AutoEnhancer: Transformer on U-Net Architecture search for Underwater Image Enhancement

Yi Tang[1], Takafumi Iwaguchi[1], Hiroshi Kawasaki[1], Ryusuke Sagawa[2], and Ryo Furukawa[3]

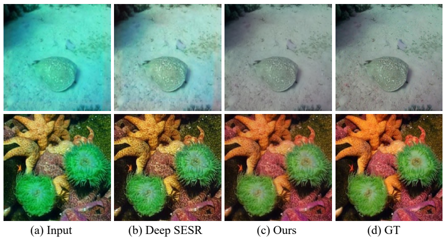
[1] Kyushu University
tang.yi.727@m.kyushu-u.ac.jp, {iwaguchi,kawasaki}@ait.kyushu-u.ac.jp
[2] National Institute of Advanced Industrial Science and Technology
ryusuke.sagawa@aist.go.jp
[3] Kindai University
furukawa@hiro.kindai.ac.jp

Abstract. Deep neural architecture plays an important role in underwater image enhancement in recent years. Although most approaches have successfully introduced different structures (e.g., U-Net, generative adversarial network (GAN) and attention mechanisms) and designed individual neural networks for this task, these networks usually rely on the designer's knowledge, experience and intensive trials for validation. In this paper, we employ Neural Architecture Search (NAS) to automatically search the optimal U-Net architecture for underwater image enhancement, so that we can easily obtain an effective and lightweight deep network. Besides, to enhance the representation capability of the neural network, we propose a new search space including diverse operators, which is not limited to common operators, such as convolution or identity, but also transformers in our search space. Further, we apply the NAS mechanism to the transformer and propose a selectable transformer structure. In our transformer, the multi-head self-attention module is regarded as an optional unit and different self-attention modules can be used to replace the original one, thus deriving different transformer structures. This modification is able to further expand the search space and boost the learning capability of the deep model. The experiments on widely used underwater datasets are conducted to show the effectiveness of the proposed method. The code is released at https://github.com/piggy2009/autoEnhancer.

Keywords: Underwater image enhancement · Neural architecture search · Transformer.

1 Introduction

Image enhancement which aims to improve the quality and recover the original information content by giving a low-quality image is a fundamental technique for image/video processing, such as video tracking [41], object recognition [19], and so on. Recently, image enhancement technologies are usually deployed in autonomous vehicles to assist the driving at night and in extreme weather [25]



(a) Input (b) Deep SESR (c) Ours (d) GT

Fig. 1. Illustration of different enhancement methods. (a) Input images. (b) results from Deep SESR enhancer [18]. (c) Our enhanced images. (d) Ground truths.

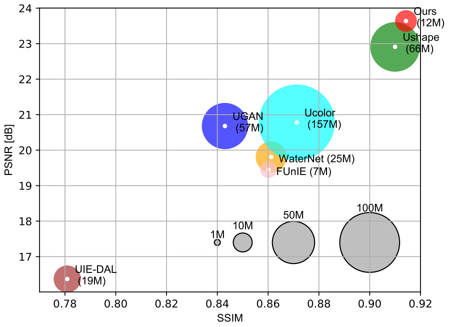


Fig. 2. Comparisons with state-of-the-art enhancers in terms of PSNR metric (Y axis), SSIM metric (X axis) and parameters (circular area) on UIEB dataset. Ours is competitive against WaterNet [28], FUnIE [19], UGAN [9], UIE-DAL [40], Ucolor [27] and Ushape [35].

or remotely operated underwater vehicle (ROV) to explore marine life and protect ecosystem [23]. Accordingly, underwater image enhancement (UIE) has recently become an interesting and meaningful research topic in computer vision tasks. However, UIE is still very challenging, because the underwater scenario is complex and diverse. For example, due to the different underwater depths, the collected images are suffering from different visual qualities, especially image brightness, which is significantly decreased with the depth. Additionally, the objects in underwater scenes are diverse. The stones, animals and plants present different colors or textures by the strong absorption and scattering, which also increases the difficulty for the enhancement algorithms to recover their original appearance from underwater scenarios.

Early approaches [6,8,26] mainly rely on a physical model, called Retinex theory [24], whose purpose is to estimate an accurate medium transmission. Then, the quality degradation values can be deduced by the medium transmission. These methods are able to improve the quality of some images in simple scenes like the shallow sea and weak ambient light. However, facing complex cases, such as turbid water, and extremely dark, these physical model-based methods always fail to estimate medium transmission, even these basic physical models are not always correct in some complicated scenes.

To enhance the image quality of severe conditions, many deep learning-based methods have been proposed [28,9,18] in the underwater image enhancement. Since it is difficult to prepare sufficient labeled data in underwater scenes, weakly supervised strategy [29] or generative adversarial network (GAN) [19,20] is used at the early stage. However, these deep models are still difficult to recover the objects' color in some complex scenes. As shown in Figure. 1, Deep SESR, which is a GAN-based method [18], cannot accurately restore the color of marine organisms. Recently, as U-Net architecture is able to effectively encode multi-level features for clear image reconstruction and easy implementation, most of the existing methods adopt this structure as a base network, and then design some specific modules for UIE. For example, the tailored attention module [27] and transformer module [35] are applied to the U-Net architecture. These modules are effective to boost performance, but their specific structures mainly rely on designers' experience and the heavy computational cost of repeated trials for validation. Additionally, in order to achieve competitive performance, the tailored modules are more and more complicated and significantly increase the parameters of models. The performance and model size of recent methods are shown in Figure. 2.

To balance the model performance and scale of parameters, our key insight is to leverage the strategy of neural architecture search (NAS) to automatically design an optimal U-Net framework for image enhancement instead of heavy structure validation experiments. First, we propose a new search space for our enhancement network. Different from [44,48] our search space is not limited to the common and lightweight operators, such as convolution, dilated convolution, etc., but also includes transformer module. Second, we propose a selectable transformer module, whose original multi-head self-attention is regarded as an optional unit, so that we can apply the NAS strategy to automatically search for an optimal self-attention module (e.g., shuffle attention [49], efficient channel attention [37]) and then to further improve the feature representation capability of the proposed network. In order to decrease the scale of parameters and use arbitrary input size, we apply convolution rather than a fully-connected operator to encode features in our selectable transformer module. Third, to allow our network to learn more color information, we introduce the images from different color spaces (i.e., RGB and Lab) as the network inputs, so that more robust color features can be extracted to improve the quality of images in the severe conditions. In the end, the contributions of our method are summarized as follows:

- We introduce a practical solution to apply Neural Architecture Search (NAS) to automatically build an end-to-end U-Net deep network for underwater image enhancement, especially for severely color degraded images.
- We present a new search space, where we are not limited to applying lightweight operators and further propose a selectable transformer module. This module grants the neural network substantial learning capability by automatically selecting suitable self-attention operators in the proposed network.
- The proposed architecture is able to encode the features from different color spaces to improve the adaptation and generalization. Besides, the comprehensive experiments prove that the proposed approach achieves competitive performance in different scenarios with the less parameters.

2 Related works

2.1 Image enhancement

The development of image enhancement can be briefly divided into two phases. In the first phase, most of the approaches exploit physical models (e.g., Retinex model [11]) to enhance image quality. For example, Ancuti et al. [2] propose a fusion-based model, which first tries to obtain the color-corrected and contrast-enhanced versions of an underwater image and then compute the corresponding weight maps to generate the final fusion result. Ancuti et al. modify the fusion model [2] and further propose a multi-scale fusion strategy for underwater image enhancement in [1]. Additionally, dynamic pixel range stretching [17], pixel distribution adjustment [12] and color correction algorithm [2] are used in underwater scenario. In [36], blurriness prior is proposed to improve the image quality but fails to recover the original color of underwater objects. These approaches can improve the quality of images to some extent, but their robustness is weak when dealing with difficult scenes. Moreover, most of these methods are suffering from heavy computation, thus affecting the efficiency of their models.

With the wide deployment of deep learning models, the community of image enhancement enters the second phase. Especially, after the proposal of fully convolutional networks (FCN) and U-Net structure, more are more efficient deep learning-based methods [34,3,42,46,21] are introduced into this community. For example, WaterGAN [30] proposes to combine GAN and U-Net to solve the problem of underwater image enhancement. Meanwhile, Li et al. [29] also propose a GAN-based weakly supervised deep models for this task. After that, Yang et al. [45] further uses conditional GAN (cGAN) to improve the image quality. GANs are widely applied to this task because there are few labeled datasets in the underwater image enhancement. It is very hard to simulate a similar underwater scenario to collect the low-quality images as inputs and its corresponding high-quality images as ground truths. Therefore, GAN is usually used with weakly supervised training strategies for deep network training. After that, WaterNet [28] not only proposes a gated fusion U-Net-based model but also collects a dataset called UIEB in this community. Moreover, their method for data collecting is novel. They try to use different enhancement methods to improve the quality of underwater images. Then, some volunteers will pick the best enhanced result as the ground truth. Following this pipeline, Ushape [35] collects a large underwater dataset, which contains 5004 images with labels. Moreover, Ushape presents a new U-Net, which includes two different transformer modules to learn robust image color and space information.

In summary, U-Net architecture is widely used in underwater image enhancement. The reasons are two-fold: 1) It can effectively extract color, content and texture features of underwater images. Then, these features are very useful to remove the noise and reconstruct a clear image by using the end-to-end U-Net architecture. 2) U-Net structure is easy to implement and extend. Generally, new modules can be directly inserted into this architecture. However, due to the diverse and complex underwater scenes, recent methods have to design more complex modules in U-Net. They are effective but increase the complexity of the models as well. Different from the existing methods that focus on designing a specific deep network, we introduce NAS to automatically obtain the optimal network for the underwater image enhancement.

2.2 Neural architecture search

The purpose of NAS is to automatically design neural networks. Early methods often use reinforcement learning [50] or evolution algorithms [43]. However, these methods require massive computing resources and cost much time during the searching. To alleviate this burden, Liu et al. [33] proposed DARTS. This method assigns weights to different operators, and then the gradient descent algorithm (SGD) is used to simultaneously optimize the corresponding parameters of different operators. DARTS is able to relieve some computation burden, but the consumption of resources is still heavy. After that, Guo et al. [14] propose a single path one-shot, namely SPOS. SPOS decouples network searching into supernet training and subnet searching. Firstly, a supernet that contains all of the optional operators should be built. Then, an evolutionary algorithm is used to search for suitable operators by using the trained supernet. Since only one path can be activated during a training iteration, the consumption of resources is much smaller than DARTS. Based on the above previous NAS models, many methods [4,13,32] adopt it to search high-performance deep networks. For example, LightTracker [44] introduce SPOS to search a lightweight tracker for industrial deployments in the object tracking task. Auto-MSFNet [48] proposes a NAS-based multi-scale fusion network for salient object detection.

3 Preliminaries

Before the introduction of the proposed method, we give a short instruction for neural architecture search (NAS). Generally, in order to strengthen the learning capability of the neural network, a complex topological structure, including lots of multiple branches or skip connections, will be introduced in the search space. However, this kind of NAS will cost a lot of computation resources and increase search time. In this paper, to balance the resources and network performance, we introduce some prior knowledge of network design and restrict the topology of neural networks. Recently, the U-Net framework is straightforward and widely deployed in the community of image enhancement. We also adopt this model and then search its specific operators/layers. To obtain the optimal subnet, we exploit One-Shot NAS [14] as our search strategy. The reason is that One-Shot NAS can save computation resources (especially GPU memory), and is easier to converge than the previous method [33].

Specifically, One-Shot NAS is regarding all network structures as different subnets of a supernet and shares their corresponding weights between the structures with the same operators/layers. The entire process of One-Shot NAS includes three stages: supernet training, subnet searching and subnet retraining. In the first stage, the supernet $N(S, W)$ is built with the search space S and network parameters W, which are shared by all the architecture candidates. We firstly need to train this supernet and obtain all the optimized parameters by using the training set:

$$W^* = \arg\min_{W} L_{train}(N(S, W)), \tag{1}$$

where L_{train} is the loss function in the training stage and W^* is the optimized parameters after minimizing the loss function.

During the subnet searching, we can use the trained supernet to search the optimal subnet s^* on the validation set:

$$s^* = \arg\max_{s \in S} Acc_{val}(N(s, W^*(s))), \tag{2}$$

where Acc_{val} represents the accuracy of subnets on the validation set based on the trained supernet with its parameters W^*. During the search stage, we choose subnets by different sampling algorithms, such as random sampling. In this paper, we follow [44] to use evolutionary algorithms.

In the final stage, the optimal subnet $M(\cdot, \cdot)$ needs to be re-trained by using the input data X from the training set to obtain the final optimized parameter $W^{'}$:

$$W^{'} = \arg\min_{W} L_{train}(M(X; W)), \tag{3}$$

4 The proposed method

Our NAS-based enhancement network inherits the widely used U-Net architecture, which consists of two components: encoder and decoder. Different from the previous U-Net architectures, the original residual structures or convolution blocks are replaced by the proposed NAS blocks so that the network can automatically select the most suitable operators and learn robust and reliable features for image enhancement. Figure. 3 shows the proposed NAS-based framework. Given the low-quality image, its two color spaces, namely RGB and Lab images, are fed into the proposed network. In the encoder, multi-level features are extracted by using downsampling operation and different receptive fields. Then, these features are further upsampled and fused to recover the final enhanced image in the decoder.

4.1 Overall pipeline

Given the underwater image $I^j \in R^{H \times W \times 3}$, where $H \times W$ represents the size of the image, j denotes the color space, we use its RGB and Lab, two different color spaces to extract features. The input images are fed into the network, which consists of different NAS blocks. For each block, as shown in Figure. 3 (a), there is a convolution operator for adjusting the channel number and a choice layer for operator selection. We also employ a skip connection in the NAS block to accelerate network convergence. This whole process can be written as below:

$$\begin{aligned} F_i^j &= Conv(F_{i-1}^j) \\ F_{i+1}^j &= op(F_i^j) + F_i^j, \qquad op \in S, \end{aligned} \tag{4}$$

where $op(\cdot)$ is an optional operation and S denotes its corresponding search space. $Conv(\cdot)$ represents the convolutional operation with kernel 3×3. F_{i-1}^j,

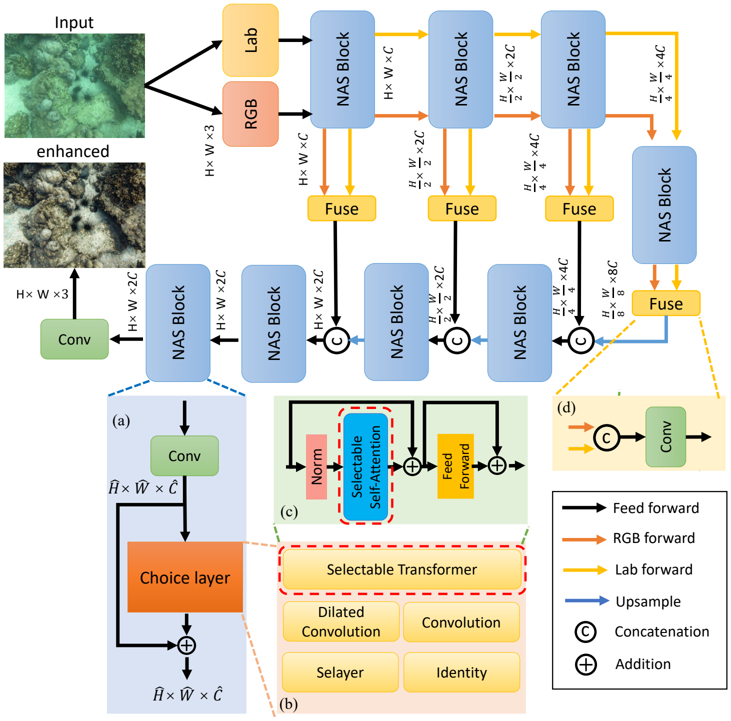


Fig. 3. The proposed framework for network searching. In the encoder, the proposed NAS blocks are employed to extract multi-level features from the image with multi-color spaces. In the decoder, we recover the enhanced images by gradually fusing the multi-level features. The core modules include: (a) NAS Block. (b) Search Space. (c) Selectable Transformer. (d) Feature fusion from two color spaces.

F_i^j, F_{i+1}^j are the feature maps from previous, current and next NAS block, respectively. Taking the first NAS block as an example, the previous feature maps F_0^j, namely the input image $I^j \in R^{H\times W\times 3}$ are fed into the network. Then, through the convolution, we can obtain the feature map $F_1^j \in R^{H\times W\times C}$, whose channel number is C. After that, the feature maps are fed into the choice layer and generate the output $F_2^j \in R^{H\times W\times C}$ for the next block. Notice that the NAS blocks in the encoder are used to extract the features from RGB and Lab color spaces. In the training phase, the parameters of the same operator are shared in a NAS block.

As shown in Figure. 3, starting from the input with $H \times W$, the size is downsampled to half of the original one and the channel number is increased to double by using pixel-unshuffle [38]. As we introduce two color spaces, their features at the same level are integrated before upsampling. Figure. 3 (d) shows its structure:

$$F_i = Conv(||_{j\in\Omega} F_i^j), \tag{5}$$

where || is the concatenation operation. Ω refers to the color spaces, including RGB and Lab color spaces in our framework.

In the decoder, the fusing feature maps from the different color spaces are further concatenated with the decoder features by skip connections to recover gradually the original resolution by the pixel-shuffle operations [38]. In the end, an extra convolution with kernel 1×1 is used to generate the final enhanced image.

4.2 Search space

The search space contains all possible candidate operators. Traditional search space [50,44] only uses some basic operators, such as convolution, pooling, etc. These operators cannot fully discover spatial and color information from the network inputs. In this paper, we further expand the search space, which includes conventional operators, such as convolution, identity, and introduce some new operators like transformer. The specific content is shown in Figure. 3 (b). It contains identity, convolution with 3×3 kernels, dilated convolution with 3×3 kernels and 2 dilations, squeeze-and-excitation block [16] and the proposed selectable transformer.

The previous transformer modules like Vision Transformer (VIT) [7] or Class-Attention in Image Transformers (CaiT) [39] try to encoder their features by image patches and fully-connected layers, which are very useful to extract robust information. However, due to the deployment of fully-connected operators in their structures, the input feature maps need to be resized into a fixed dimension. Moreover, with the increase of input size, the computation complexity grows dramatically as well. In the UIE community, Ushape [35] extends VIT structure and designs two different transformers for feature embedding, but the existing issues are still not alleviated. Each image is warped into a fixed size to meet the requirement of their tailored transformer structures. In real scenes, the images are collected by arbitrary resolutions. If the input image is with high-resolution, the computation complexity can be too heavy to be deployed in some embedded hardware. Some content information may be lost by directly warping the images.

In this paper, inspired by [15,47], we modify the original transformer structure. First, the fully-connected operators in the transformer are replaced by the convolutions. Second, we apply self-attention across channels rather than the spatial dimension, thus generating the attention map in the global context. Therefore, the feature maps with arbitrary sizes can be directly fed into the proposed transformer module. Besides, the computation complexity and the scale of parameters can be decreased remarkably. Further, combining with the NAS strategy, we design a selectable multi-head self-attention module (SMHSA) (shown in Figure. 3 (c)), where we introduce different self-attention mechanisms, including shuffle attention [49], double attention [10], spatial group-wise enhance [31], efficient channel attention [37], thus deriving different transformer structures. Besides, this modification can further expand the search space and boost the feature representation capability of the proposed network. The selectable transformer can be formulated as:

$$\begin{aligned} \hat{F}_i &= op_a(Ln(F_i)) + F_i \quad op_a \in S_a \\ F_{i+1} &= FF(Ln(\hat{F}_i)) + \hat{F}_i, \end{aligned} \tag{6}$$

where $F_i, \hat{F}_i, F_{i+1} \in R^{\hat{H}\times\hat{W}\times\hat{C}}$ represent the input, intermediate and output feature maps of the proposed selectable transformer. The size of the feature maps will not be changed through the proposed NAS block. $Ln(\cdot)$ is the layer normalization. $op_a(\cdot)$ denotes the optional self-attention operators, whose S_a is the subset the proposed the search space S, namely $S_a \subset S$. $FF(\cdot)$ refers to the feed-forward structure, whose fully-connected layers are also replaced by convolutions. To the end, the search space contains 9 operators, including (1) identity (Id), (2) convolution with 3×3 kernels ($Conv$), (3) dilated convolution with 3×3 kernels and 2 dilations ($Dconv$), (4) squeeze-and-excitation block [16] (SE), (5) transformer with transposed attention [47] (T_{ta}), (6) transformer with shuffle attention [49] (T_{sa}), (7) transformer with spatial group-wise enhance attention [31] (T_{sge}), (8) transformer with double attention [5] (T_{da}) and (9) transformer with efficient channel attention [37] (T_{eca}).

4.3 Network optimization

In our framework, following the Eq. 1 and Eq. 3, we need to optimize the supernet and optimal subnet. For the objective function, we employ a combining loss for network optimization. Given the low-quality image $I \in R^{H\times W\times 3}$ and its corresponding high-quality image $G \in R^{H\times W\times 3}$, the proposed network can generate the predicted image $P \in R^{H\times W\times 3}$. Then, the combining loss can be formulated as:

$$L = \alpha * ||G - P||_1 + \beta * ||G - P||_2 + \gamma * \sum_k ||\varphi(G) - \varphi(P)||_1, \tag{7}$$

From Eq. 7, we can see that three different loss functions are jointly used to optimize the network. The first term $||\cdot||_1$ represents L1 loss, which computes the absolute distances between the true value and the predicted value in each pixel. The second term $||\cdot||_2$ is L2 loss. It is used to minimize the error by using the squared distance. The purpose of the L1 and L2 loss functions is to optimize the low-frequency regions. To process the high-frequency information and retain the image style, we introduce perceptual loss, namely the third term in Eq. 7. $\varphi(\cdot)$ denotes the embedding function, which is the output of the k-th layer in VGG-16. Additionally, to balance the magnitude of the loss values, we introduce three loss weights α, β, γ for each term.

5 Experiments

5.1 Experimental setting

Datasets. In this paper, in order to validate the scalability and adaptation of the proposed approach, we introduce three underwater datasets and one low-light dataset in the evaluation experiments. These datasets are described as follows:

Underwater Image Enhancement Benchmark (UIEB) [28]. This dataset contains 890 paired images, but their high-quality images are generated by enhancement methods. Concretely, several enhancement methods are employed to process the low-quality images to generate the enhanced ones. After that, some volunteers will manually choose the best as the final high-quality one. In experiments, the images of UIEB are divided into the training set and testing set, in which 800 and 90 paired images are included, respectively.

Large-Scale Underwater Image (LSUI) [35]. The collection of this dataset almost follows the rule of UIEB, but LSUI is much larger than UIEB. In order to satisfy the training requirements, LSUI collects 5004 underwater images and their corresponding high-quality images. For the setting in [35], 4500 paired images are used for training. The remaining 504 images are used for testing.

Enhancement of Underwater Visual Perception (EUVP) [19]. It contains large-scale paired and unpaired collections for underwater images. These images are captured from different cameras, such as GoPro, AUV's uEye cameras, ROV's HD camera and etc. EUVP also collects some video frames from a few publicly available YouTube videos. Totally, the paired images contain 11435, 570, and 515 pairs for the training, validation, and testing, respectively. Due to the low resolution of the images in this dataset, we mainly use their test set for evaluation in our experiments.

Evaluation Metrics. In this paper, for objective comparison, we introduce two full reference evaluation measures: Peak Signal to Noise Ratio (PSNR) and Structure SIMilarity index (SSIM). They can measure color and structural similarity between the enhanced images and ground truths. Both of their scores are higher the better.

5.2 Implementation details

In this paper, we adopt PyTorch to implement the proposed approach. For our network training, we use the Adam optimizer to minimize the loss function with an initial learning rate of 1.0 $\times 10^{-4}$. The learning rate follows the "poly" adjustment policy so that it can be gradually decreased during the network training. For data augmentation, we use random cropping and random horizontal flipping. The input images are cropped into 256 × 256 and the value of image pixels is normalized to [0,1]. As the supernet training needs a GPU with large memory, we use a PC with an NVIDIA RTX A6000 GPU. The batch size can be set to 10. To balance their loss values, we introduce loss weights α, β, γ for L1, L2 and perceptual loss, respectively. In our experiments, we set them as 0.25, 1, and 0.2. As shown in Figure. 3, the channel number of the feature maps is the multiple of C, we set it to 48 in the current network. For the network testing, our network does not need to resize the images to a fixed resolution. The images of any size can be directly fed into our network.

The proposed method needs to search optimal subnet by a validation set. For the training data, there is a little different from the setting in [35]. We randomly choose 100 images from the training set for validation. Therefore, we totally use 4400, 100, and 504 images in LSUI dataset for training, validation and testing.

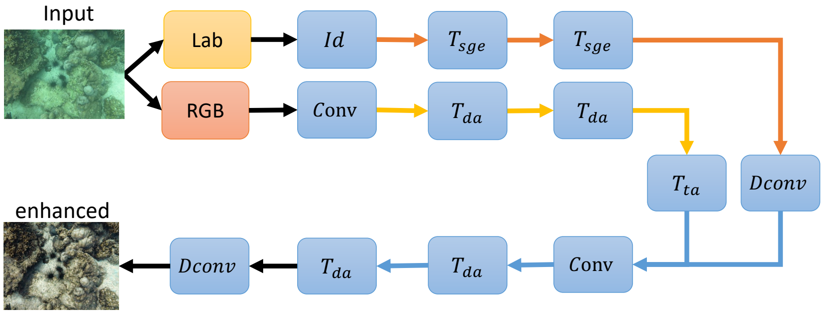


Fig. 4. The optimal network after subnet searching. During the searching, we can automatically obtain the specific operators to extract features from the RGB and Lab images in the encoder. For clear presentation, this figure only shows the chosen operators after network searching. Please understand it together with Figure. 3 and Section. 4.2.

More specifically, we firstly use 4400 images for supernet training. Then, the 100 images are exploited for subnet searching, thus obtaining the optimal subnet. The specific structure is shown in Figure. 4. After that, the subnet is retrained by the training set. Finally, we evaluate the subnet and report the experimental results on the underwater testing set of UIEB (90 paired images), LSUI (504 paired images) and EUVP (515 paired images), respectively.

5.3 Comparisons with the state-of-the-arts

Table. 1 firstly reports the quantitative results on UIEB dataset. In this table, all of the state-of-the-arts are deep learning-based methods. Notice that the proposed approach can achieve the best performance for both PSNR and SSIM metrics. Especially for the PSNR metric, our method obtains performance gains of 2.54 dB over Ushape [35]. Their model also introduces a transformer to encode the deep features. Although the performance of Ushape is competitive, the input images have to be resized to a fixed resolution. The reason is that their transformer still employs linear operators. Moreover, their transformer is based on fully-connected operators, so their model size is much higher than ours. RCTNet [22] obtains favorable performance on PSNR and SSIM as well. Moreover, its model also accepts images with an arbitrary size, but our PSNR and SSIM are better, which provides a substantial gain of 2.18 dB on PSNR. Table. 2 shows the results on LSUI dataset. This dataset contains diverse underwater scenes and object categories, so it is more difficult and challenging. As we can see, without warping the input image and their corresponding ground truths, our PSNR can still reach 26.13 dB, which obtains significant gains of 1.97 dB over Ushape, the original state-of-the-art method. Table. 3 presents the experimental results on EUVP dataset. This is an early dataset and mainly collects colorful underwater creatures. Our method still outperforms significantly the other methods in PSNR. For example, ours can surpass the previous best RCTNet [22] by 3.13 dB and the GAN-based Deep SESR [18] by 5.35 dB. In our network, we mainly ex-

Table 1. Quantitative comparison on the UIEB underwater dataset.

Method	Param.	PSNR	SSIM
WaterNet [28]	25M	19.81	0.8612
FUnIE [19]	7M	19.45	0.8602
UGAN [9]	57M	20.68	0.8430
UIE-DAL [40]	19M	16.37	0.7809
Ucolor [27]	157M	20.78	0.8713
RCTNet[4] [22]	-	22.45	0.8932
Ushape [35]	66M	22.91	0.9100
Ours	12M	**25.45**	**0.9231**

Table 2. Quantitative comparison on the LSUI underwater dataset.

Method	Param.	PSNR	SSIM
WaterNet [28]	25M	17.73	0.8223
FUnIE [19]	7M	19.37	0.8401
UGAN [9]	57M	19.79	0.7843
UIE-DAL [40]	19M	17.45	0.7912
Ucolor [27]	157M	22.91	0.8902
Ushape [35]	66M	24.16	**0.9322**
Ours	12M	**26.13**	0.8608

ploit effective but lightweight operators to construct the proposed search space. Moreover, after removing the fully-connected layers in the transformer, the network parameters can be dramatically decreased. Finally, our scale of parameters achieves 12M, which is competitive against the other state-of-the-arts as well.

Figure. 5 exhibits the visual comparison between the proposed method and the state-of-the-arts on underwater scenes. The shipwreck (the first row) demonstrates that the underwater images may suffer from multiply noises, such as color distortion, blurring, splotchy textures, etc. The previous approaches are able to eliminate some noise and recover the original content to some extent. However, their enhanced images still exist respective drawbacks. For example, FUnlE [19] remove major distorted colors, but the left bottom and right bottom corner still exist the irradicable noise region. WaterNet [28] and Ushape [35] can recover the content and texture of the original image, but the color style of the entire image is changed. Compared with their enhanced images, our result not only restores the image content but also retains the color style as much as possible.

5.4 Ablation study

In this section, to validate the effectiveness of the different components, we design the ablation studies. Table. 4 shows the experimental results. First, we validate the proposed transformer module, namely the selectable transformer. We train

[4] Their code is not released. We cannot obtain their accurate the scale of paramters.

Table 3. Quantitative comparison on the EUVP underwater dataset.

Method	Param.	PSNR	SSIM
WaterNet [28]	25M	20.14	0.6802
FUnIE [19]	7M	23.40	0.8420
UGAN [9]	57M	23.49	0.7802
Deep SESR [18]	3M	24.21	0.8401
RCTNet [22]	-	26.43	**0.8912**
Ours	12M	**29.56**	0.8818

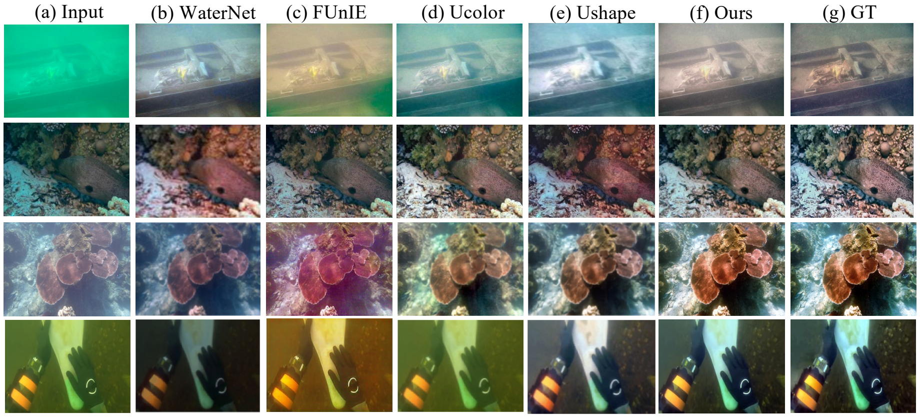


Fig. 5. Visual comparison on underwater dataset. (a) input images with low-light quality. (b) WaterNet [28]. (c) FUnIE [19] (d) Ucolor [27]. (e) Ushape [35] (f) Our enhanced images. (g) Ground truths.

the supernet without the proposed transformer and then search for the corresponding optimal subnet, whose performance is evaluated on three underwater datasets. As shown in Table. 4, without the transformer module, the measures of PSNR and SSIM are decreased dramatically. For instance, its PSNR drops by 2.61 dB, 0.91 dB and 1.41 dB on UIEB, LSUI, EUVP datasets, respectively. It is an enormous degradation. Second, we validate the effectiveness of multi-color spaces. In our framework, we use RGB and Lab two color spaces. Here, we retain RGB images to extract features and remove Lab color space. According to the results, we can see the Lab images are useful in the neural network. With the Lab color space, we can boost the PSNR from 23.69 dB to 25.45 dB on UIEB and the SSIM from 0.8666 to 0.8818 on the EUVP dataset. Although the improvement is less than the transformer modules, multi-color space is an effective part for the image enhancement task. Third, we evaluate the skip connection setting in our network. As shown in our framework, we introduce a skip connection into the proposed NAS block. The network performance will be reduced when this operation is not employed in the NAS block. For example, the PSNR and SSIM are decreased by 1.44 dB and 0.0106 on the LSUI dataset, respectively. These experiments denote that all of the proposed components are effective to enhance the quality of images.

Table 4. Effectiveness of different modules on three datasets by using PSNR and SSIM evaluation metrics. We validate the proposed search space, multi-color spaces inputs and skip connection in our NAS block.

Modules	UIEB		LSUI		EUVP	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Full modules	25.45	0.9231	26.13	0.8608	29.56	0.8818
w/o Transformer	22.02	0.8705	25.12	0.8425	27.80	0.8624
w/o Multi-color spaces	23.69	0.8965	25.70	0.8551	28.03	0.8666
w/o Skip connection	22.82	0.8926	25.39	0.8546	28.12	0.8712

Table 5. Effectiveness of different loss functions on three datasets.

Training setting	UIEB		LSUI		EUVP	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Full Loss Functions	25.45	0.9231	26.13	0.8608	29.56	0.8818
w/o L1 Loss	23.16	0.9028	25.70	0.8432	28.27	0.8578
w/o L2 Loss	23.29	0.9022	25.80	0.8551	28.73	0.8724
w/o Perceptual Loss	22.95	0.8960	25.51	0.8575	28.20	0.8720

For the network optimization, we introduce three different loss functions: L1, L2 and perceptual loss. Table. 5 shows the quantitative results by using different losses to train the network. We gradually remove L1, L2 and perceptual loss to optimize the network. From the results, we can see that all of them are useful to boost network performance. Among them, the perceptual loss can retain the original style and high-frequency information, so it is more effective to improve the quality of images.

From the our final structure (see Figure.4), we can observe: 1) The different deriving transformers have been chosen in the final structure, denoting those different operators can explore larger feature spaces and can generate more suitable feature representations for the underwater image enhancement, 2) the transformer modules are mainly chosen in the deep layers, revealing that transformers might be most suitable to encode the high-level features, and 3) using different operators are more effective than sharing the same modules for different input data, namely RGB and Lab images.

6 Conclusions

In this paper, we employ neural architecture search (NAS) technology to propose a NAS-based U-Net framework. It is able to automatically design a deep model so that it can process severely degraded images, such as turbid water or extremely dark scenes. Moreover, we introduce a search space including the common operators and the proposed selectable transformer module, which assigns the substantial learning capability to our deep model. Besides, the proposed architecture can exploit the multi-color spaces for the underwater scenarios. Finally, the extensive experiments demonstrate that the proposed framework can obtain an optimal neural network and achieve competitive performance on the widely used datasets.

References

1. Ancuti, C.O., Ancuti, C., De Vleeschouwer, C., Bekaert, P.: Color balance and fusion for underwater image enhancement. IEEE Transactions on image processing **27**(1), 379–393 (2017)
2. Ancuti, C., Ancuti, C.O., Haber, T., Bekaert, P.: Enhancing underwater images and videos by fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 81–88. IEEE (2012)
3. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6306–6314 (2018)
4. Chen, Y., Yang, T., Zhang, X., Meng, G., Xiao, X., Sun, J.: Detnas: Backbone search for object detection. Advances in Neural Information Processing Systems **32** (2019)
5. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Aˆ 2-nets: Double attention networks. Advances in neural information processing systems **31** (2018)
6. Chiang, J.Y., Chen, Y.C.: Underwater image enhancement by wavelength compensation and dehazing. IEEE transactions on image processing **21**(4), 1756–1769 (2011)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Drews, P.L., Nascimento, E.R., Botelho, S.S., Campos, M.F.M.: Underwater depth estimation and image restoration based on single images. IEEE computer graphics and applications **36**(2), 24–35 (2016)
9. Fabbri, C., Islam, M.J., Sattar, J.: Enhancing underwater imagery using generative adversarial networks. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 7159–7165. IEEE (2018)
10. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146–3154 (2019)
11. Fu, X., Zhuang, P., Huang, Y., Liao, Y., Zhang, X.P., Ding, X.: A retinex-based enhancing approach for single underwater image. In: Proceedings of the IEEE International Conference on Image Processing. pp. 4572–4576. IEEE (2014)
12. Ghani, A.S.A., Isa, N.A.M.: Underwater image quality enhancement through integrated color model with rayleigh distribution. Applied soft computing **27**, 219–230 (2015)
13. Ghiasi, G., Lin, T.Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7036–7045 (2019)
14. Guo, Z., Zhang, X., Mu, H., Heng, W., Liu, Z., Wei, Y., Sun, J.: Single path one-shot neural architecture search with uniform sampling. In: Proceedings of the European Conference on Computer Vision. pp. 544–560. Springer (2020)
15. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H.: Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704 (2021)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)

17. Iqbal, K., Odetayo, M., James, A., Salam, R.A., Talib, A.Z.H.: Enhancing the low quality images using unsupervised colour correction method. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. pp. 1703–1709. IEEE (2010)
18. Islam, M.J., Luo, P., Sattar, J.: Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. arXiv preprint arXiv:2002.01155 (2020)
19. Islam, M.J., Xia, Y., Sattar, J.: Fast underwater image enhancement for improved visual perception. IEEE Robotics and Automation Letters **5**(2), 3227–3234 (2020)
20. Kim, G., Kwon, D., Kwon, J.: Low-lightgan: Low-light enhancement via advanced generative adversarial network with task-driven training. In: Proceedings of the IEEE International Conference on Image Processing. pp. 2811–2815. IEEE (2019)
21. Kim, H.U., Koh, Y.J., Kim, C.S.: Pienet: Personalized image enhancement network. In: Proceedings of the European Conference on Computer Vision. pp. 374–390. Springer (2020)
22. Kim, H., Choi, S.M., Kim, C.S., Koh, Y.J.: Representative color transform for image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4459–4468 (2021)
23. Kimball, P.W., Clark, E.B., Scully, M., Richmond, K., Flesher, C., Lindzey, L.E., Harman, J., Huffstutler, K., Lawrence, J., Lelievre, S., et al.: The artemis under-ice auv docking system. Journal of field robotics **35**(2), 299–308 (2018)
24. Land, E.H.: The retinex theory of color vision. Scientific american **237**(6), 108–129 (1977)
25. Lee, Y., Jeon, J., Ko, Y., Jeon, B., Jeon, M.: Task-driven deep image enhancement network for autonomous driving in bad weather. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 13746–13753. IEEE (2021)
26. Li, C.Y., Guo, J.C., Cong, R.M., Pang, Y.W., Wang, B.: Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. IEEE Transactions on Image Processing **25**(12), 5664–5677 (2016)
27. Li, C., Anwar, S., Hou, J., Cong, R., Guo, C., Ren, W.: Underwater image enhancement via medium transmission-guided multi-color space embedding. IEEE Transactions on Image Processing **30**, 4985–5000 (2021)
28. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE Transactions on Image Processing **29**, 4376–4389 (2019)
29. Li, C., Guo, J., Guo, C.: Emerging from water: Underwater image color correction based on weakly supervised color transfer. IEEE Signal processing letters **25**(3), 323–327 (2018)
30. Li, J., Skinner, K.A., Eustice, R.M., Johnson-Roberson, M.: Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. IEEE Robotics and Automation letters **3**(1), 387–394 (2017)
31. Li, X., Hu, X., Yang, J.: Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. arXiv preprint arXiv:1905.09646 (2019)
32. Liu, C., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei-Fei, L.: Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 82–92 (2019)
33. Liu, H., Simonyan, K., Yang, Y.: Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)

34. Park, J., Lee, J.Y., Yoo, D., Kweon, I.S.: Distort-and-recover: Color enhancement using deep reinforcement learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5928–5936 (2018)
35. Peng, L., Zhu, C., Bian, L.: U-shape transformer for underwater image enhancement. arXiv preprint arXiv:2111.11843 (2021)
36. Peng, Y.T., Cosman, P.C.: Underwater image restoration based on image blurriness and light absorption. IEEE transactions on image processing **26**(4), 1579–1594 (2017)
37. Qilong Wang, Banggu Wu, P.Z.P.L.W.Z., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
38. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
39. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 32–42 (2021)
40. Uplavikar, P.M., Wu, Z., Wang, Z.: All-in-one underwater image enhancement using domain-adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8 (2019)
41. Uzair, M., Brinkworth, R.S., Finn, A.: Bio-inspired video enhancement for small moving target detection. IEEE Transactions on Image Processing **30**, 1232–1244 (2020)
42. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6849–6857 (2019)
43. Xie, L., Yuille, A.: Genetic cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1379–1388 (2017)
44. Yan, B., Peng, H., Wu, K., Wang, D., Fu, J., Lu, H.: Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15189 (2021)
45. Yang, M., Hu, K., Du, Y., Wei, Z., Sheng, Z., Hu, J.: Underwater image enhancement based on conditional generative adversarial network. Signal Processing: Image Communication **81**, 115723 (2020)
46. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3063–3072 (2020)
47. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. arXiv preprint arXiv:2111.09881 (2021)
48. Zhang, M., Liu, T., Piao, Y., Yao, S., Lu, H.: Auto-msfnet: Search multi-scale fusion network for salient object detection. In: Proceedings of the ACM International Conference on Multimedia. pp. 667–676 (2021)
49. Zhang, Q.L., Yang, Y.B.: Sa-net: Shuffle attention for deep convolutional neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2235–2239. IEEE (2021)

50. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)