 3c, two sequential conv layers learn to generate a single channel output that represents saliency (probabilities) for each pixel. We show the predicted saliency map as *green* intensity values; the black pixels represent background regions.

The Deep SESR learning is guided along the primary branch by a series of conv and deconv (de-convolutional) layers. As Figure 3c demonstrates, the enhanced image (LR),

(a) A residual dense block (RDB) [72].

(b) The feature extraction network (FENet).
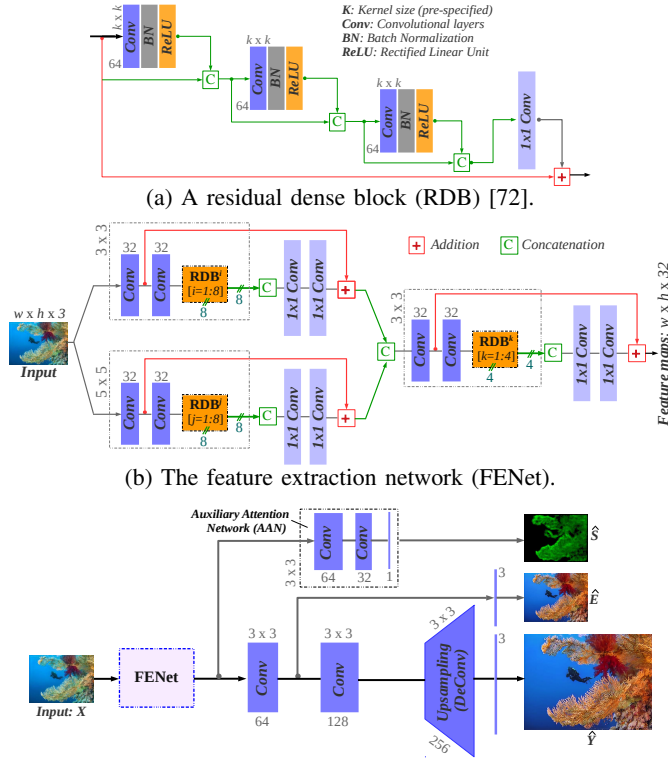
(c) The end-to-end architecture is shown. FENet-extracted feature maps are propagated along two branches: *i)* to AAN for learning saliency, and *ii)* to an intermediate convolutional layer for learning enhancement. Another convolutional layer and subsequent upsampling layers learn SESR along the main branch.

Fig. 3: Network architecture and detailed parameter specification of the proposed Deep SESR model.

and the SESR image (HR) are generated by separate output layers at different stages in the network. The enhanced image is generated from the conv layer that immediately follows FENet; it is supervised to learn enhancement by dedicated loss functions applied at the shallow output layer. The *enhanced* features are also propagated to another conv layer, followed by deconv layers for upsampling. The final SESR output is generated from upsampled features based on the given scale: $2\times$, $3\times$, or $4\times$. Other model parameters, *e.g.*, the number of filters, kernel sizes, etc., are annotated in Figure 3.

A. Loss Function Formulation

The end-to-end training of Deep SESR is supervised by seven loss components that address various aspects of learning the function $G : X \rightarrow S, E, Y$. By denoting $\hat{S}, \hat{E}, \hat{Y} = G(X)$ as the generated output, we formulate the loss terms as follows:

1) **Information Loss** for saliency prediction is measured by a standard cross-entropy function [51, 63]. It quantifies the dissimilarity in pixel intensity distributions between the generated saliency map ($\hat{S}$) and its ground truth (S). For a total of N_p pixels in $\hat{S}$, it is calculated as

$$\mathcal{L}^{AAN}_{Saliency} = \frac{1}{N_p} \sum_{p=1}^{N_p} \left[-S_p \log \hat{S}_p - (1 - S_p) \log(1 - \hat{S}_p) \right]. \quad (1)$$

2) **Contrast loss (LR)** evaluates the hue and luminance recovery in the enhanced images. The dominating green/blue hue in distorted underwater images often causes low-contrast and globally dim foreground pixels. We quantify this loss of relative strength (*i.e.*, intensity) in foreground pixels in RGB space by utilizing a differentiable function: Contrast Measurement Index (CMI) [58, 62]. The CMI measures the average intensity of foreground pixels (F_I) relative to the background (B_I) for an image I, as $CMI(I) = \frac{(F_I - B_I)}{(F_I + B_I)} \propto (F_I - B_I)$. We exploit the saliency map S (or $\hat{S}$) to find the foreground pixels in E (or $\hat{E}$), as $F_E = E \odot S$ and $F_{\hat{E}} = \hat{E} \odot \hat{S}$; here, $\odot$ denotes element-wise multiplication. Subsequently, we compute the contrast loss as

$$\mathcal{L}^{LR}_{Contrast} = \left|\left| CMI(E) - CMI(\hat{E}) \right|\right|_2. \quad (2)$$

An immediate consequence of using $\mathcal{L}^{LR}_{Contrast}$ is that AAN can directly influence learning enhancement despite being on a separate branch. Such coupling also provides better training stability (otherwise AAN tends to converge too early and starts over-fitting). Moreover, in Figure 4a, we show the distributions of CMI for training samples of the UFO-120 dataset, which suggests that the distorted samples' CMI scores are skewed to much lower values compared to the ground truth. Hence, $\mathcal{L}^{LR}_{Contrast}$ forces the CMI distribution to shift toward higher values for learning contrast enhancement.
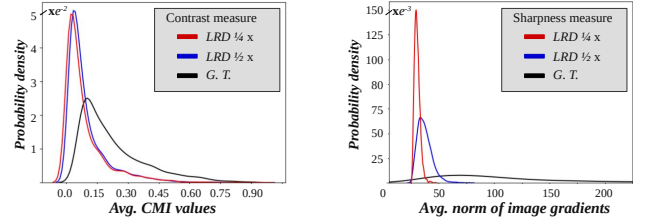
(a) Contrast measure: $CMI(I)$.   (b) Sharpness measure: $|\nabla I|$.
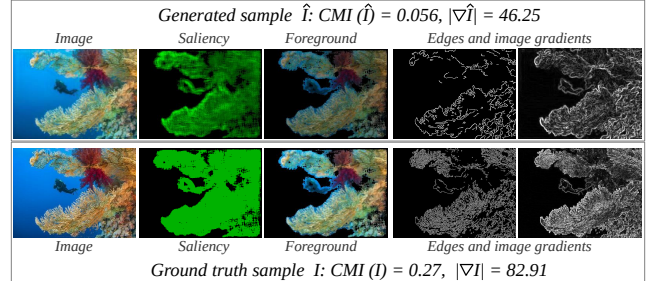
(c) Image contrast and sharpness properties of a particular sample compared to its ground truth measurement.

Fig. 4: The lack of contrast and sharpness in LRD samples of UFO-120 dataset (compared to their ground truth) are shown in (a) and (b); as seen, distributions for LRD samples are densely skewed to lower values, whereas the ground truth distributions span considerably higher values. A qualitative interpretation of this numeric disparity is illustrated in (c).

3) **Color Loss (LR/HR)** evaluates global similarity of the enhanced ($\hat{E}$) and SESR ($\hat{Y}$) output with respective ground truth measurements in RGB space. The standard $\mathcal{L}_2$ loss terms

are: $\mathcal{L}_2^{LR} = \left|\left|E - \hat{E}\right|\right|_2$, and $\mathcal{L}_2^{HR} = \left|\left|Y - \hat{Y}\right|\right|_2$. Additionally, we formulate two *perceptual loss* functions that are particularly designed for learning underwater image enhancement and super-resolution. First, we utilize two wavelength-dependent chrominance terms: $C_{rg} = (r - g)$, and $C_{yb} = \frac{1}{2}(r + g) - b$, which are core elements of the Underwater Image Colorfulness Measure (UICM) [55, 49]. By denoting Δr, Δg, and Δb, as the per-channel numeric differences between $\hat{E}$ and E, we formulate the loss as:

$$\mathcal{L}_P^{LR} = \left|\left|4(\Delta r - \Delta g)^2 + (\Delta r + \Delta g - 2\Delta b)^2\right|\right|_2. \quad (3)$$

On the other hand, being inspired by [17, 34], we evaluate the perceptual similarity at HR as

$$\mathcal{L}_P^{HR} = \left|\left|\frac{(512 + \bar{R})}{256}\Delta R^2 + 4\Delta G^2 + \frac{(767 - \bar{R})}{256}\Delta B^2\right|\right|_2. \quad (4)$$

Here, $\bar{R} = (R_Y + R_{\hat{Y}})/2$, whereas ΔR, ΔG, and ΔB are the per-channel disparities between $\hat{Y}$ and Y. Finally, we adopt the color loss terms for enhancement and SESR as

$$\mathcal{L}_{Color}^{LR} = 0.25\ \mathcal{L}_P^{LR} + 0.75\ \mathcal{L}_2^{LR}, \text{ and} \quad (5)$$

$$\mathcal{L}_{Color}^{HR} = 0.25\ \mathcal{L}_P^{HR} + 0.75\ \mathcal{L}_2^{HR}, \text{ respectively.} \quad (6)$$

4) **Content loss (LR/HR)** forces the generator to restore a similar *feature content* as the ground truth in terms of high-level representation. Such feature preservation has been found to be very effective for image enhancement, style transfer, and SISR problems [31, 34]; as suggested in [38], we define the image content function $\Phi_{VGG}(\cdot)$ as high-level features extracted by the last `conv` layer of a pre-trained VGG-19 network. Then, we formulate the content loss for enhancement and SESR as

$$\mathcal{L}_{Content}^{LR} = \left|\left|\Phi_{VGG}(E) - \Phi_{VGG}(\hat{E})\right|\right|_2, \text{ and} \quad (7)$$

$$\mathcal{L}_{Content}^{HR} = \left|\left|\Phi_{VGG}(Y) - \Phi_{VGG}(\hat{Y})\right|\right|_2, \text{ respectively.} \quad (8)$$

5) **Sharpness loss (HR)** measures the blurriness recovery in SESR output by exploiting local image gradients. The literature offers several solutions for evaluating image sharpness based on norm/histogram of gradients or frequency-domain analysis. In particular, the notions of Just Noticeable Blur (JNB) [22] and Perceptual Sharpness Index (PSI) [21] are widely used; they apply non-linear transformation and thresholding on local contrast or gradient-based features to quantify perceived blurriness based on the characteristics of human visual system. However, we found better results and numeric stability by using the norm of image gradients directly; specifically, we use the standard 3×3 Sobel operator [24] for computing spatial gradient $\nabla I = \sqrt{I_x^2 + I_y^2}$ for an image I. Subsequently, we formulate the sharpness loss for SESR as

$$\mathcal{L}_{Sharpness}^{HR} = \left|\left|\ |\nabla Y|^2 - |\nabla \hat{Y}|^2\ \right|\right|_1. \quad (9)$$

In Figure 4b, we present a statistical validity of $\mathcal{L}_{sharpness}^{LR}$ as a loss component; also, edge gradient features for a particular sample are provided in Figure 4c. As shown, numeric disparities for the norm of gradients between distorted images and their HR ground truth are significant, which we quantify by $\mathcal{L}_{sharpness}^{LR}$ to encourage sharper image generation.

B. End-to-end Training Objective

We use a linear combination of the above-mentioned loss components to formulate the unified objective function as

$$G^* = \arg\min_{G} \left\{\lambda_s^{AAN}\mathcal{L}_{Saliency}^{AAN} + \mathcal{L}_{SESR}^{LR} + \mathcal{L}_{SESR}^{HR}\right\}; \quad (10)$$

where $\mathcal{L}_{SESR}^{LR}$ and $\mathcal{L}_{SESR}^{HR}$ are expressed by

$\mathcal{L}_{SESR}^{LR} = \lambda_c^{LR}\mathcal{L}_{Color}^{LR} + \lambda_f^{LR}\mathcal{L}_{Content}^{LR} + \lambda_t^{LR}\mathcal{L}_{Contrast}^{LR}$, and
$\mathcal{L}_{SESR}^{HR} = \lambda_c^{HR}\mathcal{L}_{Color}^{HR} + \lambda_f^{HR}\mathcal{L}_{Content}^{HR} + \lambda_g^{HR}\mathcal{L}_{Sharpness}^{HR}$.

Here, $\lambda_{\odot}^{\square}$ symbols are scaling factors that represent the contributions of respective loss components; their values are empirically tuned as hyper-parameters.

V. EXPERIMENTAL RESULTS

A. Implementation Details

As mentioned in Section III-B, Deep SESR training is supervised by paired data of the form $(\{X\}, \{S, E, Y\})$. We use TensorFlow libraries [1] to implement the optimization pipeline (of Eq. 10); a Linux host with two Nvidia™ GTX 1080 graphics cards are used for training. Adam optimizer [41] is used for the global iterative learning with a rate of 10^{-4} and a momentum of 0.5; the network converges within 23-26 epochs of training in this setup (with a batch-size of 2). In the following sections, we present the experimental results based on qualitative analysis, quantitative evaluations, and ablation studies. Since there are no existing SESR methods, we compare the Deep SESR performance separately with SOTA image enhancement and super-resolution models. Note that, all models in comparison are trained on the same train-validation splits (of respective datasets) by following their recommended parameter settings. Also, for datasets other than UFO-120, the AAN (and $\mathcal{L}_{Contrast}^{LR}$) is not used by Deep SESR as their ground truth saliency maps are not available.

B. Evaluation: Enhancement

We first qualitatively analyze the Deep SESR-generated images in terms of color, contrast, and sharpness. As Fig. 5 illustrates, the enhanced images are perceptually similar to the respective ground truth. Specifically, the greenish underwater hue is rectified, true pixel colors are mostly restored, and the global image sharpness is recovered. Moreover, the generated saliency map suggests that it focused on the right foreground regions for contrast improvement. We further demonstrate the contributions of each loss-term: $\mathcal{L}_{Contrast}^{LR}$, $\mathcal{L}_P^{LR}$, $\mathcal{L}_{Color}^{LR}$, and $\mathcal{L}_{Content}^{LR}$ for learning the enhancement. We observe that the color rendition gets impaired without $\mathcal{L}_P^{LR}$ and $\mathcal{L}_{Color}^{LR}$, whereas, $\mathcal{L}_{Content}^{LR}$ contributes to learning finer texture details. We also notice a considerably low-contrast image generation without $\mathcal{L}_{Contrast}^{LR}$, which validates the utility of saliency-driven contrast evaluation via CMI (see Section IV-A).

Next, we compare the perceptual image enhancement performance of Deep SESR with the following models: (i) relative global histogram stretching (RGHS) [29], (ii) unsupervised color correction (UCM) [33], (iii) multi-scale fusion (MS-Fusion) [4], (iv) multi-scale Retinex (MS-Retinex) [71], (v)
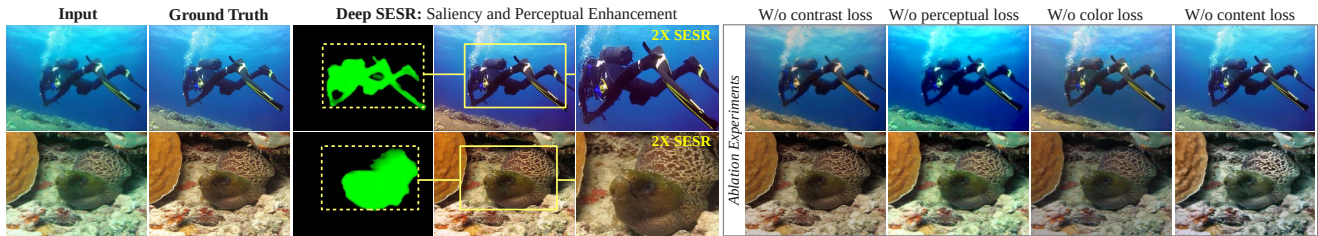
Fig. 5: Each row demonstrates perceptual enhancement and saliency prediction by Deep SESR on respective LRD input images; the corresponding results of an ablation experiment shows contributions of various loss-terms in the learning.

TABLE I: Quantitative performance comparison for enhancement: scores are shown as $mean \pm \sqrt{variance}$; the first and second best scores (in each row) are colored red, and blue, respectively.

|	Dataset	RGHS	UCM	MS-Fusion	MS-Retinex	Water-Net	UGAN	Fusion-GAN	FUnIE-GAN	**Deep SESR**
PSNR	UFO-120	20.05 ± 3.1	20.99 ± 2.2	21.32 ± 3.3	21.69 ± 3.6	22.46 ± 1.9	23.45 ± 3.1	24.07 ± 2.1	25.15 ± 2.3	27.15 ± 3.2
	EUVP	20.12 ± 2.9	20.55 ± 1.8	19.85 ± 2.4	21.27 ± 3.1	20.14 ± 2.3	23.67 ± 1.5	23.77 ± 2.4	26.78 ± 1.1	25.25 ± 2.1
	UImNet	19.98 ± 1.8	20.48 ± 2.2	19.59 ± 3.2	22.63 ± 2.5	21.02 ± 1.6	23.88 ± 2.1	23.12 ± 1.9	24.68 ± 2.4	25.52 ± 2.7
SSIM	UFO-120	0.75 ± 0.06	0.78 ± 0.07	0.79 ± 0.09	0.75 ± 0.10	0.79 ± 0.05	0.80 ± 0.08	0.82 ± 0.07	0.82 ± 0.08	0.84 ± 0.03
	EUVP	0.69 ± 0.11	0.73 ± 0.14	0.70 ± 0.05	0.69 ± 0.15	0.68 ± 0.18	0.67 ± 0.11	0.68 ± 0.05	0.86 ± 0.05	0.75 ± 0.07
	UImNet	0.61 ± 0.08	0.67 ± 0.06	0.64 ± 0.11	0.74 ± 0.04	0.71 ± 0.07	0.79 ± 0.08	0.75 ± 0.07	0.77 ± 0.06	0.81 ± 0.05
UIQM	UFO-120	2.36 ± 0.33	2.41 ± 0.53	2.76 ± 0.45	2.69 ± 0.59	2.83 ± 0.48	3.04 ± 0.28	2.98 ± 0.28	3.09 ± 0.51	3.13 ± 0.45
	EUVP	2.45 ± 0.46	2.48 ± 0.77	2.51 ± 0.36	2.48 ± 0.09	2.55 ± 0.06	2.70 ± 0.31	2.58 ± 0.07	2.95 ± 0.38	2.98 ± 0.28
	UImNet	2.32 ± 0.48	2.38 ± 0.42	2.79 ± 0.55	2.84 ± 0.37	2.92 ± 0.35	3.32 ± 0.55	3.19 ± 0.27	3.23 ± 0.32	3.26 ± 0.36
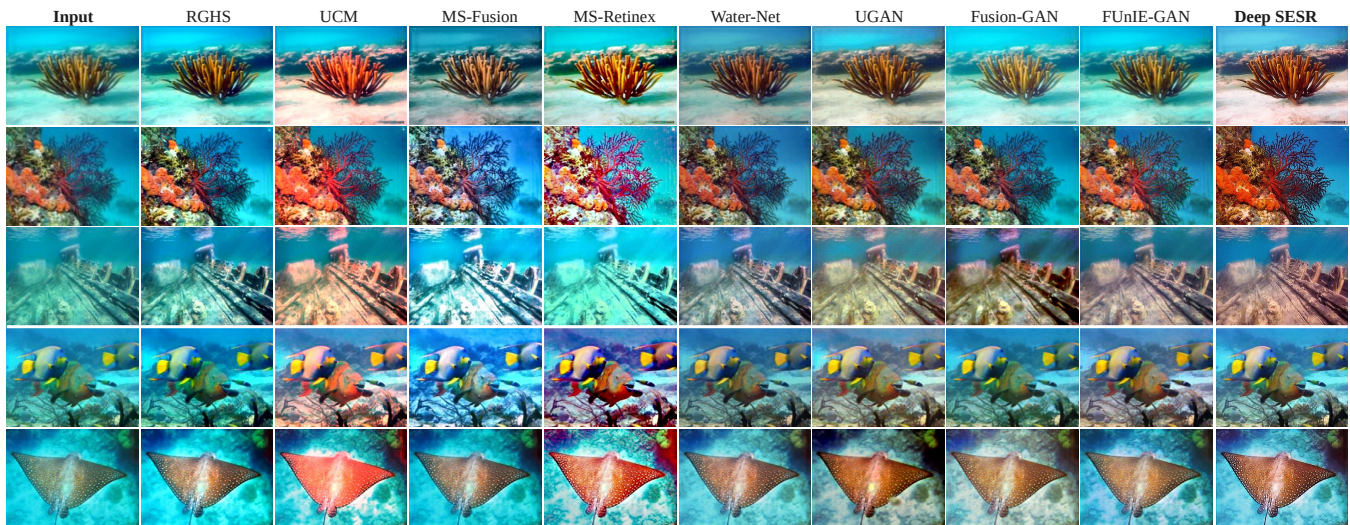
Fig. 6: Qualitative comparison of Deep SESR-enhanced images with SOTA models: RGHS [29], UCM [33], MS-Fusion [4], MS-Retinex [71], Water-Net [43], UGAN [20], Fusion-GAN [44], and FUnIE-GAN [35].

Water-Net [43], (vi) UGAN [20], (vii) Fusion-GAN [44], and (viii) FUnIE-GAN [35]. The first four are physics-based models and the rest are learning-based models; they provide SOTA performance for underwater image enhancement in RGB space (without requiring scene depth or optical waterbody measures). Their performance is quantitatively evaluated on common test sets of each dataset based on standard metrics [35, 49]: peak signal-to-noise ratio (PSNR) [28], structural similarity measure (SSIM) [65], and underwater image quality measure (UIQM) [55]. The PSNR and SSIM quantify reconstruction quality and structural similarity of generated images with respect to ground truth, whereas the UIQM evaluates image qualities based on colorfulness, sharpness, and contrast. The evaluation is summarized in Table I; moreover, a few qualitative comparisons are shown in Fig. 6.

As Fig. 6 demonstrates, UCM and MS-Retinex often suffer from over-saturation, whereas RGBH, MS-Fusion, and Water-Net fall short in hue rectification. In comparison, the color restoration and contrast enhancement of UGAN, Fusion-GAN, and FUnIE-GAN are generally better. In addition to achieving comparable color recovery and hue rectification, the Deep SESR-generated images are considerably sharper. Since the boost in performance is rather significant for UFO-120 dataset (suggested by the results of Table I), it is likely that the additional knowledge about foreground pixels through $\mathcal{L}^{LR}_{Contrast}$ helps in this regard. Deep SESR achieves competitive and often better performance in terms of PSNR and SSIM as well. In particular, it generally attains better UIQM scores; we postulate that $\mathcal{L}^{LR}_{P}$ contributes to this enhancement, as it is designed to improve the UICM (see Section IV-A). Further ablation investigations reveal a 9.47% drop in UIQM values without using $\mathcal{L}^{LR}_{P}$ in the learning objective.

TABLE II: Quantitative performance comparison for super-resolution: scores are shown as *mean* $\pm \sqrt{variance}$; the first and second best scores (in each column per-dataset) are colored red, and blue, respectively. ($\oslash$ Does not support $3\times$ scale)

|	Model	PSNR 2×	PSNR 3×	PSNR 4×	SSIM 2×	SSIM 3×	SSIM 4×	UIQM 2×	UIQM 3×	UIQM 4×
UFO-120	SRCNN	24.75 ± 3.7	22.22 ± 3.9	19.05 ± 2.3	.72 ± .07	.65 ± .09	.56 ± .12	2.39 ± 0.35	2.24 ± 0.17	2.02 ± 0.47
	SRResNet	25.23 ± 4.1	23.85 ± 2.8	19.13 ± 2.4	.74 ± .08	.68 ± .07	.56 ± .05	2.42 ± 0.37	2.18 ± 0.26	2.09 ± 0.30
	SRGAN	26.11 ± 3.9	23.87 ± 4.2	21.08 ± 2.3	.75 ± .06	.70 ± .05	.58 ± .09	2.44 ± 0.28	2.39 ± 0.25	2.26 ± 0.17
	RSRGAN	25.25 ± 4.3	23.15 ± 4.1	20.25 ± 2.4	.79 ± .08	.71 ± .08	.58 ± .04	2.41 ± 0.29	2.38 ± 0.31	2.27 ± 0.22
	SRDRM	26.23 ± 4.4	$\oslash$	22.26 ± 2.5	.79 ± .09	$\oslash$	.59 ± .05	2.45 ± 0.43	$\oslash$	2.28 ± 0.35
	SRDRM-GAN	26.26 ± 4.3	$\oslash$	22.21 ± 2.4	.78 ± .08	$\oslash$	.58 ± .13	2.42 ± 0.30	$\oslash$	2.27 ± 0.44
	Deep SESR	28.57 ± 3.5	26.86 ± 4.1	24.75 ± 2.8	.85 ± .09	.75 ± .06	.66 ± .05	3.09 ± 0.41	2.87 ± 0.39	2.55 ± 0.35
USR-248	SRCNN	24.88 ± 4.4	24.01 ± 3.5	23.75 ± 3.2	.73 ± .08	.70 ± .10	.69 ± .12	2.38 ± 0.38	2.31 ± 0.29	2.21 ± 0.68
	SRResNet	24.96 ± 3.7	23.39 ± 5.2	22.21 ± 3.6	.74 ± .07	.71 ± .11	.70 ± .08	2.42 ± 0.48	2.33 ± 0.58	2.27 ± 0.70
	SRGAN	25.76 ± 3.5	25.02 ± 3.9	24.36 ± 4.3	.77 ± .06	.75 ± .05	.69 ± .13	2.53 ± 0.42	2.65 ± 0.44	2.75 ± 0.66
	RSRGAN	25.11 ± 2.9	24.96 ± 4.7	24.15 ± 2.9	.75 ± .06	.72 ± .09	.71 ± .09	2.42 ± 0.35	2.49 ± 0.56	2.55 ± 0.47
	SRDRM	26.16 ± 3.5	$\oslash$	24.96 ± 3.3	.77 ± .10	$\oslash$	.72 ± .11	2.47 ± 0.69	$\oslash$	2.35 ± 0.51
	SRDRM-GAN	26.77 ± 4.1	$\oslash$	24.77 ± 3.4	.82 ± .07	$\oslash$	.70 ± .12	2.87 ± 0.55	$\oslash$	2.81 ± 0.56
	Deep SESR	27.03 ± 2.9	25.92 ± 3.5	24.59 ± 3.8	.88 ± .05	.76 ± .05	.71 ± .08	3.15 ± 0.44	3.04 ± 0.37	2.96 ± 0.28
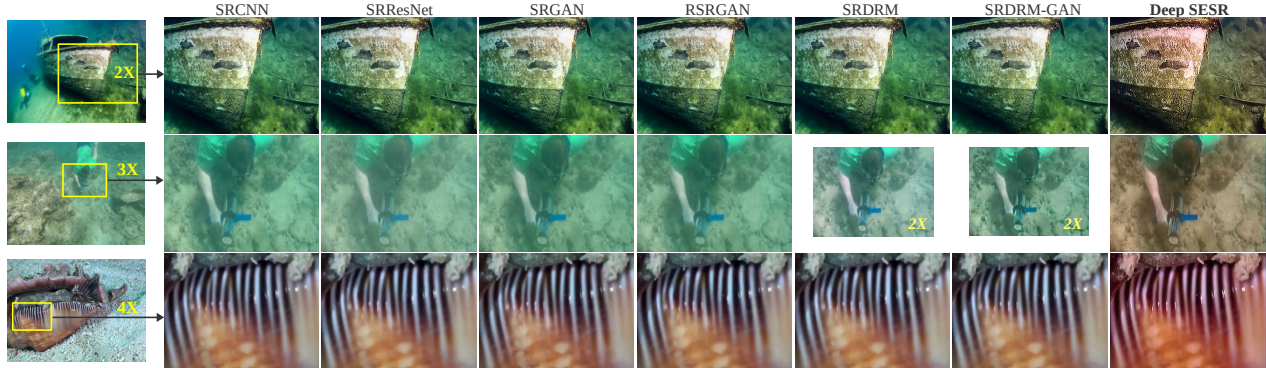
Fig. 7: Qualitative comparison for SISR performance of Deep SESR with existing solutions and SOTA models: SRCNN [18], SRResNet [42], SRGAN [42], RSRGAN [13], SRDRM [34], and SRDRM-GAN [34].

C. Evaluation: Super-Resolution

We follow similar experimental procedures for evaluating the super-resolution performance of Deep SESR. We consider the existing underwater SISR models named RSRGAN [13], SRDRM [34], and SRDRM-GAN [34] for performance comparison. We also include the standard (terrestrial) SISR models named SRCNN [18], SRResNet [42], and SRGAN [42] in the evaluation as benchmarks. We compare their $2\times$, $3\times$, and $4\times$ SISR performance on two large-scale datasets: UFO-120, and USR-248. The results are presented in Table II, and a few samples are shown in Fig. 7. Note that, the test images of USR-248 dataset are left undistorted for a fair comparison.

TABLE III: Deep SESR performance on the UFO-120 test dataset; set-wise mean scores are shown for $2\times/3\times/4\times$ SESR.

	PSNR	SSIM	UIQM
Set-U	28.55/26.77/24.25	0.86/0.75/0.66	3.07/2.89/2.54
Set-F	27.93/26.33/24.87	0.85/0.73/0.63	3.10/2.84/2.52
Set-O	28.95/27.15/25.45	0.84/0.79/0.68	3.09/2.86/2.58

As Table II demonstrates, Deep SESR outperforms other models in comparison by considerable margins on UIQM. This is due to the fact that it enhances perceptual image qualities in addition to spatial resolution. As shown in Fig. 7, Deep SESR generates much sharper and better quality HR images from both distorted and undistorted LR input patches, which contributes to its competitive PSNR and SSIM scores on the


Fig. 8: Color and texture recovery of Deep SESR: comparison shown with two best-performing SISR models (as of Table II).

USR-248 dataset. Fig. 8 further demonstrates that it does not introduce noise by unnecessary over-correction, which is a prevalent limitation of existing solutions. Lastly, we observe similar performance trends for all three types of spatial down-sampling, *i.e.*, for Set-U, Set-F, and Set-O (see Section III-B); we present the relative quantitative scores in Table III.

VI. GENERALIZATION PERFORMANCE

Due to the ill-posed nature of modeling underwater image distortions without scene-depth and optical waterbody measurements, learning-based solutions often fail to generalize beyond supervised data. In addition to the already-presented results, we demonstrate the color and texture recovery of Deep SESR on unseen natural images in Fig. 9. As seen in Fig. 9a, Deep SESR-enhanced pixel intensities are perceptually similar to a comprehensive physics-based approximation [7]. Addi-
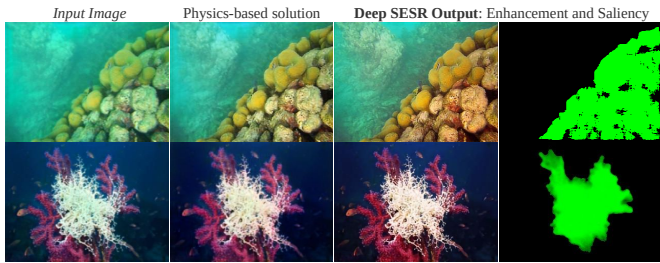
(a) Comparison with a physics-based color restoration method [7] that uses spectral waterbody measures and haze-lines prior.
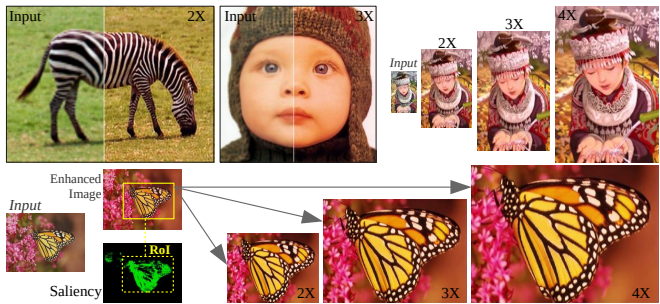
(b) Performance for $2\times$, $3\times$, and $4\times$ SESR on terrestrial images.

Fig. 9: Demonstration of generalization performance of Deep SESR model (trained on UFO-120 dataset).

tionally, it generates the respective HR images and saliency maps, and still offers more than 10 times faster run-time.

Deep SESR also provides reasonable performance on terrestrial images. As demonstrated in Fig. 9b, the color and texture enhancement of unseen objects (*e.g.*, grass, face, clothing, etc.) are perceptually coherent. Moreover, as Table IV indicates, its performance in terms of sharpness and contrast recovery for $2\times$, $3\times$, and $4\times$ SISR are competitive with SOTA benchmark results [46, 72]. Note that, much-improved performance can be achieved by further tuning and training on terrestrial datasets. Nevertheless, these results validate that the proposed architecture has the capacity to learn a generalizable solution of the underlying SESR problem.

TABLE IV: Deep SESR performance on terrestrial test data; blue (and boldfaced) scores represent 3% (and 1%) margins with SOTA benchmark results for $2\times/3\times/4\times$ SISR [46, 72].

	PSNR	SSIM
Set5 [8]	29.87 / **28.77** / 26.14	0.925 / **0.908** / 0.855
Set14 [69]	28.78 / 27.34 / 26.89	**0.914** / 0.801 / 0.756
Sun80 [60]	25.73 / 23.18 / 21.05	**0.802** / 0.755 / 0.704

VII. Operational Feasibility & Design Choices

Deep SESR's on-board memory requirement is only 10 MB, and it offers a run-time of 129 milliseconds (ms) per-frame, *i.e.*, 7.75 frames-per-second (FPS) on a single-board computer: Nvidia^{TM}AGX Xavier. As shown in Table V, it provides much faster speeds for the following design choices:

1) **Learning $\hat{E}$ and $\hat{S}$ on separate branches** facilitates a faster run-time when HR perception is not required. Specifically, we can decouple the $X \to S, E$ branches from the frozen model, which operates at 10.02 FPS (22% faster) to perform enhancement and saliency prediction. As shown in Fig. 10,

the predicted saliency map can be exploited for automatic RoI selection by using density gradient estimation techniques such as mean-shift [16]. The SESR output corresponding to the RoI can be generated with an additional 25 ms of processing time.

TABLE V: Run-time comparison for various design choices of Deep SESR (on Nvidia^{TM}AGX Xavier).

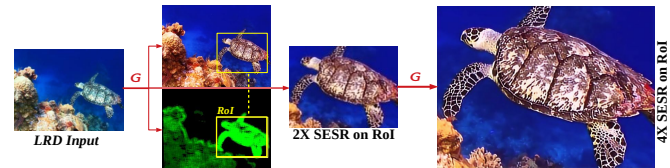|	$X \to S, E$	$X \to S, E, Y$
With FENet-1d	87.3 ms (11.45 FPS)	113 ms (8.85 FPS)
With FENet-2d	99.8 ms (10.02 FPS)	129 ms (7.75 FPS)

Fig. 10: Demonstration of automatic RoI selection based on local intensity values in the saliency map; Deep SESR can be applied again on the enhanced RoI for a detailed perception.

2) **FENet-1d and FENet-2d** are two design choices for the FENet (see Fig. 3b); FENet-2d is the default architecture that learns 3×3 and 5×5 filters in two parallel branches, whereas, FENet-1d refers to using a single branch of 3×3 filters. As shown in Table V, faster feature extraction by FENet-1d facilitates a 12.5% speed-up for Deep SESR. However, we observe a slight drop in performance, *e.g.*, $1.8\%/1.5\%/1.8\%$ lower scores for PSNR/SSIM/UIQM on UFO-120 dataset. Nevertheless, the generated images are qualitatively indistinguishable and the trade-off is admissible in practical applications.

Overall, Deep SESR offers use-case-specific design choices and ensures computational efficiency with robust SESR performance. These features make it suitable for near real-time robotic deployments; further demonstration is provided in the supplementary material (see https://youtu.be/wEkTu2CPW-g).

VIII. Conclusion

In this paper, we introduce the problem of simultaneous enhancement and super-resolution (SESR) and present an efficient learning-based solution for underwater imagery. The proposed generative model, named Deep SESR, can learn $2\times-4\times$ SESR and saliency prediction on a shared feature space. We also present its detailed network architecture, associated loss functions, and end-to-end training pipeline. Additionally, we contribute over 1500 annotated samples to facilitate large-scale SESR training on the UFO-120 dataset. We perform a series of qualitative and quantitative experiments, which suggest that Deep SESR: *i)* provides SOTA performance on underwater image enhancement and super-resolution, *ii)* exhibits significantly better generalization performance on natural images than existing solutions, *iii)* provides competitive results on terrestrial images, and *iv)* achieves fast inference on single-board platforms. The inspiring performance, computational efficiency, and availability of application-specific design choices make Deep SESR suitable for near real-time use by visually-guided underwater robots. In the future, we seek to incorporate $6\times-8\times$ spatial upscaling capability into the model with reasonable performance trade-offs.

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, et al. TensorFlow: A System for Large-scale Machine Learning. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 265–283, 2016.

[2] D. Akkaynak and T. Treibitz. A Revised Underwater Image Formation Model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6723–6732, 2018.

[3] D. Akkaynak and T. Treibitz. Sea-Thru: A Method for Removing Water From Underwater Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1682–1691, 2019.

[4] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert. Enhancing Underwater Images and Videos by Fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 81–88, 2012.

[5] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

[6] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent Mixture Density Network for Spatiotemporal Visual Attention. In *International Conference on Learning Representations (ICLR)*, 2017.

[7] D. Berman, D. Levy, S. Avidan, and T. Treibitz. Underwater Single Image Color Restoration using Haze-Lines and a New Quantitative Dataset. *arXiv preprint arXiv:1811.01343*, 2018.

[8] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel. Low-complexity Single-image super-resolution based on Nonnegative Neighbor Embedding. In *British Machine Vision Conference*, pages 135.1–135.10, 2012.

[9] B. Bingham, B. Foley, H. Singh, R. Camilli, K. Delaporta, R. Eustice, et al. Robotic Tools for Deep Water Archaeology: Surveying an Ancient Shipwreck with an Autonomous Underwater Vehicle. *Journal of Field Robotics (JFR)*, 27(6):702–717, 2010.

[10] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. B. Williams. True Color Correction of Autonomous Underwater Vehicle Imagery. *Journal of Field Robotics (JFR)*, 33(6):853–874, 2016.

[11] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. DehazeNet: An End-to-end System for Single Image Haze Removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.

[12] Y. Chen, B. Yang, M. Xia, W. Li, K. Yang, and X. Zhang. Model-based Super-resolution Reconstruction Techniques for Underwater Imaging. In *Photonics and Optoelectronics Meetings (POEM): Optoelectronic Sensing and Imaging*, volume 8332, page 83320G. International Society for Optics and Photonics, 2012.

[13] Y. Chen, J. Sun, W. Jiao, and G. Zhong. Recovering Super-Resolution Generative Adversarial Network for Underwater Images. In *International Conference on Neural Information Processing*, pages 75–83. Springer, 2019.

[14] Z. Cheng, Q. Yang, and B. Sheng. Deep Colorization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 415–423, 2015.

[15] Y. Cho, J. Jeong, and A. Kim. Model-assisted Multiband Fusion for Single Image Enhancement and Applications to Robot Vision. *IEEE Robotics and Automation Letters (RA-L)*, 3(4):2822–2829, 2018.

[16] D. Comaniciu and P. Meer. Mean Shift Analysis and Applications. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1197–1203. IEEE, 1999.

[17] CompuPhase. Perceptual Color Metric. https://www.compuphase.com/cmetric.htm, 2019. Accessed: 12-12-2019.

[18] C. Dong, C. C. Loy, K. He, and X. Tang. Image Super-resolution using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015.

[19] D. R. Edgington, K. A. Salamy, M. Risi, RE. Sherlock, D. Walther, and C. Koch. Automated Event Detection in Underwater Video. In *Oceans*, volume 5, pages 2749–2753. IEEE, 2003.

[20] C. Fabbri, M. J. Islam, and J. Sattar. Enhancing Underwater Imagery using Generative Adversarial Networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 7159–7165. IEEE, 2018.

[21] C. Feichtenhofer, H. Fassold, and P. Schallauer. A Perceptual Image Sharpness Metric based on Local Edge Gradient Analysis. *IEEE Signal Processing Letters*, 20 (4):379–382, 2013.

[22] R. Ferzli and L. J. Karam. A No-reference Objective Image Sharpness Metric based on the Notion of Just Noticeable Blur (JNB). *IEEE Transactions on Image Processing*, 18(4):717–728, 2009.

[23] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X. Zhang, and X. Ding. A Retinex-based Enhancing Approach for Single Underwater Image. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4572–4576. IEEE, 2014.

[24] W. Gao, X. Zhang, L. Yang, and H. Liu. An Improved Sobel Edge Detection. In *International Conference on Computer Science and Information Technology*, volume 5, pages 67–71. IEEE, 2010.

[25] Y. Girdhar, P. Giguere, and G. Dudek. Autonomous Adaptive Exploration using Realtime Online Spatiotemporal Topic Modeling. *International Journal of Robotics Research (IJRR)*, 33(4):645–657, 2014.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[27] K. He, J. Sun, and X. Tang. Single Image Haze Removal using Dark Channel Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2010.

[28] A. Hore and D. Ziou. Image Quality Metrics: PSNR vs. SSIM. In *International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.

[29] D. Huang, Y. Wang, W. Song, J. Sequeira, and S. Mavromatis. Shallow-water Image Enhancement using Relative Global Histogram Stretching Based on Adaptive Parameter Acquisition. In *International Conference on Multimedia Modeling*, pages 453–465. Springer, 2018.

[30] Z. Hui, X. Wang, and X. Gao. Fast and Accurate Single Image Super-resolution via Information Distillation Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 723–731, 2018.

[31] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. DSLR-quality Photos on Mobile Devices with Deep Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3277–3285, 2017.

[32] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, volume 37, pages 448–456, 2015.

[33] K. Iqbal, M. Odetayo, A. James, R. A. Salam, and A. Z. H. Talib. Enhancing the Low Quality Images using Unsupervised Colour Correction Method. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 1703–1709. IEEE, 2010.

[34] M. J. Islam, S. S. Enan, P. Luo, and J. Sattar. Underwater Image Super-Resolution using Deep Residual Multipliers. *arXiv preprint arXiv:1909.09437*, 2019.

[35] M. J. Islam, Y. Xia, and J. Sattar. Fast Underwater Image Enhancement for Improved Visual Perception. *arXiv preprint arXiv:1903.09766*, 2019.

[36] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[37] D. J. Jobson, Z. Rahman, and G. A. Woodell. A Multiscale Retinex for Bridging the Gap between Color Images and the Human Observation of Scenes. *IEEE Transactions on Image processing*, 6(7):965–976, 1997.

[38] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual Losses for Real-time Style Transfer and Super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016.

[39] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate Image Super-resolution using Very Deep Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[40] J. Kim, J. Kwon Lee, and K. Mu Lee. Deeply-recursive Convolutional Network for Image Super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016.

[41] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations (ICLR)*, 2015.

[42] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic Single Image Super-resolution using a Generative Adversarial Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[43] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. An Underwater Image Enhancement Benchmark Dataset and Beyond. In *IEEE Transactions on Image Processing (TIP)*, pages 1–1. IEEE, 2019.

[44] H. Li, J. Li, and W. Wang. A Fusion Adversarial Underwater Image Enhancement Network with a Public Test Dataset. *arXiv preprint arXiv:1906.06819*, 2019.

[45] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson. WaterGAN: Unsupervised Generative Network to Enable Real-time Color Correction of Monocular Underwater Images. *IEEE Robotics and Automation Letters (RA-L)*, 3(1):387–394, 2018.

[46] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. Feedback Network for Image Super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, 2019.

[47] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced Deep Residual Networks for Single Image Super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 136–144, 2017.

[48] P. Liu, G. Wang, H. Qi, C. Zhang, H. Zheng, and Z. Yu. Underwater Image Enhancement With a Deep Residual Framework. *IEEE Access*, 7:94614–94629, 2019.

[49] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. ISSN 1558-2205. doi: 10.1109/TCSVT.2019.2963772.

[50] H. Lu, Y. Li, S. Nakashima, H. Kim, and S. Serikawa. Underwater Image Super-resolution by Descattering and Fusion. *IEEE Access*, 5:670–679, 2017.

[51] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical Question-image co-attention for Visual question Answering. In *Advances in Neural Information Processing Systems*, pages 289–297, 2016.

[52] A. Maldonado-Ramírez and L. A. Torres-Méndez. Robotic Visual Tracking of Relevant Cues in Underwater Environments with Poor Visibility Conditions. *Journal of Sensors*, 2016.

[53] X. Mao, Q. Li, H. Xie, R. YK. Lau, Z. Wang, and S. Paul Smolley. Least Squares Generative Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017.

[54] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference on Machine Learning (ICML)*, pages 807–814,

2010.

[55] K. Panetta, C. Gao, and S. Agaian. Human-visual-system-inspired Underwater Image Quality Measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2016.

[56] E. Quevedo, E. Delory, GM. Callicó, F. Tobajas, and R. Sarmiento. Underwater Video Enhancement using Multi-camera Super-resolution. *Optics Communications*, 404:94–102, 2017.

[57] M. S. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single Image Super-resolution through Automated Texture Synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4491–4500, 2017.

[58] A. Shaus, S. Faigenbaum-Golovin, B. Sober, and E. Turkel. Potential Contrast–A New Image Quality Measure. *Electronic Imaging*, 2017(12):52–58, 2017.

[59] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised Map Inference for Image Super-resolution. In *International Conference on Learning Representations (ICLR)*, 2017.

[60] L. Sun and J. Hays. Super-resolution from Internet-scale Scene Matching. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2012.

[61] T. Tong, G. Li, X. Liu, and Q. Gao. Image Super-resolution using Dense Skip Connections. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4807, 2017.

[62] M. Trivedi, A. Jaiswal, and V. Bhateja. A Novel HVS Based Image Contrast Measurement Index. In *International Conference on Signal and Image Processing 2012 (ICSIP 2012)*, pages 545–555. Springer, 2013.

[63] W. Wang and J. Shen. Deep Visual Attention Prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

[64] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced Super-resolution Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[65] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P Simoncelli, et al. Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[66] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, and Q. Liao. Deep learning for Single Image Super-resolution: A Brief Review. *IEEE Transactions on Multimedia*, 2019.

[67] D. Yu, J. Fu, T. Mei, and Y. Rui. Multi-level Attention Networks for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4709–4717, 2017.

[68] X. Yu, Y. Qu, and M. Hong. Underwater-GAN: Underwater Image Restoration via Conditional Generative Adversarial Network. In *International Conference on Pattern Recognition(ICPR)*, pages 66–75. Springer, 2018.

[69] R. Zeyde, M. Elad, and M. Protter. On Single Image Scale-up using Sparse-Representations. In *International Conference on Curves and Surfaces*, pages 711–730. Springer, 2010.

[70] L. Zhang, B. He, Y. Song, and T. Yan. Underwater Image Feature Extraction and Matching based on Visual Saliency Detection. In *OCEANS*, pages 1–4. IEEE, 2016.

[71] S. Zhang, T. Wang, J. Dong, and H. Yu. Underwater Image Enhancement via Extended Multi-scale Retinex. *Neurocomputing*, 245:1–9, 2017.

[72] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual Dense Network for Image Super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.

[73] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-image Translation using Cycle-consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.