Spectrum-inspired Low-light Image Translation for Saliency Detection

Kitty Varghese[1], Sudarshan Rajagopalan[2], Mohit Lamba[1], Kaushik Mitra[1]

[1] Indian Institute of Technology, Madras, India

[2] Madras Institute of Technology, India

Abstract

Saliency detection methods are central to several real-world applications such as robot navigation and satellite imagery. However, the performance of existing methods deteriorate under low-light conditions because training datasets mostly comprise of well-lit images. One possible solution is to collect a new dataset for low-light conditions. This involves pixel-level annotations, which is not only tedious and time-consuming but also infeasible if a huge training corpus is required. We propose a technique that performs classical band-pass filtering in the Fourier space to transform well-lit images to low-light images and use them as a proxy for real low-light images. Unlike popular deep learning approaches which require learning thousands of parameters and enormous amounts of training data, the proposed transformation is fast and simple and easy to extend to other tasks such as low-light depth estimation. Our experiments show that the state-of-the-art saliency detection and depth estimation networks trained on our proxy low-light images perform significantly better on real low-light images than networks trained using existing strategies.

CCS Concepts

• **Computing methodologies → Interest point and salient region detections**; **Interest point and salient region detections**.

Keywords

Low-light and salient object detection

Figure 1: Saliency and depth estimation networks perform poorly for low-light images, see *(d)* & *(j)*, because datasets mainly comprise of well-lit images. We propose a simple transformation from well-lit to low-light images. Training existing models on our proxy low-light images significantly boosts the model's performance on *real* low-light images, see *(f)* & *(l)*.

1 Introduction

Saliency detection models aim to identify prominent subjects in a scene, which is useful in several tasks such as robot navigation [21, 37], satellite imagery [15, 47], video summarization [18], foreground annotation [5], and action recognition [39, 40]. In a real-world scenario, these applications require the saliency detection model to perform well in both good and bad lighting conditions. But, past studies in this domain [10, 24, 36] have focused mainly on

good lighting conditions with their effectiveness deteriorating for low-light images, as shown in Fig. 1.

An obvious solution is to pre-process low-light images using existing restoration methods [13, 19, 41] and then feed them to saliency detection models trained for well-lit images. But our experiments indicate that this does not yield satisfactory results, see Fig. 3. Another alternative is to create a new dataset for low-light conditions. This can be done by manually annotating salient objects in existing low-light datasets [4, 6] or by retouching well-lit saliency detection datasets [34] in image editing softwares like Adobe Lightroom and GIMP [38, 42]. Either-way, this could be laborious, time-consuming, and perhaps even infeasible when a large amount of training data is required.

To alleviate the above challenges, several image translation [2, 33, 35] and domain adaptation [22, 46] methods have been proposed. For example, HiDT [2] adopts an encoder-decoder architecture to decompose a well-lit image into its style and content and consequently uses adversarial learning to transform well-lit images into low-light images. Nonetheless, such GAN-based solutions are difficult to train and susceptible to problems such as mode collapse [12]. Recently, Yang *et al.* [46] proposed a simple domain adaptation technique, called Fourier Domain Adaptation (FDA), wherein they swap the low frequencies of the source and target domain images. In the present context, source domain represents well-lit images while target domain represents low-light images. However, FDA is likely to introduce ringing artifacts in the transformed image due to the Gibbs phenomenon [31], leading to sub-optimal results, as discussed in Sec. 4.6.

To alleviate above problems, we propose a transformation that fuses the amplitude spectrum of a well-lit image with that of a low-light image using band-pass filtering, as shown in Fig. 2. We keep the phase spectrum as it is, because it contains structural information about the source image [32]. During band-pass filtering of the amplitude spectrum, we also perform a windowing operation to facilitate smooth transition of frequencies and to curb ringing artifacts. The proxy low-light image is finally obtained by computing the inverse Fourier transform of the fused amplitude response and the phase spectrum of the well-lit image. These transformed well-lit images into proxy images are then used to train existing networks for real low-light conditions. Our proposed approach is computationally and memory efficient as it requires tuning a couple of hyper-parameter and needs only $3 - 4$ real low-light images for the transformation of well-lit images into proxy images. This is in contrast with popular deep-learning-based models which require training hundreds of parameters and a lot of images. For the aforementioned reason, our proposed transformation can be easily generalised to other computer vision tasks in low-light conditions. We show that networks trained using our proxy images perform significantly better on real low-light images for downstream computer vision tasks such as saliency prediction and depth estimation.

Our contributions can be summarised as below:

- We propose a technique for transforming well-lit images into proxy low-light images, which can then be used to train existing networks for real low-light conditions.
- Unlike popular deep-learning-based solutions, our approach requires tuning only a couple of hyper-parameters and a

handful of real low-light images. Thus, the proposed transformation can be easily generalized to other computer vision tasks.
- We demonstrate both qualitatively and quantitatively that the state-of-the-art saliency detection and depth estimation networks trained on our proxy low-light images perform significantly better on real low-light images.

2 Related Works

Saliency prediction models can be classified as bottom-up and top-down models. Bottom up saliency models use low-level features and are stimuli driven as discussed in [17]. Work by Goferman et al. [11] detects saliency by computing the local and global contrast. Kim et al. [20] in their work used a regression based model and color transform to calculate local and global saliency. These bottom up saliency networks often fail in detecting salient objects when the background is cluttered and in low contrast regions.

Whereas, top-down models use high level features to detect salient objects. Xu et al. [43] in their work predict saliency maps using a support vector machine (SVM) model. A covariance based CNN model was used by Mu et al. [28] to learn saliency values in image patches. Dong et al. [7] used feature fusion and feature aggregation in their bidirectional collaboration network (BCNet) for detecting salient objects. It is observed that top down saliency networks demand high computational requirements, yet they fail to predict accurate boundaries of salient objects in low-light conditions. Thus, we see that low-light saliency detection is a largely unexplored problem. We propose a method to address this problem by generating proxy low-light images from well-lit images.

Past works have also explored image translation methods to solve similar problems but not saliency detection in low light conditions. We give a brief overview of them. Park et al. [33], used unpaired image-to-image translation using contrastive learning for domain adaptation. Anokhin et al. [2], used the style and content representation of an image to translate into desired domain. Long et al. [26] used per-pixel regression for classification to solve image-to-image translation. Li et al. [23] used PatchGAN architecture to locate style statistics. Isola et al. [16] used Pix2pix to map functions between input and output images. However, most of these methods use deep networks which are data hungry and need a lot of training time. Recently, Yang et al. [46] proposed Fourier domain adaptation (FDA) which overcomes these limitations as they do not need a large training corpus.

3 Spectrum inspired low-light image translation

3.1 Method Overview

We propose a method to convert well-lit images into proxy images. Our main objective is to reduce the domain gap for downstream computer vision applications by fusing the statistics of low-light and well-lit images. This enables networks to perform downstream vision tasks in low-light conditions even in the absence of real low-light datasets. We do not place much emphasis on making the proxy images look visually indistinct from real low-light images.
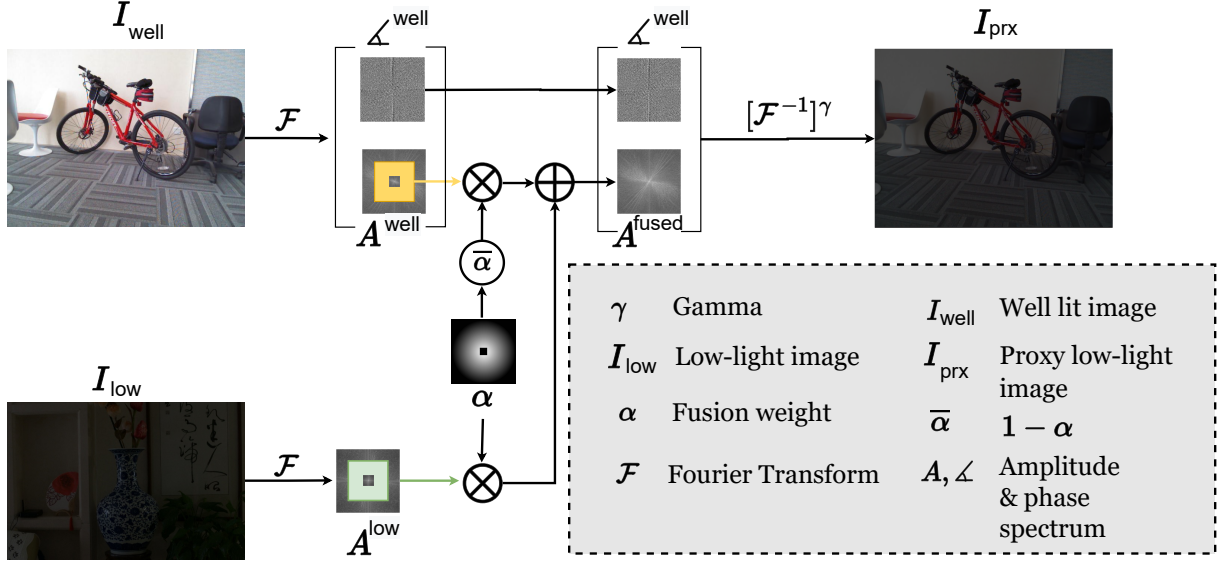
Figure 2: Block diagram of the proposed method.

Our method takes inspiration from the fact that in the Fourier representation of an image, it is the phase that carries most relevant information needed to restore the image, and changes made to the amplitude spectrum do not alter higher-level semantics. We thus retain the phase spectrum of the well-lit image as it is. The amplitude spectrum of the well-lit image, on the other hand, is fused with the amplitude spectrum of a real low-light image using weighted averaging. Further, to preserve the colors we use band-pass filtering and adopt 2D windowing for suppressing the ringing artifacts. Using our method mitigates the problem of building a large real low-light dataset which may be time consuming and laborious. Since, our method mainly involves modification of the spectral characteristics of images, the computation efficiency depends mainly on that of the FFT algorithm. This makes it very fast compared to training neural networks for image translation and has a very low memory footprint (See Sec. 4.3).

3.2 Low-light and well-lit fusion

Fig. 2 shows the various steps involved in our transformation pipeline. Given any real well-lit image $I_{\text{well}} \in \mathbb{R}^{H \times W \times 3}$, we randomly choose a real low-light image I_{low} from a pool of real low-light images and resize it to I_{well}'s resolution. We next decompose the images into their respective amplitude and phase spectrums using the 2D Fourier Transform $\mathcal{F}$ as

$$A^{\text{well}}, \angle^{\text{well}} = \mathcal{F}(I_{\text{well}}) \text{ and } A^{\text{low}}, \angle^{\text{low}} = \mathcal{F}(I_{\text{low}}). \quad (1)$$

The image semantics are better preserved in the phase response [32] and so we do not modify $\angle^{\text{well}}$. We however, compute a weighted average of A^{well} and A^{low} to obtain the fused amplitude spectrum A^{fused}. For the fusion, more weightage is given to A^{well} for high frequencies and to A^{low} for low frequencies (See Eq. 2). We do this to ensure that the proxy image I_{prx} has the semantics of I_{well} and the style of I_{low} [30].

$$A_{m,n}^{\text{fused}} = \alpha_{m,n} \cdot A_{m,n}^{\text{low}} + (1 - \alpha_{m,n}) \cdot A_{m,n}^{\text{well}} \quad (2)$$

Algorithm 1 Proxy Dataset Generation

Input: $\mathcal{D}_{\text{well}}$: dataset of well-lit images; $\mathcal{D}_{\text{low}}$: pool of real low-light images.

Hyperparameters: $\lambda_l, \lambda_u, \gamma$.

Remarks: $\mathcal{D}_{\text{low}}$ can have unpaired images with respect to $\mathcal{D}_{\text{well}}$ and should have at least 1 real low-light image, i.e. $|\mathcal{D}_{\text{low}}| \geq 1$.

Output: $\mathcal{D}_{\text{prx}}$: dataset of proxy images.

1:   $\mathcal{D}_{\text{prx}} = \{\}$
2:   **for** I_{well} in $\mathcal{D}_{\text{well}}$ **do**
3:      **if** $|\mathcal{D}_{\text{low}}| > 1$ **then**
4:         Sample a real low-light image, i.e. $I_{\text{low}} \sim \mathcal{D}_{\text{low}}$
5:      **else**
6:         $I_{\text{low}} = \mathcal{D}_{\text{low}}$
7:      **end if**
8:      $I_{\text{low}} = \text{resize}(I_{\text{low}}, \text{size} = \dim(I_{\text{well}}))$
9:      $A^{\text{well}}, \angle^{\text{well}} = \text{DFT}(I_{\text{well}})$
10:     $A^{\text{low}}, \angle^{\text{low}} = \text{DFT}(I_{\text{low}})$
11:     Define $\mathcal{R} = \mathcal{R}_u - \mathcal{R}_l$ where $\mathcal{R}_u, \mathcal{R}_l$ are given by Eq. 5
12:     Compute mask α_B as defined in Eq. 3
13:     $A^{\text{fused}} = \alpha_B \cdot A^{\text{low}} + (1 - \alpha_B) \cdot A^{\text{well}}$
14:     $I_{\text{prx}} = \left[IDFT(A^{\text{fused}}, \angle^{\text{well}}) \right]^{\gamma}$
15:     Append I_{prx} to $\mathcal{D}_{\text{prx}}$
16:   **end for**
17:   **return** $\mathcal{D}_{\text{prx}}$

During fusion it is also necessary to ensure a smooth transition of frequencies, otherwise the proxy image I_{prx} will have significant ringing artifacts due to Gibbs effect [31]. Our fusion weights $\alpha_{m,n}$

Least Salient — **Most Salient**

| | Real well-lit image | Real low-light image | a) **Tr:** Well-lit **Te:**EnLow | b)**Tr:**CUT **Te:**Low-light | c) **Tr:**HiDT **Te:**Low-light | d)**Tr:**CoMoGAN **Te:**Low-light | e) **Tr:** FDA **Te:**Low-light | f) **Tr:** Ours **Te:**Low-light | g) **Tr:** Well-lit **Te:**Well-lit |
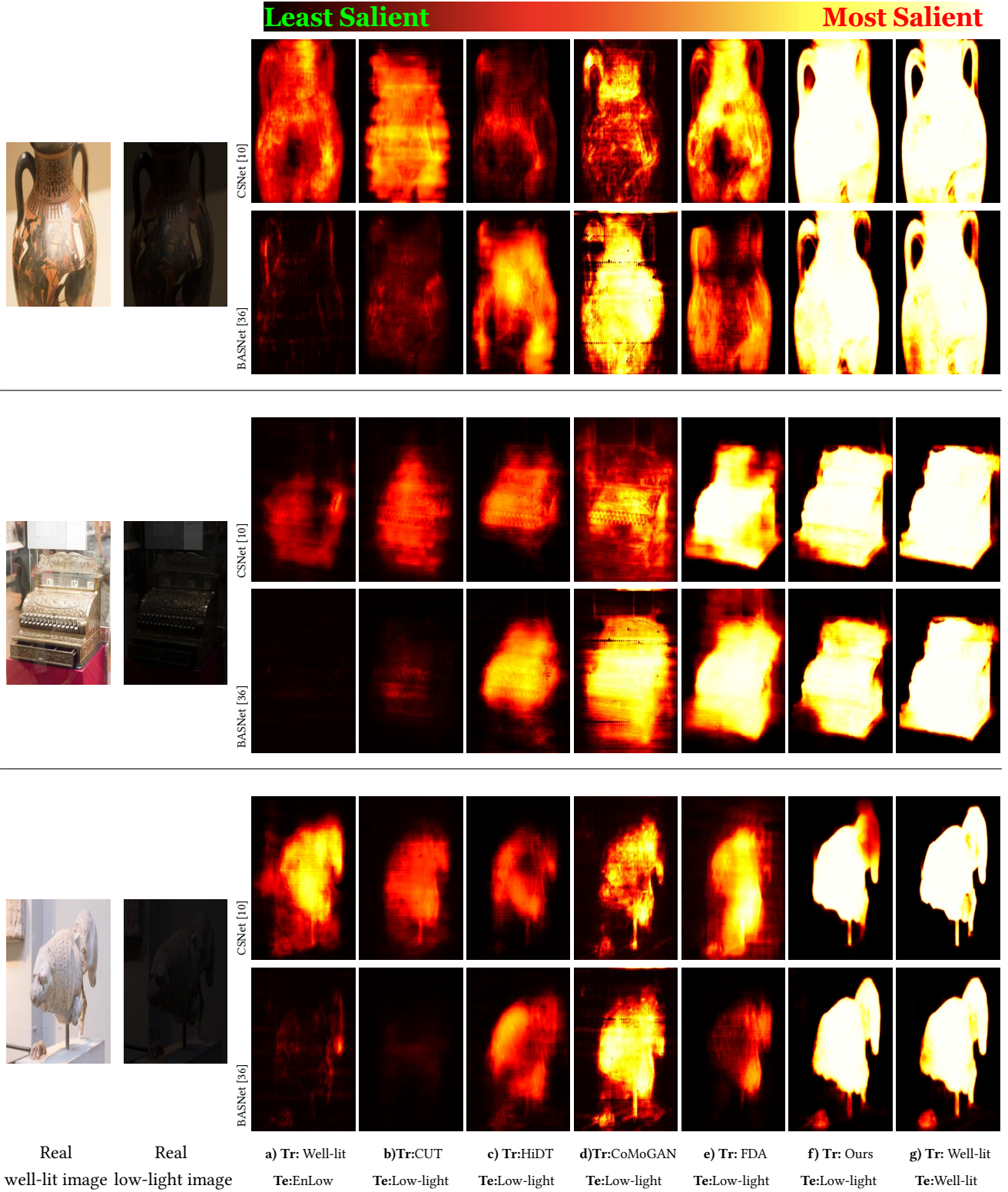
Figure 3: [Tr: *Training***; Te:** *Testing***; EnLow:** *Enhanced low-light using Zero-DCE* **[13]] Saliency Detection by CSNet [10] and BASNet [36] on** *real* **low-light images from the SICE dataset [4].** *(a)***: Enhancing low-light images barely improves the performance of the networks trained for well-lit images.** *(b), (c), (d), (e)***: Marginal improvements are observed when the networks are trained on images simulated using CUT [33], HiDT [2], CoMoGAN [35] and FDA [46].** *(f)***: Training models on our proxy low-light images significantly improves saliency detection on real low-light images and the predictions are close to** *(g)***.**

Table 1: Quantitative results for saliency detection averaged over SICE's [4] real low-light images. The best result is in bold and second best is underlined. Our proposed strategy significantly outperforms existing methods.

	CUT [33]	HiDT [2]	CoMoGAN [35]	Zero-DCE [13]	FDA [46]	Ours
			BASNet [36]			
E-measure↑	0.391	0.453	0.423	0.512	0.599	**0.602**
S-measure↑	0.323	0.344	0.401	0.382	0.568	**0.831**
F-measure↑	0.596	0.609	0.731	0.712	0.874	**0.921**
MAE↓	0.462	0.311	0.243	0.296	0.168	**0.092**
			CSNet [10]			
E-measure↑	0.498	0.518	0.621	0.611	0.587	**0.675**
S-measure↑	0.388	0.417	0.532	0.503	0.631	**0.801**
F-measure↑	0.621	0.693	0.756	0.732	0.755	**0.923**
MAE↓	0.321	0.249	0.221	0.256	0.201	**0.105**

are inspired from the classical Blackman windowing [31]. We empirically found that it is also necessary to retain the DC frequencies of I_{well}, otherwise the overall contrast of I_{prx} is destroyed (see Fig. 6). We therefore compute fusion over a band of frequencies and not over the entire spectrum. Formally, $\alpha_{m,n}$ is computed as

$$\alpha_{m,n} = \begin{cases} w_{m,n} & \forall\, m, n \in \mathcal{R}_u - \mathcal{R}_l \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where,

$$w_{m,n} = \left[0.42 + 0.5\cos\left(\frac{2\pi m}{\lambda_u \cdot H}\right) + 0.08\cos\left(\frac{4\pi m}{\lambda_u \cdot H}\right)\right]$$
$$\times \left[0.42 + 0.5\cos\left(\frac{2\pi n}{\lambda_u \cdot W}\right) + 0.08\cos\left(\frac{4\pi n}{\lambda_u \cdot W}\right)\right] \tag{4}$$

$$\mathcal{R}_l \leftarrow m \in [-\lambda_l \tfrac{H}{2}, \lambda_l \tfrac{H}{2}], \text{ and } n \in [-\lambda_l \tfrac{W}{2}, \lambda_l \tfrac{W}{2}]$$
$$\mathcal{R}_u \leftarrow m \in [-\lambda_u \tfrac{H}{2}, \lambda_u \tfrac{H}{2}], \text{ and } n \in [-\lambda_u \tfrac{W}{2}, \lambda_u \tfrac{W}{2}]$$
$$0 \le \lambda_l < \lambda_u < 1 \tag{5}$$

Finally, I_{prx} is obtained using the inverse Fourier transform as shown in Eq. 6. $\gamma > 1$ controls the overall brightness of I_{prx}. Increasing the value of γ yields a darker proxy low-light image I_{prx}.

$$I_{\text{prx}} = \left[\mathcal{F}^{-1}(A^{\text{fused}}, \angle^{\text{well}})\right]^{\gamma} \tag{6}$$

Empirically, we observed that visual artifacts begin to appear as we increase the value of λ_l and λ_u. Therefore, for our simulation, we used $\lambda_l = 0.01$ and $\lambda_u = 0.1$ (See Sec. 4.6). Our proposed method can be iteratively applied to all well-lit images belonging to a dataset. For this, only few real low-light images are required for transformation. The details for transforming such well-lit datasets are given in algorithm 1. Also for this algorithm to work, we do not require a paired set of well-lit and low-lit images, and they can belong to cameras of different make and model or even depict different scenes.

Table 2: Comparison of the training time and number of parameters used by various methods to translate well lit images into proxy low-light images. Compared to other methods which have millions of parameters, FDA and our strategy contain only a couple of hyper-parameters. Thus FDA and our method do not require several hours of training time.

	CUT	HiDT	CoMoGAN	FDA	Ours
Parameters	18.7M	9.8M	56.8M	1	2
Train Time (in hrs)	24	24	48	N/A	N/A

4 Experiments

4.1 Experimental Settings

To evaluate the proposed technique for salient object detection we use the NLPR [34], LIME [14], and SICE [4] datasets. The NLPR dataset contains 1000 well-lit images of size 640×480 with corresponding GT annotations for salient objects. LIME has 10 real low-light images from which we used 5 images to translate well-lit images into low-light images. The SICE dataset contains 589 well-lit images with corresponding real low-light images of resolutions varying from 3000×2000 to 6000×4000. Proxy low-light images generated using NLPR well-lit images are used for training state-of-the-art saliency detection models CSNet [10] and BASNet [36] while real low-light images of SICE dataset are reserved for testing. Due to the absence of GT annotation for real low-light images, we consider the saliency predictions of BASNet and CSNet trained for well-lit conditions on SICE's well-lit images as the ground truth respectively.
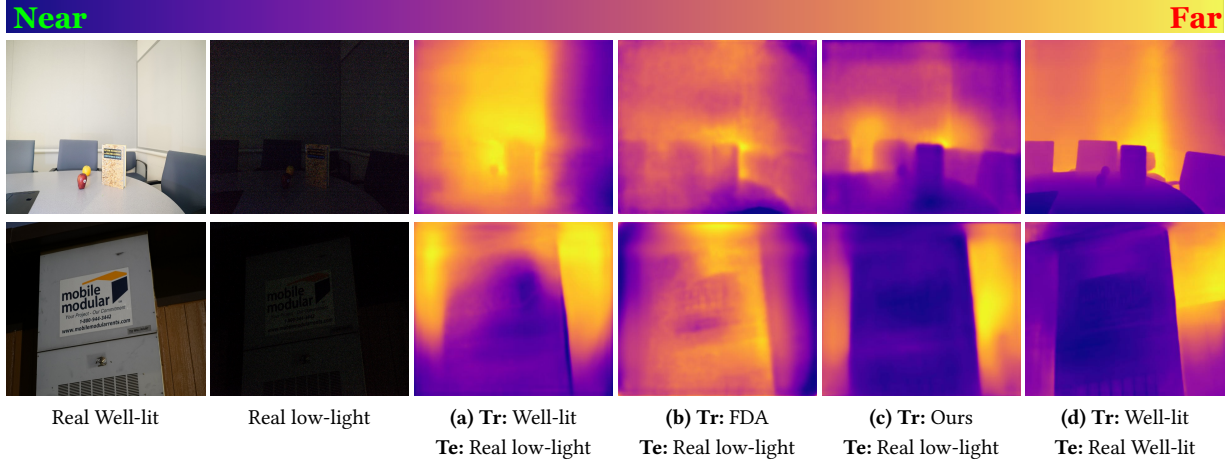
Near Far

| Real Well-lit | Real low-light | **(a) Tr:** Well-lit **Te:** Real low-light | **(b) Tr:** FDA **Te:** Real low-light | **(c) Tr:** Ours **Te:** Real low-light | **(d) Tr:** Well-lit **Te:** Real Well-lit |

Figure 4: Depth estimation using AdaBins [3] on *real* low-light images from the SID dataset [6]. *(a)*: AdaBins when trained on well-lit images degenerates for low-light conditions. *(b)*: Training AdaBins using FDA barely improves the performance. *(c)*: Training AdaBins on our proxy low-light images significantly improves depth estimation for real low-light images. Our results are close to ground truth shown in *(d)*.

Table 3: Quantitative comparison for depth estimation on real low-light images [6]. The best result is in bold and second best is <u>underlined</u>. Our method outperforms FDA.

Trained On	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	REL$\downarrow$	RMSE$\downarrow$
Well-lit [29]	<u>0.456</u>	0.71	0.878	0.389	0.725
FDA [46]	0.454	<u>0.794</u>	<u>0.939</u>	<u>0.318</u>	<u>0.644</u>
Ours	**0.523**	**0.833**	**0.961**	**0.276**	**0.569**

We compare the performance of our method with HiDT [2], CUT [33], CoMoGAN [35] and FDA [46]. HiDT, CUT and CoMo-GAN are GAN based deep learning networks for image translation, while FDA uses classical signal processing for domain adaptation. The low-light images generated by all these methods from the well-lit NLPR dataset are then used to re-train BASNet and CSNet. FDA and our method uses 5 real low-light images from the LIME dataset for low-light image conversion. CUT has to be re-trained for this task since it was not designed for well-lit to low-light transformation. As 5 images are too less for training GAN based models, additional 3000 images from the Ex-Dark dataset [25] are used when training GAN based models. We also tried fine-tuning HiDT and CoMoGAN, but as they are specifically designed for low-light translation, the performance of pre-trained models is better and we use them for all comparisons.

We additionally compare with Zero-DCE [13] which is used to enhance low-light images as a pre-processing step. We could not compare with works of Xu *et al.* [45], [27], [44] since neither their code nor their dataset is publicly available.

We use PyTorch running on a CPU with 32GB RAM and a 12GB K80 GPU for implementing the proposed method. Unless stated otherwise, lower-frequency (λ_l), upper-frequency (λ_u) and gamma (γ) are set to 0.01, 0.10 and 3.5, respectively. Other parameters such as

the loss function, optimiser and data augmentations are as mentioned in the available codes of above stated methods.

4.2 Qualitative and Quantitative comparisons

In Fig. 3 we visually compare the saliency maps generated by BAS-Net and CSNet in different situations. We observe that the simple pre-processing step of enhancing low-light images using Zero-DCE before feeding them to BASNet [36] and CSNet [10] trained on well-lit images yields unsatisfactory results. Marginal improvements are observed if well-lit images are first translated to low-light images using HiDT [2], CUT [33] and CoMoGAN [35] and then used to re-train BASNet and CSNet. This is mainly because, adversarial training is often susceptible to training instabilities and unnatural artifacts in the generated images. Training using FDA proxy images yields better predictions compared to other methods, but is still quite inferior to ground truth. This is because, as discussed in Sec. 4.6, FDA transformed images have considerable ringing artifacts. Predictions using our transformation not only outperform all existing methods but are almost at par with ground truth. Our superiority is also supported by Table. 1 where we outperform existing methods on all four metrics, namely, E-Measure [9], S-measure [8], F-measure [1] and Mean-Absolute-Error (MAE).
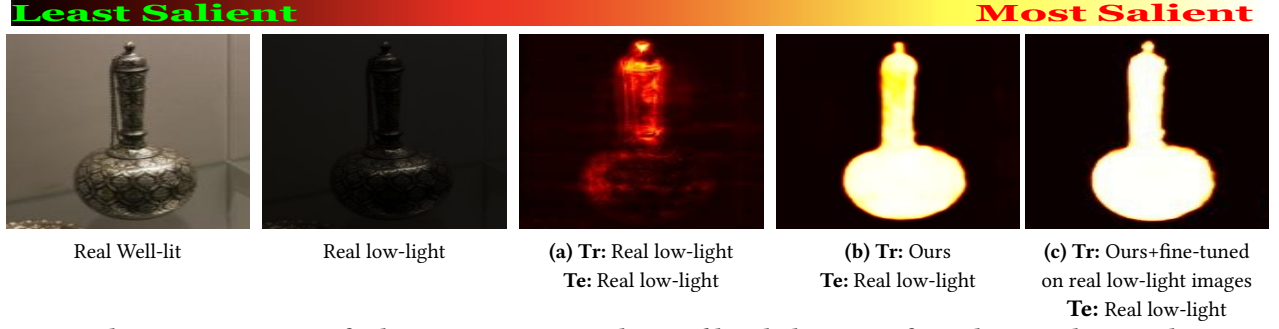
Least Salient   **Most Salient**

| Real Well-lit | Real low-light | **(a) Tr:** Real low-light **Te:** Real low-light | **(b) Tr:** Ours **Te:** Real low-light | **(c) Tr:** Ours+fine-tuned on real low-light images **Te:** Real low-light |

Figure 5: Qualitative comparison of saliency maps generated on *real* low-light images from the SICE dataset when CSNet is trained on: *(a)* real low-light images, *(b)* our proxy low-light images, *(c)* fine-tuning (b) on real low-light images. Without using our synthetic images, it is not possible to get good performance under low-light conditions because of the absence of publicly available large-scale datasets for low-light saliency detection.

Table 4: Quantitative comparison for CSNet trained on: (a) real low-light images from SICE, (b) our proxy images and (c) our proxy images followed by fine tuning on real low-light images from SICE. The best result is in bold and second best is underlined. Training CSNet on real low-light images yields poor results due to the absence of large-scale datasets for low-light saliency detection. However, using our synthetic images to increase the training size significantly improves performance as indicated in columns 2 and 3.

Trained On	**Real low-light images**	**Ours**	**Ours+fine-tuned on real low-light images**
S-measure↑	0.619	<u>0.801</u>	**0.821**
F-measure↑	0.823	<u>0.923</u>	**0.939**

4.3 Time-Complexity

Table. 2 reports the training time required by CUT, HiDT, CoMo-GAN, FDA and the proposed method for generating proxy low-light images. This includes the time needed for training GAN based methods. We see that GAN based methods take at least 48× more time than FDA and Ours to transform images. Compared to deep learning networks, which have millions of learnable parameters, the proposed transformation has only 2 hyper-parameters i.e., λ_l and λ_u. FDA has only one hyper-parameter, β, which is comparable to λ_u in our algorithm. If γ is also considered, hyper-parameter count for FDA and ours increase by one. Thus, our method not only exhibits qualitative and quantitative superiority but is also fast with a low number of parameters.

4.4 Generalizabilty

Our method is easy to generalize to other computer vision tasks. We demonstrate this by extending our pipeline for depth estimation under extreme low-light conditions. Specifically, we re-train a recent depth estimation network AdaBins [3] on our proxy low-light images generated using well lit images present in the NYU dataset [29] and then test it on real extreme low-light images from the SID dataset [6]. The NYU dataset consists of 640×480 well-lit images with ground truth depth annotations and the SID dataset consists of 4256×2848 real night-time images with their corresponding well-lit images. For this experiment we use only the low-light images captured with 0.1s exposure. For transforming NYU well-lit images we used just *one* real low-light image from the SID dataset with lower-frequency (λ_l), upper-frequency (λ_u) and gamma (γ) set

to 0.01, 0.1 and 6 respectively. We have increased the γ from 3.5 to 6 as SID images are much more dark than SICE dataset. Similar settings are used for the FDA pipeline. For benchmarking, we compute GT depth by passing the well-lit SID images through the original AdaBins trained for well-lit images. The qualitative results can be found in Fig. 4 and quantitative results in Table. 3 where we use the same metrics as used in the AdaBins paper.

4.5 Training on real low-light images

There is no publicly available large scale dataset to train networks for low-light saliency detection. We however show that such networks can be first trained on our proxy images and then fine-tuned on a limited number of real low-light images to improve performance. We do this by evaluating the performance of CSNet under three scenarios: (i) training on a limited number of real low-light images from the SICE [4] dataset, (ii) training on our proxy image dataset obtained from the well-lit NLPR saliency dataset which has large number of images and (iii) by fine-tuning the network obtained in (ii) using limited number of real low-light images from (i).

The NLPR dataset consists of well-lit images with corresponding ground truth saliency maps but lacks low-light images. On the other hand, the SICE dataset has well-lit and low-light pairs but lacks ground truth saliency maps. Thus as described in Sec. 4.1, for (i) we treated the saliency maps generated by passing well-lit SICE images through CSNet trained for well-lit conditions as the ground truth. After discarding the images for which the ground truth maps were not appropriate by manual inspection, we finally obtained
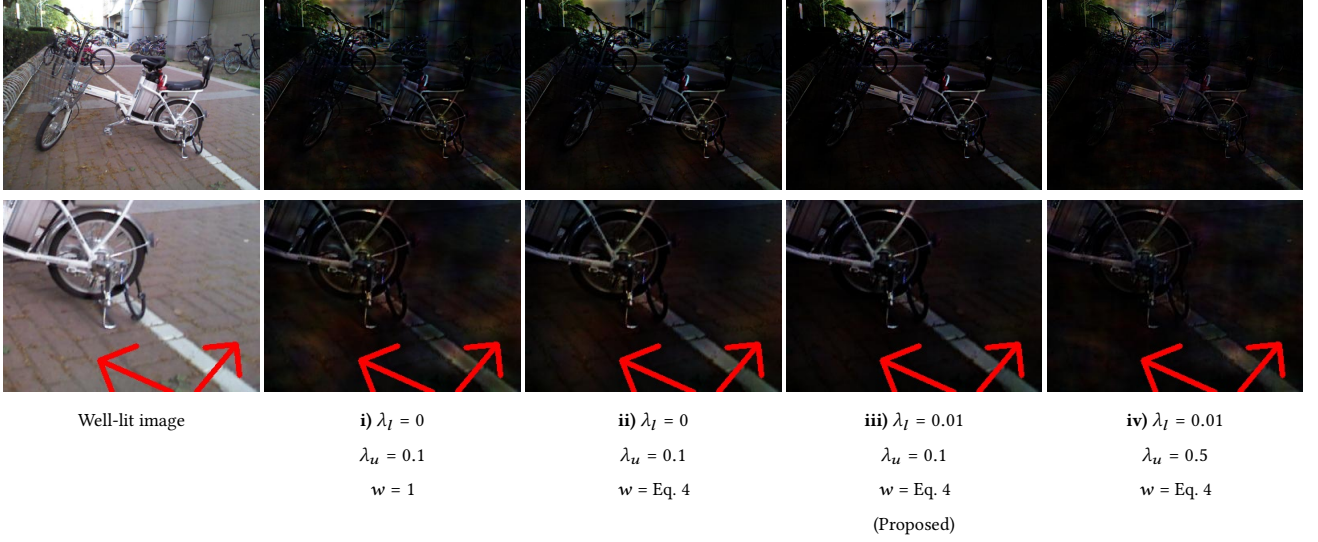
Figure 6: Ablation study showing the effect of λ_l, λ_u and w in generating I_{prx}. γ was set to 2.5 for all the images. Color and ringing artifacts can be observed in i). However, our windowing technique suppresses these ringing artifacts as shown in ii). But, color artifacts are still present in ii) which are indicated by the red arrows. These color artifacts are diminished by using our proposed band-pass filtering instead of low-pass filtering as shown in iii). Using a large value of λ_u degrades the visual quality as shown in iv).

156 real low-light images with ground truth saliency. For (ii) we translated well-lit NLPR images into proxy low-light images while retaining original saliency ground truth (see Sec. 4.1 for details).

Table. 4 and Fig. 5 respectively present the quantitative and qualitative results for the different scenarios. The poor performance of the network in Fig.5(a) is due to the limited number of real low-light images available for training. However, using our proxy images for pre-training and then fine-tuning with these limited number of real low-light images (in our case 156) boosts the network's performance as shown in Fig.5(d).

4.6 Ablation Studies

Fig. 6 shows the ablation studies conducted on our method by choosing well-lit images from the NLPR dataset and a real low-light image from the SID [6] dataset. In Fig. 6 i) we do not use weighted averaging for fusion and instead in Eq. 4 we set $w = 1$ which causes sharp discontinuities at the cut-off frequencies $\frac{\lambda_u H}{2}$ and $\frac{\lambda_u W}{2}$. We additionally do not retain the DC frequencies of I_{well} by setting $\lambda_l = 0$. Clearly, the transformed images lack contrast and exhibit severe ringing artifacts. Except for the γ correction, Fig. 6 i) is same as FDA. In Fig. 6 ii) we enforce a smooth fusion of well-lit and low-light images by using w as defined in Eq. 4. This helps limit the Gibbs phenomenon leading to removal of ringing artifacts visible in Fig. 6 i). The colors in Fig. 6 ii), however, continue to be poor. For example in the second row in Fig. 6 ii), the color of the road as indicated by the red arrow has reddish-brown patches. In Fig. 6 iii) we use band-pass filtering instead of low-pass filtering by slightly increasing λ_l from 0 to 0.01. Clearly band-pass filtering leads to better color restorations. Finally in Fig. 6 iv) we use a large value of λ_u which consequently degrades the semantics of I_{well} in the generated proxy low-light image. This is expected because a large value of λ_u implies that even the high frequencies of real low-light

image, which mostly capture the semantics of low-light image, are fused into the frequency spectrum of well-lit image. We, however, only wish to incorporate the style of low-light images and not their semantics into the well-lit images. As Fig. 6 iii) qualitatively yields better low-light proxy images, we fix λ_l and λ_u to 0.01 and 0.1 respectively.

5 Conclusion

Existing saliency detection datasets mostly consist of well-lit images which make models trained on these datasets unsuitable for saliency detection under low-light conditions. Alleviating this problem generally involves using GAN based models which are computationally expensive and difficult to train. We thus proposed a classical computer vision method to generate proxy low-light images from well-lit images which can be used to train models for saliency estimation under real low-light conditions. We used band-pass filtering in the Fourier domain for translating well-lit images into proxy low-light images. During filtering, we ensured a smooth fusion of frequencies which suppressed the ringing artifacts. Our method has only a few hyper-parameters and is thus easy to generalize for different computer vision applications such as depth estimation. Specifically, we showed that models trained on our proxy low-light images outperformed existing low-light image translation methods for saliency and depth estimation under real low-light conditions.

Acknowledgments

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 1597–1604.

[2] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. 2020. High-resolution daytime translation without domain labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7488–7497.

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. AdaBins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4009–4018.

[4] Jianrui Cai, Shuhang Gu, and Lei Zhang. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing* 27, 4 (2018), 2049–2062.

[5] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Xiaojie Guo, and Qi Tian. 2015. Saliency-aware nonparametric foreground annotation based on weakly labeled data. *IEEE transactions on neural networks and learning systems* 27, 6 (2015), 1253–1265.

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3291–3300.

[7] Bo Dong, Yan Zhou, Chuanfei Hu, Keren Fu, and Geng Chen. 2021. BCNet: Bidirectional collaboration network for edge-guided salient object detection. *Neurocomputing* 437 (2021), 58–71.

[8] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*. 4548–4557.

[9] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421* (2018).

[10] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. 2020. Highly efficient salient object detection with 100k parameters. In *European Conference on Computer Vision*. Springer, 702–721.

[11] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. 2011. Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence* 34, 10 (2011), 1915–1926.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[13] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1780–1789.

[14] Xiaojie Guo, Yu Li, and Haibin Ling. 2016. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing* 26, 2 (2016), 982–993.

[15] Jianming Hu, Xiyang Zhi, Wei Zhang, Longfei Ren, and Lorenzo Bruzzone. 2020. Salient Ship Detection via Background Prior and Foreground Constraint in Remote Sensing Images. *Remote Sensing* 12, 20 (2020), 3370.

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

[17] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20, 11 (1998), 1254–1259.

[18] Hugo Jacob, Flávio LC Pádua, Anisio Lacerda, and Adriano Pereira. 2017. A video summarization approach based on the emulation of bottom-up mechanisms of visual attention. *Journal of Intelligent Information Systems* 49, 2 (2017), 193–211.

[19] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing* 30 (2021), 2340–2349.

[20] Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim. 2014. Salient region detection via high-dimensional color transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 883–890.

[21] Gábor Kovács, Yasuharu Kunii, Takao Maeda, and Hideki Hashimoto. 2019. Saliency and spatial information-based landmark selection for mobile robot navigation in natural environments. *Advanced Robotics* 33, 10 (2019), 520–535.

[22] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10285–10295.

[23] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*. Springer, 702–716.

[24] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3089–3098.

[25] Yuen Peng Loh and Chee Seng Chan. 2019. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding* 178 (2019), 30–42.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[27] Nan Mu, Xin Xu, and Xiaolong Zhang. 2019. Salient object detection in low contrast images via global convolution and boundary refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[28] Nan Mu, Xin Xu, Xiaolong Zhang, and Hong Zhang. 2018. Salient object detection using a covariance-based CNN model in low-contrast images. *Neural Computing and Applications* 29, 8 (2018), 181–192.

[29] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

[30] Mark Nixon and Alberto Aguado. 2019. *Feature extraction and image processing for computer vision*. Academic press.

[31] Alan V Oppenheim, John R Buck, and Ronald W Schafer. 2001. *Discrete-time signal processing. Vol. 2*. Upper Saddle River, NJ: Prentice Hall.

[32] Alan V Oppenheim and Jae S Lim. 1981. The importance of phase in signals. *Proc. IEEE* 69, 5 (1981), 529–541.

[33] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*. Springer, 319–345.

[34] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. 2014. Rgbd salient object detection: a benchmark and algorithms. In *European conference on computer vision*. Springer, 92–109.

[35] Fabio Pizzati, Pietro Cerri, and Raoul de Charette. 2021. CoMoGAN: continuous model-guided image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14288–14298.

[36] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7479–7489.

[37] Han Wang, Chen Wang, and Lihua Xie. 2020. Online visual place recognition via saliency re-identification. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5030–5036.

[38] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. 2019. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6849–6857.

[39] Xuanhan Wang, Lianli Gao, Jingkuan Song, and Hengtao Shen. 2016. Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition. *IEEE Signal Processing Letters* 24, 4 (2016), 510–514.

[40] Xiaofang Wang and Chun Qi. 2020. Detecting action-relevant regions for action recognition using a three-stage saliency detection technique. *Multimedia Tools and Applications* 79, 11 (2020), 7413–7433.

[41] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560* (2018).

[42] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560* (2018).

[43] Xin Xu, Nan Mu, Hong Zhang, and Xiaowei Fu. 2015. Salient object detection from distinctive features in low contrast images. In *2015 IEEE international conference on image processing (ICIP)*. IEEE, 3126–3130.

[44] Xin Xu and Jie Wang. 2018. Extended non-local feature for visual saliency detection in low contrast images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.

[45] Xin Xu, Shiqin Wang, Zheng Wang, Xiaolong Zhang, and Ruimin Hu. 2020. Exploring Image Enhancement for Salient Object Detection in Low Light Images. *arXiv preprint arXiv:2007.16124* (2020).

[46] Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4085–4095.

[47] Libao Zhang and Qiaoyue Sun. 2018. Saliency detection and region of interest extraction based on multi-image common saliency analysis in satellite images. *Neurocomputing* 283 (2018), 150–165.