DEEP COMPLEX SPATIO-SPECTRAL NETWORKS WITH COMPLEX VISUAL INPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex-valued neural networks have attracted growing attention for their ability to handle complex-valued data with enhanced representational capacity. However, their potential in computer vision remains relatively untapped. In this paper, we introduce Deep Complex Spatio-Spectral Network (DCSNet), a fully complex-valued token-based, end-to-end neural network designed for foreground extraction tasks. Additionally, our DCSNet encoder can be used for image classification. We also propose an invertible real-to-complex (R2C) transform, which generates two complex-valued input channels, complex intensity and complex hue, while producing complex-valued images with distinct real and imaginary components. DCSNet operates in both spatial and spectral domains by leveraging complex-valued inputs and complex Fourier transform. As a result, the complex-valued representation is maintained throughout DCSNet, and we avoid the information loss typically associated with Real↔Complex transformations. Extensive experiments show that DCSNet surpasses existing complex-valued methods across various tasks on both real and complex-valued data and achieves competitive performance compared to state-of-the-art real-valued methods, establishing a robust framework for handling both data types effectively.

1 INTRODUCTION

In the evolving landscape of deep learning, complex-valued neural networks (iCNNs) have shown great potential by enabling richer and more expressive representations. Despite their promise, iCNNs remain relatively underinvestigated in addressing key computer vision challenges. Addressing problems, such as image classification and foreground extraction, necessitates a prominent understanding of the global and local context in the input. Despite complex-valued networks showing promising outcomes Löwe et al.; Stanic et al. (2023); Singhal et al. (2022), the application of complex-valued networks remains understudied due to a lack of complex-valued input and sub-optimal complex-valued architectures. When tackling these issues, we are faced with a few challenges that need to be addressed: (i) A suitable way is required to have complex-valued inputs from real-valued RGB images. (ii) No prior work has shown promising results while maintaining the complex-valued nature of the input throughout different tasks. (iii) For foreground extraction, there is no objective function to handle complex-valued output.

Despite the challenges at hand, our motivation to explore complex-valued representation for computer vision tasks stems from the success demonstrated in diverse domains where complex-valued inputs are readily available, such as MRI Cole et al. (2021); Vasudeva et al. (2022), radar signals Gao et al. (2018); Georgiou & Koutsougeras (1992), and audio signals Hayakawa et al. (2018); Hu et al. (2020). Moreover, studies conducted in Arjovsky et al. (2016); Danihelka et al. (2016); Jojoa et al. (2022); Löwe et al., highlight the superior performance of complex-valued neural networks (iCNNs) compared to their real-valued counterparts. Additionally, iCNNs exhibit biological inspiration Reichert & Serre (2014) and greater generalization capacity Hirose & Yoshida (2012). Recent works Stanic et al. (2023); Halimeh & Kellermann (2022) also highlight the importance of complex-valued representation in images as well as audio, further fueling our exploration of iCNNs for computer vision tasks.

More specifically, we develop a complex-valued color transform R2C (real-to-complex), which converts real-valued images to complex-valued ones. We observe that any color vector in RGB space

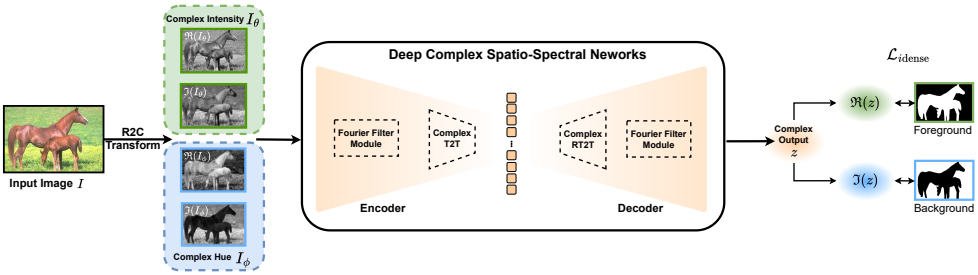


Figure 1: We introduce a novel token-based Deep Complex Spatio-Spectral Networks (DCSNet), which leverage (i) complex R2C transform for producing complex-valued inputs, (ii) Fourier Filter module for capturing global context using complex Fourier transform, (iii) Complex T2T for progressively reducing tokens and its inverse Complex RT2T, and (iv) our novel complex-valued objective function $\mathcal{L}_{idense}$ to optimize complex-valued output for foreground extraction.

can be projected to a grayscale line. The shortest angle between the two can provide us with the projection of color vector (on grayscale line) and the deflection ($\perp$ to grayscale line). Interestingly, the projection vector and the deflection vector are orthogonal, which creates an argand plane representing a complex number with the projection as the real part and deflection as the imaginary part. Similarly, we observe that the plane orthogonal to the grayscale line containing the color point can be considered an argand plane. Using the polar form, we can locate the point as a complex number using the perpendicular distance of the color point from grayscale line and a reference angle. Following this, we can generate two complex-valued representations for any color.

Furthermore, we introduce DCSNet, a novel end-to-end deep complex spatio-spectral network designed to operate on complex-valued images generated through the R2C transformation. It is first token-based approach utilizing complex-valued representation. The core component of DCSNet is the Fourier Filter Module, which transforms complex-valued tokens from the spatial domain to the frequency domain, applies a learnable Fourier filter, and subsequently maps the filtered results back to the spatial domain. Note that complex Fourier transform provides positive and negative frequency representation of complex-valued input. This architecture enables the network to capture both spatial and spectral domain information effectively, leading to preserved complex-valued representations. For foreground extraction tasks, we further propose a novel objective function $\mathcal{L}_{idense}$ that optimizes the separation of foreground and background in the predicted complex-valued outputs. This is achieved by decomposing the output into real and imaginary components, enabling more effective supervision and improved performance in handling complex-valued predictions. We give an overview of our proposed approach in Fig. 1. Our experimental analysis reveals that encoder of DCSNet beats all other complex-valued methods on large-scale real-valued dataset (ImageNet) and complex-valued datasets while maintaining a significantly lesser number of parameters. Moreover, DCSNet outperforms other complex-valued methods on multiple foreground extraction tasks, including salient object detection and blur detection. It highlights the efficiency and versatility of DCSNet for both classification and foreground extraction tasks on both complex and real-valued datasets.

Our contributions in this paper are as follows: (i) We propose R2C (real-to-complex), a novel complex-valued color transformation. (ii) We propose first token-based complex-valued network, DCSNet, which maintains the complex-valued information throughout. (iii) We propose a loss minimization strategy to handle complex-valued dense outputs. (iv) Our experimental analysis shows that our approach significantly improves for real and complex-valued data over existing methods across multiple tasks.

2 Related Work

Complex-valued Deep Learning: The integration of complex numbers as weights in deep learning introduces novel possibilities for delving into two-dimensional spectra, as highlighted in previous works Tygert et al. (2016); Hirose & Yoshida (2011). Moreover, the significance of phase information in the firing rate of neurons is underscored by studies such as Reichert & Serre (2014) and Jiang et al. (2019), emphasizing the potential advantages of employing complex-valued representa-

tions in neural networks. Specifically, the observed behavior of synchronized neurons with similar phases firing together, contrasted with asynchronous neurons with differing phases causing interference, bears a closer resemblance to the dynamics of biological neurons. This synchronization of inputs through neurons draws parallels to the gating mechanism found in both deep feedforward and recurrent neural networks Srivastava et al. (2015); Van den Oord et al. (2016); Kim & Adalı (2003).

Recent studies have showcased the superior generalization capacity of complex-valued networks, as evidenced by prior research Jojoa et al. (2022); Hirose & Yoshida (2012); Singhal et al. (2022). Notably, complex-valued autoencoders have outperformed slot-attention in the domain of object-centric learning Löwe et al. (2022); Stanic et al. (2023). Moreover, the application of complex-valued networks has proven beneficial in diverse areas, including saliency prediction Jiang et al. (2019; 2020) and iris recognition Nguyen et al. (2022). The findings presented in Cheung et al. (2019) further support the notion that employing a complex-valued vector can enhance the learning process for addressing multiple tasks.

Fourier Transform in Vision: The application of Fourier transform has been a cornerstone in digital image processing for decades, as acknowledged by seminal works in the field Gonzalez (2009); Pitas (2000). With the advent of Convolutional Neural Networks (CNNs) revolutionizing vision tasks He et al. (2016); Krizhevsky et al. (2012), there is a growing body of research integrating Fourier transform into deep learning methodologies Ding et al. (2017); Lee et al. (2018); Li et al. (2020); Yang & Soatto (2020). Some approaches employ discrete Fourier transform to transition images into the frequency domain, leveraging frequency information to enhance task performance Coates et al. (2011); Yang & Soatto (2020). Others exploit the convolution theorem, employing fast Fourier transform (FFT) to accelerate CNNs Ding et al. (2017); Li et al. (2020). In this study, we introduce a novel methodology using learnable Fourier filters to learn the global context in the Fourier domain, drawing inspiration from frequency filters in digital image processing Pitas (2000). Additionally, we capitalize on specific properties of FFT to reduce computational costs and parameter count.

3 Proposed Method

3.1 R2C: Complex-valued color transform

In order to obtain complex-valued images from a real-valued one, our goal is to define a transformation $T : \mathbb{R}^{d_1} \rightarrow \mathbb{C}^{d_2}$, where d_1 & d_2 are dimensions of the input and output respectively. Since Real-valued images are typically in RGB format, we have $d_1 = H \times W \times 3$, where H, and W are the height and width of the image. So, our target transformation function becomes $T : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{C}^{H \times W \times k}$, where k is the number of complex channels in the complex image.

Defining the transformation function on the pixel level is relatively more straightforward. Let us consider $\hat{T}(p) : \mathbb{R}^3 \rightarrow \mathbb{C}^k, \forall p \in I_{RGB}$. If we find k for each pixel p, we will ultimately have our transformation T. Each pixel p in the image I_{RGB} has its corresponding R, G & B values: $I_{RGB}(p) = \{I_r(p), I_g(p), I_b(p)\}$. Note that all the pixels in I_{RGB} image are located in the three-dimensional RGB space. Given their corresponding $I_r, I_g \& I_b$ values, one can quickly locate them in this space. We use properties of RGB space and simple linear algebra to create a complex-valued representation of the image.

In RGB space (Fig. 2), we consider an isotropic vector O (grayscale line) passing through the origin C and making equal angles with each axis. For a given pixel p in this space, we have a plane P, which has O as a normal vector intersecting at point E and contains pixel p at point F. Also, The plane P intersects the red, blue, and green axes at B, A, and D. Let us try to determine k for pixel p.

3.1.1 Complex Intensity Channel (I_θ):

It is crucial to notice that pixels lying on the vector O will represent grayscale color since $I_r(p) = I_g(p) = I_b(p)$. If we project other pixels on vector O, we can obtain a projected grayscale version of the image I_{rgb}.

Taking this observation into account, in Fig 2, we see that vector $v = \overrightarrow{CF}$ for pixel p makes an angle $\theta = \angle ECF$. v has two orthogonal components in the direction of $\overrightarrow{CE}$(projection) &

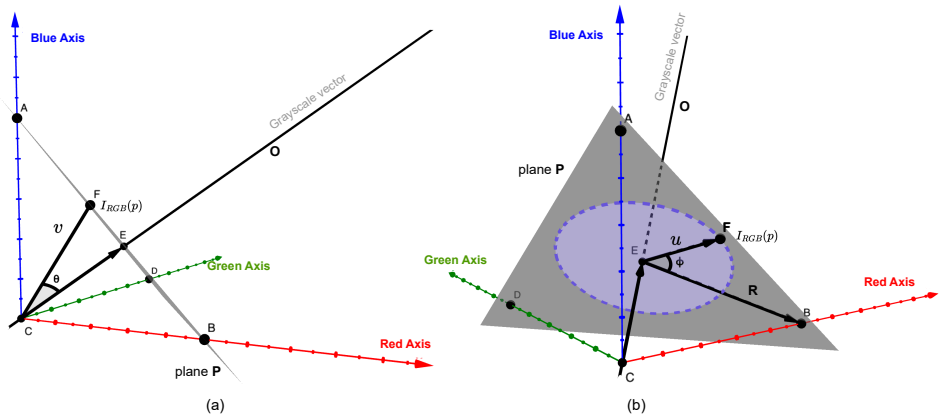


Figure 2: Proposed R2C color transformation: Given an RGB color $I_{RGB}(p) = \{I_R(p), I_G(p), I_B(p)\}$ for a pixel p. We have vectors $v = \overrightarrow{CF}$ & $u = \overrightarrow{EF}$. (a) Using θ, we project v onto the grayscale vector O, giving us projection $||v|| \cos\theta$ and deflection $||v|| \sin\theta$, which gives us the real and imaginary components of the first complex number. (b) Using ϕ between the reference vector R and vector u. In the argand plane P (assuming R as the real axis), we locate p, giving us the second complex number.

$\overrightarrow{EF}$(deflection). Using θ and $||v||$, we can decompose v into its two orthogonal components, forming real and imaginary components of a complex value. Real component of this complex value I_θ being $||v|| \cos\theta$ and imaginary component being $||v|| \sin\theta$. It results in the first complex value of pixel p as follows:

$$I_\theta(p) = ||v||e^{i\theta} = ||v|| \cos\theta + i||v|| \sin\theta \tag{1}$$

Here, $||.||$ represents the norm of the vector.

3.1.2 Complex Hue Channel (I_ϕ):

Note that in Fig 2, $u = \overrightarrow{EF}$ for pixel p lies in a plane that is $\perp$ to O, denoted as P. Assuming P as an argand plane, we can locate p using a complex number. For this, we need a reference axis. In Fig 2, we take this as a reference vector $R = \overrightarrow{EB}$, from E to the intersection of plane P and the red axis, given as B. Here, we assume P as an argand plane with R as the real axis and $R^\perp$ as the imaginary axis.

If we find the $\angle FEB = \phi$, we can easily locate p in this argand plane using polar coordinates $(||u||, \phi)$. For computing ϕ, we first find the shortest angle ϕ' between u and R using $\cos\phi' = \frac{u.R}{||u||||R||}$. Now we define ϕ as:

$$\phi = \begin{cases} \phi' & \text{if } I_b \geq I_g \\ 2\pi - \phi' & \text{else} \end{cases} \tag{2}$$

Now, using the polar coordinate, we can easily locate p in argand plane P. This leads to the second complex value of p as:

$$I_\phi(p) = ||u||e^{i\phi} = ||u|| \cos\phi + i||u|| \sin\phi \tag{3}$$

3.1.3 iRGB input:

From above, we found that $k = 2$ and established $\hat{T}(p) = \mathbb{R}^3 \rightarrow \mathbb{C}^{k=2}, \forall p \in I_{rgb}$. This results in a complex transformation $T : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{C}^{H \times W \times 2}$ such that $T(I_{RGB}) = I_{iRGB}$. Using T, we can convert our real-valued image I_{RGB} to complex-valued image I_{iRGB} with two complex-valued color channels. We use that to form the set required for the complex representation of $I_{iRGB}(p)$ as follows:

$$I_{iRGB}(p) = \{I_\theta(p), I_\phi(p)\} \tag{4}$$

Algorithm 1 R2C Color Transformation

Input: A Real-valued RGB Image, I_{RGB}
Output: A Complex-valued transformed Image, I_{iRGB}
for $p \in I_{RGB}$ **do** ▷ for each pixel in the image
  ▷ Find angle between pixel vector v and vector O.
  $\cos\theta = \frac{O \cdot v}{||O|| \times ||v||}$
  $\sin\theta = \sqrt{1 - \cos\theta}$
  $I_\theta(p) = ||v||(\cos\theta + i\sin\theta)$ ▷ Intensity channel
  ▷ Find angle between between vector u and vector R.
  $\phi' = \cos^{-1}(\frac{R \cdot u}{||R|| \times ||u||})$
  $I_{hsv} = rgb2hsv(I_{rgb})$
  if $\sin(hue(I_{hsv})) \geq 0$ **then**
    $\phi = \phi'$
  else
    $\phi = 2\pi - \phi'$
  end if
  $I_\phi(p) = ||u||(\cos\phi + i\sin\phi)$ ▷ Color channel
end for
return $I_{iRGB} = \{I_\theta, I_\phi\}$ ▷ Final complex image

We can separate complex-valued image I_{iRGB} into its real and imaginary component as: $I_{iRGB} = I_{re} + iI_{im}$, where $I_{re} = \{||v||\cos\theta, ||u||\cos\phi\}$ and $I_{im} = \{||v||\sin\theta, ||u||\sin\phi\}$. Algorithm 1 presents the algorithm for the above-presented R2C transform. Note that R2C is also invertible; we provide a detailed explanation for inverse R2C in Appendix A. Moreover, we considered using the Fourier transform to get complex input, but the complex input generated from the Fourier transform proved unsuitable. We present empirical results and discuss them in Appendix B.

3.2 DCSNet: Deep Complex Spatio-Spectral Networks

Recent advances in transformers Dosovitskiy et al. (2021); Yuan et al. (2021) demonstrate that self-attention based models can achieve good performances in solving various tasks. Following this approach, complex-valued transformer-based approaches Eilers & Jiang (2023); Yang et al. (2020); Dong et al. (2021) try to utilize self-attention. However, due to the nature of complex domain, they suffer from increased computation and poor results for image-related tasks. The proposed DCSNet architecture removes self-attention in favor of Fourier filters, enabling the model to retain information entirely within the complex domain while preserving global contextual information. As illustrated in Figure 3, DCSNet takes a complex input of size $H \times W$, divides it into patches, and unfolds these patches into tokens. The token length is progressively reduced in the spatial domain, while the Fourier filter operates in the spatial frequency domain. For classification tasks, the encoder is utilized directly, whereas for foreground extraction tasks, the process is reversed in the decoder to progressively increase the token length.

3.3 DCSNet Encoder

We propose an encoder part that learns to generate image embeddings that can be used to generate foreground extraction output. However, unlike previous methods, we cannot use existing pre-trained encoders because we propose a new architecture that utilizes complex-valued information end-to-end. Hence, we also train DCSNet encoder on image classification, more details in experiments.

Initially, we have our complex input image $I \in \mathbb{C}^{H \times W \times 2}$. We apply complex convolution and generate complex-valued patches of size $= \frac{H}{4} \times \frac{W}{4}$. We introduce a Complex T2T module, which reduces the number of complex tokens, and a Fourier filter module, which acts as a global convolution operation in the frequency domain while maintaining the complex-valued nature of the input.

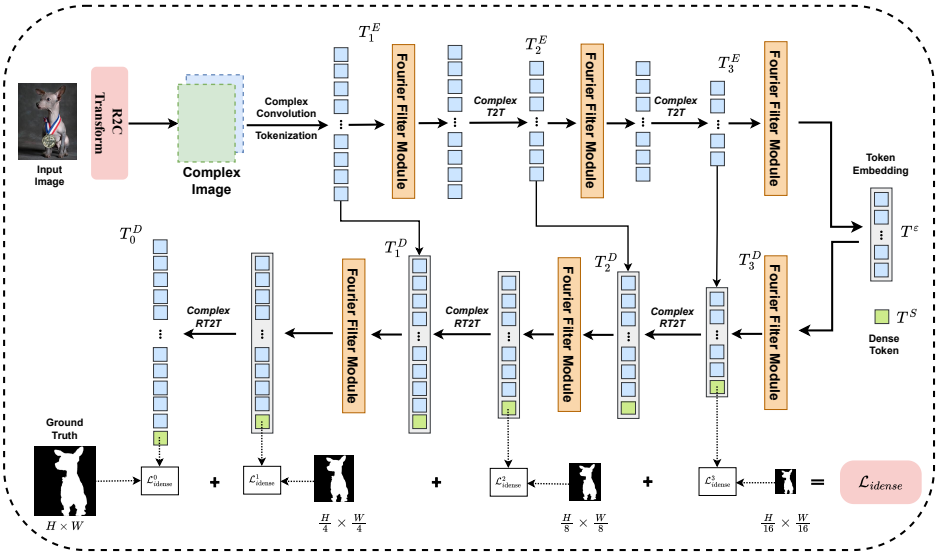


Figure 3: The overall architecture of our proposed DCSNet. It first encodes input image patch sequences to generate tokens to multiple resolutions (T_1^E, T_2^E, T_3^E), using Complex T2T. Then, we add a saliency token (T^S) to the output (T^ε) of the encoder. Finally, the decoder progressively upsamples the tokens using Complex RT2T while predicting saliency map at each step. We also optimize the generated map at each step using our proposed loss to improve the predicted map.

3.3.1 Fourier Filter Module

We propose a Fourier filter module consisting of a Fourier filter and a complex MLP layer. Before each layer, we apply layer normalization. In the Fourier filter layer, given the input feature $x \in \mathbb{C}^{H \times W \times D}$, we first perform 2D DFT (see Appendix C) along the spatial dimension to convert x to the frequency domain $X = F[x] \in \mathbb{C}^{H \times W \times D}$, where $F[\cdot]$ denotes 2D complex DFT. The output X of DFT is a complex tensor and represents x in the frequency domain. We can now apply a learnable filter $K \in \mathbb{C}^{H \times W \times D}$ to X by simple element-wise multiplication.

$$\hat{X} = K \odot X \tag{5}$$

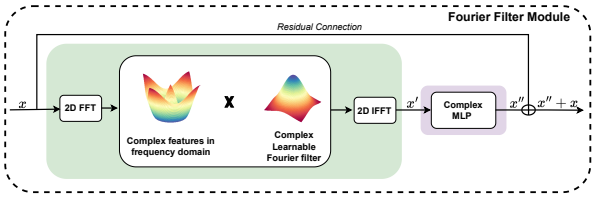


Figure 4: Fourier Filter Module takes a complex-valued input, applies a learnable complex filter in the frequency domain, and gives a complex-valued output.

where $\odot$ denotes element-wise multiplication. Since K has the same dimension as X, it will act as a global filter in the frequency domain. Finally, we use inverse DFT to transform back to the spatial domain as $x' = F^{-1}[\hat{X}]$. We illustrate this process in Fig. 4. Fourier filter in the frequency domain is equivalent to a convolution with filter size $H \times W$ in the spatial domain. Hence, unlike convolution in the spatial domain, which focuses on local features due to the small filter size, the Fourier filter module focuses on the global context as shown by Rao et al. (2021). Moreover, keeping the operation in a complex domain helps us preserve additional complex information and both positive and negative frequencies in the spatial frequency domain.

3.3.2 Complex T2T Module

Given a sequence of patch tokens T with length l from the previous layer, following T2T-ViT in Yuan et al. (2021), we introduce and iteratively apply the Complex T2T module (Fig. 5), which is composed of a re-structurization step and a soft split step, to model the local structure information in T and obtain a new sequence of tokens.

This module consists of a structurization and a de-structurization step. In structurization step, $T \in \mathbb{C}^{l \times c}$ is reshaped to a 2D image $I \in \mathbb{C}^{h \times w \times c}$, where $l = h \times w$, to recover spatial structures. After the structurization step, we apply the de-structurization step, where I is first split into $k \times k$ patches with s overlapping. p zero-padding is also utilized to pad image boundaries. Then, the

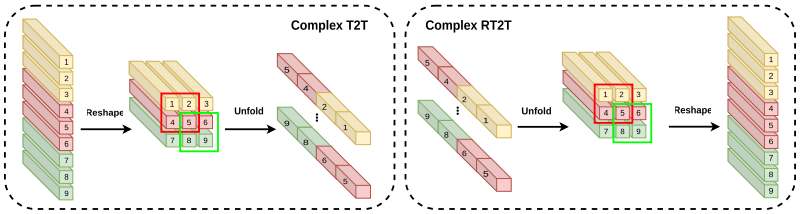


Figure 5: Complex T2T and Complex RT2T reduce and increase the number of tokens, respectively, while maintaining the complex nature of tokens.

image patches are unfolded to a sequence of tokens $T' \in \mathbb{C}^{l' \times ck^2}$, where the sequence length l' is computed as $l' = h' \times w' = \lfloor \frac{h+2p-k}{k-s} + 1 \rfloor \times \lfloor \frac{w+2p-k}{k-s} + 1 \rfloor$.

3.4 DCSNet Decoder

As shown in Fig. 3, after obtaining complex feature embeddings T^{ε} , we want to predict the respective salient masks $M \in \{0,1\}^{H \times W \times 1}$ for each input I. The encoded length of T^{ε} is quite small $l^{\varepsilon} = [\frac{H}{16}, \frac{W}{16}]$. We use the reverse version of the Complex T2T module, Complex RT2T. Complex RT2T upsamples each token into multiple subtokens while maintaining common information between them. We leverage low-level tokens from the encoder to provide additional information simultaneously. Progressively, we upsample embedded tokens T^{ε} and add information from T_1 and T_2 using concatenation and complex linear projection.

3.4.1 Complex RT2T module

In order to upsample the encoded information in tokens, we introduce the reverse version of the complex T2T module. Specifically, we first project the input patch tokens to reduce their embedding dimension from $d = 384$ to $c = 64$. Then, we use another complex linear projection to expand the embedding dimension from c to ck^2. Next, similar to the de-structurization step in complex T2T, each token is seen as a $k \times k$ image patch, and neighboring patches have s overlapping. Then, we can fold the tokens as an image using p zero-padding. Finally, we reshape the image to match the upsampled tokens, as shown in Fig. 5.

3.4.2 Token-based mask generation

Inspired by existing transformer architecture Dosovitskiy et al. (2021); Yuan et al. (2021); Eilers & Jiang (2023); Liu et al. (2021), we add a dense token as shown in Fig. 3 In doing so, we design a complex-valued dense token $T^S \in \mathbb{C}^{1 \times d}$, where d is the embedding dimension. At each level in the decoder, we add T^S to the encoded patch tokens $T_i^D, i \in \{0,1,2,3\}$. When the modified tokens are processed through the Fourier filter module, the dense token learns dense information from an image by progressively extracting feature information from the tokens. For foreground extraction, we do not use any self-attention module. We again send our obtained tokens T_1^D to a Fourier filter module, and the third complex RT2T module upsamples the tokens from $1/4$ to full resolution.

3.5 Objective function

Since our architecture is complex-valued, we must ensure that the loss function can handle the complex output. For classification tasks, we employ the complex loss function proposed by Yadav & Jerripothula (2023). However, existing loss functions cannot handle complex outputs directly for foreground extraction. To tackle this problem, we propose a modified binary cross entropy loss $\mathcal{L}_{\text{idense}}$. We obtain a complex-valued output z_j from our network for j^{th} input. Given the real-valued ground truth y_j, we construct a complex-valued ground truth by

considering the foreground dense map as the real part and the background dense map as the imaginary part, as shown in Fig. 6. It turns out to be a complex-valued map with the foreground pixel having value = 1 and the background pixel having value = i, just like real-valued ground-truth dense maps have foreground pixel value = 1 and background pixel value = 0. we develop a loss $\mathcal{L}_{\text{idense}}$ which can minimize the comlpex-valued output of our network z_j with help of complex-valued ground truth saliency map $y_j + i(1 - y_j)$. We formulate our $\mathcal{L}_{\text{idense}}$ as follows:

$$\mathcal{L}_{i\text{dense}}(z_j) = \sum_{j=1}^{N} \sum_{k=0}^{3} \mathcal{L}_{BCE}(y_j^k, \Re(z_j^k)) + \mathcal{L}_{BCE}((1 - y_j^k), \Im(z_j^k)) \tag{6}$$

where $\mathcal{L}_{BCE}$ is binary cross entropy loss, k indicates token levels T_k^D in decoder, N is total number of input images, and $\Re$ and $\Im$ are real and imaginary components of complex output z_j. To generate the final binary mask from complex-valued output z_j, we take the average of real and complement of the imaginary component as $mask = (\Re(z_j) + (1 - \Im(z_j)))/2$.

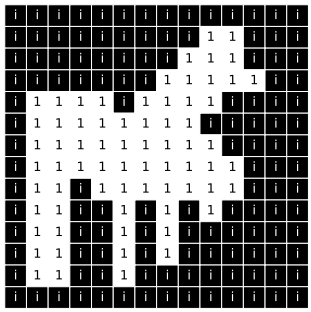


Figure 6: Complex grountruth, foreground pixels are denoted by 1, and background pixels as 0

4 EXPERIMENTS

In this section, we present the evaluation of our proposed Deep Complex Spatio-Spectral Networks (DCSNet) and compare their performance with other methods (both real and complex-valued) for multiple task.

4.1 DATASETS & METHODS

We conduct experiments on both classification task and foreground extraction tasks. For both types of tasks, we present results on both real-valued and complex-valued datasets.

Real-valued Datasets: For classification, we evaluate our model on the large-scale real-valued ImageNet dataset Deng et al. (2009) and conduct additional ablation studies on the CIFAR10 dataset Krizhevsky et al.. Accuracy is used as the evaluation metric for comparison with existing complex-valued methods. For foreground extraction, we assess the performance of DCSNet on salient object detection, defocus blur detection, and shadow detection using multiple datasets. Furthermore, we provide a comprehensive comparison of real-valued and complex-valued methods across these tasks.

We follow the approach outlined in Liu et al. (2021) for comparison with real-valued methods on salient object detection. We evaluate our model on five widely-used benchmark datasets: ECSSD Yan et al. (2013) (1,000 images), PASCAL-S Li et al. (2014) (850 images), HKU-IS Li & Yu (2015) (4,447 images), DUT-O Yang et al. (2013) (5,168 images), and DUTS Wang et al. (2017) (10,553 images). We use four commonly employed evaluation metrics across these datasets: S_m Fan et al. (2017), maxF (maximum F-measure), E_ξ^{max} Fan et al. (2018) (maximum enhanced-alignment measure), and MAE (mean absolute error).

We use the CUHKShi et al. (2014) and DUT Zhao et al. (2018) datasets for defocus blur detection. The CUHK dataset contains 704 defocus images, while DUT comprises 1,100 images. Our method is trained on 604 images from CUHK and 600 from DUT. We evaluate the remaining test images from both datasets using F-measure (F_β) and mean absolute error (MAE) as evaluation metrics. For shadow detection, we assess our model on the SBU Vicente et al. (2016) and ISTD Wang et al. (2018) datasets. We train the model separately on each dataset and report the performance using the balanced error rate (BER).

Complex-valued Datasets: For the classification task, we evaluate our model on two complex-valued datasets: MSTAR Ross et al. (1998) and S1SLC_CVDL Mohammadi Asiyabi et al. (2023), using accuracy as the evaluation metric for comparison. For foreground extraction, we utilize the Simulated InSAR Building dataset Chen (2020) for building detection, treating the layover class as the foreground. The dataset consists of 270 training images and 42 test images. We use mean Intersection over Union (mIoU) as the evaluation metric and compare our results with other complex-valued methods.

Table 1: Comparing classification accuracy(%) of our encoder with only other complex-valued method to show results on Real-valued (ImageNet-1kDeng et al. (2009)) and Complex-valued (MSTARRoss et al. (1998), S1SLC_CVDLMohammadi Asiyabi et al. (2023)) datasets.

(a) Real-valued Dataset

Model	Top-1 Acc. (%)
ResNet50He et al. (2016)	76.1
Vit-S/16Dosovitskiy et al. (2021)	78.1
GFNet-xsRao et al. (2023)	78.6
ResMLP-12Touvron et al. (2022)	76.6
ResNet152 (DCN)Trabelsi et al. (2018)	72.6
ResNet152 (FCCN)Yadav & Jerripothula (2023)	77.3
DCSNet Encoder	**78.8**

(b) Complex-valued Dataset

Dataset	DCN	FCCN	DCSNet Encoder
MSTAR	96.1	97.4	**97.7**
S1SLC_CVDL	93.2	89.8	**91.6**

Table 2: Quantitative comparison of our proposed DCSNet with other real-valued RGB SOD methods on five benchmark datasets. "-R" and "-R2" means the ResNet50 and Res2Net backbone, respectively. The values in red are best, and the ones in blue are second best. Our DCSNet performs best or second best 55% of the time.

		DUTSWang et al. (2017)				ECSSDYan et al. (2013)				HKU-ISLi & Yu (2015)				PASCAL-SLi et al. (2014)				DUT-OYang et al. (2013)				Average			
Method	Param(M)	$S_m \uparrow$	maxF$\uparrow$	$E_\xi^{max} \uparrow$	MAE$\downarrow$	$S_m \uparrow$	maxF$\uparrow$	$E_\xi^{max} \uparrow$	MAE$\downarrow$	$S_m \uparrow$	maxF$\uparrow$	$E_\xi^{max} \uparrow$	MAE$\downarrow$	$S_m \uparrow$	maxF$\uparrow$	$E_\xi^{max} \uparrow$	MAE$\downarrow$	$S_m \uparrow$	maxF$\uparrow$	$E_\xi^{max} \uparrow$	MAE$\downarrow$	$S_m \uparrow$	maxF$\uparrow$	$E_\xi^{max} \uparrow$	MAE$\downarrow$
Real-valued																									
PiCANetLiu et al. (2018)	47.22	0.863	0.840	0.915	0.040	0.916	0.929	0.953	0.035	0.905	0.913	0.951	0.031	0.846	0.824	0.882	0.072	0.826	0.767	0.865	0.054	0.871	0.854	0.913	0.046
BASNetQin et al. (2019)	87.06	0.866	0.838	0.902	0.047	0.916	0.931	0.951	0.037	0.909	0.919	0.952	0.032	0.837	0.819	0.868	0.083	0.836	0.779	0.872	0.057	0.872	0.856	0.909	0.051
PoolNetLiu et al. (2019)	68.26	0.879	0.853	0.917	0.041	0.917	0.929	0.948	0.042	0.916	0.920	0.955	0.032	0.852	0.830	0.880	0.076	0.832	0.769	0.869	0.056	0.879	0.864	0.914	0.049
EGNet-RZhao et al. (2019a)	111.64	0.887	0.866	0.926	0.039	0.925	0.936	0.955	0.037	0.0918	0.923	0.956	0.031	0.852	0.825	0.874	0.080	**0.841**	0.778	0.878	**0.053**	0.886	0.868	0.918	0.048
MINet-RPang et al. (2020)	162.38	0.884	0.864	0.926	**0.037**	0.925	0.938	0.957	**0.034**	0.919	0.926	0.960	0.031	0.856	0.831	0.883	0.071	0.833	0.769	0.869	0.056	0.885	0.868	0.919	0.046
LDF-RWei et al. (2020)	25.15	0.892	**0.877**	0.930	**0.034**	0.925	0.938	0.954	**0.034**	0.920	**0.929**	0.958	**0.028**	0.861	**0.839**	0.888	**0.067**	0.839	**0.782**	0.879	**0.052**	0.892	0.876	0.921	**0.043**
CSF-R2Gao et al. (2020)	36.53	0.890	0.869	0.929	**0.037**	**0.931**	0.942	**0.960**	**0.033**	-	-	-	-	0.863	**0.839**	0.885	0.073	0.838	0.775	0.869	0.055	0.892	0.874	0.911	0.049
GateNet-RZhao et al. (2020)	128.63	0.891	**0.874**	0.932	0.038	0.924	0.935	0.955	0.038	**0.921**	0.926	0.959	0.031	0.863	0.836	0.886	0.071	0.840	**0.782**	0.878	0.055	0.888	0.874	0.922	0.047
VSTLiu et al. (2021)	44.48	**0.896**	**0.877**	**0.939**	**0.037**	**0.932**	**0.944**	**0.964**	**0.034**	**0.928**	**0.937**	**0.968**	0.030	**0.873**	**0.850**	**0.900**	**0.067**	**0.850**	**0.800**	**0.888**	0.058	**0.904**	**0.894**	**0.932**	0.045
Complex-valued																									
FCCNYadav & Jerripothula (2023)	48.61	0.811	0.750	0.872	0.070	0.874	0.876	0.918	0.082	0.867	0.857	0.919	0.058	0.807	0.783	0.854	0.100	0.787	0.695	0.827	0.079	0.824	0.813	0.878	0.078
SCVUNetWei et al. (2023)	54.15	0.824	0.769	0.874	0.063	0.882	0.887	0.922	0.062	0.877	0.872	0.929	0.052	0.815	0.795	0.865	0.089	0.788	0.697	0.875	0.077	0.831	0.827	0.893	0.069
DCSNet (ours)	39.12	**0.894**	**0.874**	**0.941**	0.039	0.927	**0.945**	**0.960**	**0.034**	0.917	0.924	**0.967**	**0.029**	**0.866**	**0.850**	**0.903**	**0.062**	0.839	0.776	**0.880**	0.056	**0.893**	**0.895**	**0.930**	**0.044**

We resize input images to 256×256 and then randomly crop 224×224 image regions as the model input and use random flipping as data augmentation. We randomly initialize weights, set the batch size to 8, and use Adam Kingma & Ba (2015) optimizer. For implementating complex-valued operations, we use the PyTorch library which provides native support for complex tensor and complex operations including convolution and complex FFT.

4.2 Results

Comparison on classification: Our proposed DCSNet encoder is the first token-based approach for complex-valued image classification. When comparing our encoder on image classification (Table 1a and 1b), we observe that DCSNet beats FCCNYadav & Jerripothula (2023) on both large-scale real-valued dataset (ImageNet) and complex-valued datasets (MSTARRoss et al. (1998)& S1SLC_CVDLMohammadi Asiyabi et al. (2023)). We provide additional comparisons with FCCN in Appendix E. We obtain significantly better results than Yadav & Jerripothula (2023) while maintaining fewer parameters, marking the best result for complex-valued image classification on large-scale datasets for both real and complex-valued datasets.

Comparison on foreground extraction: To show the applicability and efficiency of our proposed method, we also compare our DCSNet with 9 other real-valued salient object detection methods on five different datasets for a more extensive comparison with real-valued methods. We also present results and compare them with two recent complex-valued methods: FCCN & SCVUNet. For FCCN, we follow a CNN-based encoder-decoder approach for foreground extraction, while SCVUNet is directly utilized for the real-valued dataset. We present our results in Table 2, which shows that DCSNet performs either best or second best **55**% of the time. We also present qualitative results in Appendix F. Similarly, for both defocus blur detection, and shadow detection, we outperform existing real-valued and complex-valued methods. We present the comparison on both tasks in Tab. 3& 4.

When comparing complex-valued data for foreground extraction (Tab. 5), we see a similar pattern, i.e., DCSNet outperforms existing complex-valued methods decisively. Our proposed fully complex-valued method obtains results comparable to existing real-valued methods. It marks the first milestone for the application of complex-valued methods in foreground extraction tasks.

Table 3: Comparison with real and complex-valued methods on defocus blur detection. Red: best, Blue: second best.

Method	Param (M)	DUT		CUHK	
		$\mathcal{F}_\beta \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{F}_\beta \uparrow$	$\mathcal{M} \downarrow$
Real-valued					
DeFusionNetTang et al. (2019)	-	0.823	0.118	0.818	0.117
BTBNetZhao et al. (2018)	-	0.827	0.138	0.889	0.082
CENetZhao et al. (2019b)	-	0.817	0.135	0.906	0.059
DAD Zhao et al. (2021b)	44.13	0.794	0.153	0.884	0.079
EFENetZhao et al. (2021a)	43.61	**0.854**	**0.094**	**0.914**	**0.053**
Complex-valued					
FCCNYadav & Jerripothula (2023)	48.61	0.860	0.104	0.898	0.081
SCVUNetWei et al. (2023)	54.15	0.848	0.096	0.901	0.078
DCSNet (ours)	39.12	**0.894**	**0.058**	**0.907**	**0.045**

Table 4: Comparison with real and complex-valued approaches on shadow detection. Red: best, Blue: second best.

Method	Param (M)	ISTD BER ↓	SBU BER ↓
Real-valued			
Stacked CNNVicente et al. (2016)	-	8.60	-
BDRARZhu et al. (2018)	42.45	2.69	3.89
DSCHu et al. (2018)	122.49	3.42	5.59
DSDZheng et al. (2019)	58.15	2.17	3.45
MTMTChen et al. (2020)	44.12	1.72	3.15
FDRNet Zhu et al. (2021)	-	**1.55**	**3.04**
Complex-valued			
FCCNYadav & Jerripothula (2023)	48.61	1.78	3.22
SCVUNetWei et al. (2023)	54.15	1.91	3.25
DCSNet (ours)	39.12	**1.49**	**3.05**

Table 5: Comparison on Complex-valued Building InSAR datasetChen (2020) for foreground extraction. Red: best, Blue: second best.

Method	Param (M)	Building mIoU ↑
FCCNYadav & Jerripothula (2023)	48.61	0.82
SCVUNetWei et al. (2023)	54.15	**0.86**
DCSNet (ours)	39.12	**0.89**

4.3 Ablation Study

We conduct three ablation studies to highlight the importance of our contributions. The first two ablation studies are conducted on four benchmark datasets for SOD. The third study is conducted on CIFAR10 following Yadav & Jerripothula (2023) for image classification. In addition to our encoder, we take a smaller version with fewer parameters to observe performance variation. All the models for the ablation study are trained from scratch in order to maintain fairness.

Effects of complex input on foreground extraction: In this ablation study, we analyze the contribution of various inputs and channels in I_{iRGB} input by providing them one at a time. Results are shown in Table 6. Each I_{iRGB} channel performs well; however, we get the best results when both are taken together. Even while using other inputs to DCSNet, i.e., RGB and iHSVYadav & Jerripothula (2023), we observe that our complex input performs better for foreground extraction.

Effects of $\mathcal{L}_{\text{idense}}$ & T^S: To analyze the importance of dense loss $\mathcal{L}_{\text{idense}}$, we use binary cross entropy loss and only optimize the real component of complex-valued output. Similarly, to observe the importance of dense token T^S, we remove it from our model. We present the result of these ablations in Table 6.

Table 6: Results of ablation study that highlights the effect of both channels in I_{iRGB} and various other inputs. We also highlight the importance of our loss $\mathcal{L}_{idense}$, and dense token(T^S).

Dataset	DUTS				ECSSD				PASCAL-S				DUT-O			
	$S_m \uparrow$	maxF↑	$E_\xi^{max} \uparrow$	MAE↓	$S_m \uparrow$	maxF↑	$E_\xi^{max} \uparrow$	MAE↓	$S_m \uparrow$	maxF↑	$E_\xi^{max} \uparrow$	MAE↓	$S_m \uparrow$	maxF↑	$E_\xi^{max} \uparrow$	MAE↓
baseline (VST)	0.732	0.644	0.785	0.130	0.819	0.808	0.866	0.112	0.742	0.691	**0.866**	0.157	0.731	0.631	0.782	0.131
ours (RGB)	0.731	0.641	0.774	0.147	0.820	0.801	0.857	0.114	0.732	0.658	0.847	0.139	0.731	0.625	0.781	0.136
ours (iHSV)	0.735	0.648	0.780	0.128	0.824	0.813	0.864	0.105	0.742	0.710	0.813	0.128	0.738	0.631	0.789	0.112
ours (I_θ)	0.729	0.629	0.790	0.123	0.810	0.791	0.865	0.109	0.743	0.701	0.794	0.144	0.735	0.634	0.792	0.122
ours (I_ϕ)	0.726	0.631	0.791	0.120	0.805	0.788	0.866	0.112	0.745	0.692	0.789	0.139	0.729	0.636	0.785	0.119
ours (w/o $\mathcal{L}_{idense}$)	0.731	0.630	0.789	0.122	0.824	0.807	0.872	0.097	**0.747**	0.702	0.794	0.143	0.733	0.633	0.788	0.123
ours (w/o T^S)	0.734	0.649	0.787	0.113	0.825	0.818	0.878	0.091	0.749	0.709	0.703	0.138	0.730	0.646	0.792	0.114
ours (iRGB)	**0.740**	**0.654**	**0.801**	**0.107**	**0.831**	**0.825**	**0.884**	**0.083**	**0.747**	**0.713**	0.801	**0.131**	**0.739**	**0.645**	**0.795**	**0.109**

Table 7: Improvements over baseline two variants of GFNetRao et al. (2021) -xs and -ti. Here, we study the results of another complex-valued input, iHSVYadav & Jerripothula (2023), and the effect of each complex-valued channel in Image classification on the CIFAR-10 dataset.

Model	GFNet-xs		DCSNet Encoder				GFNet-ti		DCSNet Encoder-small			
Input	(RGB)	(HSV)	(iHSV)	(I_θ)	(I_ϕ)	(iRGB)	(RGB)	(HSV)	(iHSV)	(I_θ)	(I_ϕ)	(iRGB)
Acc (%)	93.6	93.3	93.5	92.7	91.4	**94.3**	92.1	91.9	92.2	91.7	90.3	**92.8**

Effects of complex input on image classification: This study observes the effect of various complex-valued inputs and each channel of I_{iRGB} for image classification. We compare these results with two variants of real-valued model GFNetRao et al. (2021). Our DCSNet encoders contain a similar number of parameters as variants of GFNet. From Table 7, we can observe the role of proposed complex-valued input.

5 Conclusion

In this work, we have presented DCSNet, a fully complex-valued token-based network for Foreground extraction and image classification, which operates both in spatial and frequency domain. It takes complex-valued input generated from our R2C transform and optimizes the complex-valued dense output using our proposed loss function. While maintaining complex-valued information throughout, our model outperforms previous complex-valued methods on various tasks and both real and complex-valued data, presenting a robust approach using complex-valued representation.

REFERENCES

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pp. 1120–1128. PMLR, 2016.

Ali Borji and Laurent Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *CoRR*, abs/1505.03581, 2015. URL `http://arxiv.org/abs/1505.03581`.

Jiankun Chen. Simulated insar building dataset for cvcmff net, 2020. URL `https://dx.doi.org/10.21227/2csm-3723`.

Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 5610–5619. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00565. URL `https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_A_Multi-Task_Mean_Teacher_for_Semi-Supervised_Shadow_Detection_CVPR_2020_paper.html`.

Brian Cheung, Alexander Terekhov, Yubei Chen, Pulkit Agrawal, and Bruno Olshausen. Superposition of many models into one. *Advances in neural information processing systems*, 32, 2019.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Elizabeth Cole, Joseph Cheng, John Pauly, and Shreyas Vasanawala. Analysis of deep complex-valued convolutional neural networks for mri reconstruction and phase-focused applications. *Magnetic resonance in medicine*, 86(2):1093–1109, 2021.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pp. 3488–3493. IEEE, 2016. doi: 10.1109/ICPR.2016.7900174. URL `https://doi.org/10.1109/ICPR.2016.7900174`.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Trans. Image Process.*, 27(10):5142–5154, 2018. doi: 10.1109/TIP.2018.2851672. URL `https://doi.org/10.1109/TIP.2018.2851672`.

Ivo Danihelka, Greg Wayne, Benigno Uria, Nal Kalchbrenner, and Alex Graves. Associative long short-term memory. In *International Conference on Machine Learning*, pp. 1986–1994. PMLR, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, Geng Yuan, Xiaolong Ma, Yipeng Zhang, Jian Tang, Qinru Qiu, Xue Lin, and Bo Yuan. Circnn: accelerating and compressing deep neural networks using block-circulant weight matrices. In Hillery C. Hunter, Jaime Moreno, Joel S. Emer, and Daniel Sánchez (eds.), *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2017, Cambridge, MA, USA, October 14-18, 2017*, pp. 395–408. ACM, 2017. doi: 10.1145/3123939.3124552. URL `https://doi.org/10.1145/3123939.3124552`.

Yihong Dong, Ying Peng, Muqiao Yang, Songtao Lu, and Qingjiang Shi. Signal transformer: Complex-valued attention and meta-learning for signal recognition. *CoRR*, abs/2106.04392, 2021. URL `https://arxiv.org/abs/2106.04392`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at

scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Florian Eilers and Xiaoyi Jiang. Building blocks for a complex-valued transformer architecture. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095349. URL https://doi.org/10.1109/ICASSP49357.2023.10095349.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 4558–4567. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.487. URL https://doi.org/10.1109/ICCV.2017.487.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In Jérôme Lang (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 698–704. ijcai.org, 2018. doi: 10.24963/IJCAI.2018/97. URL https://doi.org/10.24963/ijcai.2018/97.

Jingkun Gao, Bin Deng, Yuliang Qin, Hongqiang Wang, and Xiang Li. Enhanced radar imaging using a complex-valued convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 16(1):35–39, 2018.

Shanghua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pp. 702–721. Springer, 2020. doi: 10.1007/978-3-030-58539-6_42. URL https://doi.org/10.1007/978-3-030-58539-6_42.

George M Georgiou and Cris Koutsougeras. Complex domain backpropagation. *IEEE transactions on Circuits and systems II: analog and digital signal processing*, 39(5):330–334, 1992.

Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Process.*, 19(1):185–198, 2010. doi: 10.1109/TIP.2009.2030969. URL https://doi.org/10.1109/TIP.2009.2030969.

Mhd Modar Halimeh and Walter Kellermann. Complex-valued spatial autoencoders for multichannel speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pp. 261–265. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747528. URL https://doi.org/10.1109/ICASSP43922.2022.9747528.

Daichi Hayakawa, Takashi Masuko, and Hiroshi Fujimura. Applying complex-valued neural networks to acoustic modeling for speech recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1725–1731. IEEE, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Akira Hirose and Shotaro Yoshida. Comparison of complex-and real-valued feedforward neural networks in their generalization ability. In *International Conference on Neural Information Processing*, pp. 526–531. Springer, 2011.

Akira Hirose and Shotaro Yoshida. Generalization characteristics of complex-valued feedforward neural networks in relation to signal coherence. *IEEE Transactions on Neural Networks and Learning Systems*, 23(4):541–551, 2012. doi: 10.1109/TNNLS.2012.2183613.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. doi: 10.1109/CVPR.2007.383267. URL https://doi.org/10.1109/CVPR.2007.383267.

Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7454–7462. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00778. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Direction-Aware_Spatial_Context_CVPR_2018_paper.html.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. 2020.

Lai Jiang, Zhe Wang, Mai Xu, and Zulin Wang. Image saliency prediction in transformed domain: A deep complex neural network method. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 8521–8528. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33018521. URL https://doi.org/10.1609/aaai.v33i01.33018521.

Lai Jiang, Mai Xu, Shanyi Zhang, and Leonid Sigal. Deepct: A novel deep complex-valued network with learnable transform for video saliency prediction. *Pattern Recognition*, 102:107234, 2020.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1072–1080. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298710. URL https://doi.org/10.1109/CVPR.2015.7298710.

Mario Jojoa, Begonya Garcia-Zapirain, and Winston Percybrooks. A fair performance comparison between complex-valued and real-valued neural networks for disease detection. *Diagnostics*, 12 (8):1893, 2022.

Taehwan Kim and Tülay Adalı. Approximation by fully complex multilayer perceptrons. *Neural computation*, 15(7):1641–1666, 2003.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

A Krizhevsky, I Sutskever, and GE Hinton. 2012 alexnet, 2012.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Jaehan Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 330–339. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00042. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Lee_Single-Image_Depth_Estimation_CVPR_2018_paper.html.

Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 5455–5463. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299184. URL https://doi.org/10.1109/CVPR.2015.7299184.

Shaohua Li, Kaiping Xue, Bin Zhu, Chenkai Ding, Xindi Gao, David S. L. Wei, and Tao Wan. FALCON: A fourier transform based approach for fast and secure convolutional neural network predictions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 8702–8711. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00873. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Li_FALCON_A_Fourier_Transform_Based_Approach_for_Fast_and_Secure_CVPR_2020_paper.html.

Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 280–287. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.43. URL https://doi.org/10.1109/CVPR.2014.43.

Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3917–3926. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00404. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_A_Simple_Pooling-Based_Design_for_Real-Time_Salient_Object_Detection_CVPR_2019_paper.html.

Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3089–3098. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00326. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Liu_PiCANet_Learning_Pixel-Wise_CVPR_2018_paper.html.

Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 4702–4712. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00468. URL https://doi.org/10.1109/ICCV48922.2021.00468.

Sindy Löwe, Phillip Lippe, Maja Rudolph, and Max Welling. Complex-valued autoencoders for object discovery. *Transactions on Machine Learning Research*.

Sindy Löwe, Phillip Lippe, Maja Rudolph, and Max Welling. Complex-valued autoencoders for object discovery. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=1PfcmFTXoa.

Reza Mohammadi Asiyabi, Mihai Datcu, Andrei Anghel, and Holger Nies. S1slc_cvdl: A complex-valued annotated single look complex sentinel-1 sar dataset for complex-valued deep networks, 2023. URL https://dx.doi.org/10.21227/nm4g-yd98.

Kien Nguyen, Clinton Fookes, Sridha Sridharan, and Arun Ross. Complex-valued iris recognition network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):182–196, 2022.

Junting Pan, Cristian Canton-Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró-i-Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *CoRR*, abs/1701.01081, 2017. URL http://arxiv.org/abs/1701.01081.

Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9410–9419. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00943. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Pang_Multi-Scale_Interactive_Network_for_Salient_Object_Detection_CVPR_2020_paper.html.

Ioannis Pitas. *Digital image processing algorithms and applications*. John Wiley & Sons, 2000.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jägersand. Basnet: Boundary-aware salient object detection. In *IEEE Conference on Computer Vision*

and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 7479–7489. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00766. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Qin_BASNet_Boundary-Aware_Salient_Object_Detection_CVPR_2019_paper.html.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 980–993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/07e87c2f4fc7f7c96116d8e2a92790f5-Abstract.html.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jie Zhou, and Jiwen Lu. Gfnet: Global filter networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10960–10973, 2023. doi: 10.1109/TPAMI.2023.3263824. URL https://doi.org/10.1109/TPAMI.2023.3263824.

David P. Reichert and Thomas Serre. Neuronal synchrony in complex-valued deep networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6115.

Timothy D Ross, Steven W Worrell, Vincent J Velten, John C Mossing, and Michael Lee Bryant. Standard sar atr evaluation experiments using the mstar public release data set. In *Algorithms for Synthetic Aperture Radar Imagery V*, volume 3370, pp. 566–573. SPIE, 1998.

Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 2965–2972. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.379. URL https://doi.org/10.1109/CVPR.2014.379.

Utkarsh Singhal, Yifei Xing, and Stella X Yu. Co-domain symmetry for complex-valued deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 681–690, 2022.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Aleksandar Stanic, Anand Gopalakrishnan, Kazuki Irie, and Jürgen Schmidhuber. Contrastive training of complex-valued autoencoders for object discovery. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/2439ec22091b9d6cfbebf3284b40116e-Abstract-Conference.html.

Chang Tang, Xinzhong Zhu, Xinwang Liu, Lizhe Wang, and Albert Y. Zomaya. Defusionnet: Defocus blur detection via recurrently fusing and refining multi-scale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2700–2709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00281. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Tang_DeFusionNET_Defocus_Blur_Detection_via_Recurrently_Fusing_and_Refining_Multi-Scale_CVPR_2019_paper.html.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5314–5321, 2022.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. In *International Conference on Learning Representations*, 2018.

Mark Tygert, Joan Bruna, Soumith Chintala, Yann LeCun, Serkan Piantino, and Arthur Szlam. A mathematical motivation for complex-valued convolutional networks. *Neural computation*, 28 (5):815–825, 2016.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Bhavya Vasudeva, Puneesh Deora, Saumik Bhattacharya, and Pyari Mohan Pradhan. Compressed sensing mri reconstruction with co-vegan: Complex-valued generative adversarial network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 672–681, 2022.

Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pp. 816–832. Springer, 2016. doi: 10.1007/978-3-319-46466-4_49. URL https://doi.org/10.1007/978-3-319-46466-4_49.

Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1788–1797. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00192. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Stacked_Conditional_Generative_CVPR_2018_paper.html.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3796–3805. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.404. URL https://doi.org/10.1109/CVPR.2017.404.

Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Trans. Image Process.*, 27(5):2368–2378, 2018. doi: 10.1109/TIP.2017.2787612. URL https://doi.org/10.1109/TIP.2017.2787612.

Chenxi Wei, Zhenyuan Ji, Maosheng Wei, Yun Zhang, and Haoxuan Yuan. Scv-unet: Saliency-combined complex-valued u-net for sar ship target segmentation. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6948–6951, 2023. doi: 10.1109/IGARSS52108.2023.10283275.

Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 13022–13031. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01304. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wei_Label_Decoupling_Framework_for_Salient_Object_Detection_CVPR_2020_paper.html.

Saurabh Yadav and Koteswar Rao Jerripothula. Fccns: Fully complex-valued convolutional networks using complex-valued color model and loss function. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10689–10698, October 2023.

Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pp. 1155–1162. IEEE Computer Society, 2013. doi: 10.1109/CVPR.2013.153. URL https://doi.org/10.1109/CVPR.2013.153.

Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pp. 3166–3173. IEEE Computer Society,

2013. doi: 10.1109/CVPR.2013.407. URL https://doi.org/10.1109/CVPR.2013.407.

Muqiao Yang, Martin Q. Ma, Dongyu Li, Yao-Hung Hubert Tsai, and Ruslan Salakhutdinov. Complex transformer: A framework for modeling complex-valued sequence. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 4232–4236. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9054008. URL https://doi.org/10.1109/ICASSP40776.2020.9054008.

Yanchao Yang and Stefano Soatto. FDA: fourier domain adaptation for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 4084–4094. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00414. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yang_FDA_Fourier_Domain_Adaptation_for_Semantic_Segmentation_CVPR_2020_paper.html.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 538–547. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00060. URL https://doi.org/10.1109/ICCV48922.2021.00060.

Jianming Zhang and Stan Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(5):889–902, 2016. doi: 10.1109/TPAMI.2015.2473844. URL https://doi.org/10.1109/TPAMI.2015.2473844.

Jiaxing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8778–8787. IEEE, 2019a. doi: 10.1109/ICCV.2019.00887. URL https://doi.org/10.1109/ICCV.2019.00887.

Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3080–3088. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00325. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhao_Defocus_Blur_Detection_CVPR_2018_paper.html.

Wenda Zhao, Bowen Zheng, Qiuhua Lin, and Huchuan Lu. Enhancing diversity of defocus blur detectors via cross-ensemble network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8905–8913. Computer Vision Foundation / IEEE, 2019b. doi: 10.1109/CVPR.2019.00911. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zhao_Enhancing_Diversity_of_Defocus_Blur_Detectors_via_Cross-Ensemble_Network_CVPR_2019_paper.html.

Wenda Zhao, Xueqing Hou, You He, and Huchuan Lu. Defocus blur detection via boosting diversity of deep ensemble networks. *IEEE Trans. Image Process.*, 30:5426–5438, 2021a. doi: 10.1109/TIP.2021.3084101. URL https://doi.org/10.1109/TIP.2021.3084101.

Wenda Zhao, Cai Shang, and Huchuan Lu. Self-generated defocus blur detection via dual adversarial discriminators. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 6933–6942. Computer Vision Foundation / IEEE, 2021b. doi: 10.1109/CVPR46437.2021.00686. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Self-Generated_Defocus_Blur_Detection_via_Dual_Adversarial_Discriminators_CVPR_2021_paper.html.

Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In Andrea Vedaldi, Horst Bischof, Thomas

Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pp. 35–51. Springer, 2020. doi: 10.1007/978-3-030-58536-5_3. URL https://doi.org/10.1007/978-3-030-58536-5_3.

Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. Distraction-aware shadow detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5167–5176. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00531. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zheng_Distraction-Aware_Shadow_Detection_CVPR_2019_paper.html.

Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pp. 122–137. Springer, 2018. doi: 10.1007/978-3-030-01231-1_8. URL https://doi.org/10.1007/978-3-030-01231-1_8.

Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson W. H. Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 4682–4691. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00466. URL https://doi.org/10.1109/ICCV48922.2021.00466.

A INVERTIBILITY OF R2C TRANSFORM

As mentioned in the main manuscript, R2C transform is invertible. Given the function transform function $T : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{C}^{H\times W\times 2}$, we need to show that it is invertible, which implies it is a one to one function. In Algorithm. 2, we present a pseudo code for inverse R2C color transform. We focus on obtaining the vector $v = \overrightarrow{CF}$, as shown in Fig 7. Once it is obtained, we will have the location of pixel p at F.

We provide a pseudo code for the inverse R2C transform. Given the complex image I_{iRGB}, we convert each pixel value from complex to real. First, we find $u = \overrightarrow{EF}$ (Fig. 7) using its components along R and $R^{\perp}$, then using the component of v along $O = \overrightarrow{CE}$, we obtain $v = \overrightarrow{CE} + \overrightarrow{EF}$ and hence the point F which is the location of pixel p in RGB space. We also provide a code demo for R2C transform in the supplementary zip folder.

Algorithm 2 Pseudo code for inverse R2C Color Transformation

Input: A complex-valued Image, I_{iRGB}
Output: A real-valued Image, I_{RGB}
  for $p \in I_{iRGB}$ **do** ▷ for each pixel in the complex image
    ▷ Our goal is to find vector $v = \overrightarrow{CF} = \overrightarrow{CE} + \overrightarrow{EF}$
    ▷ Let us assume $u = \overrightarrow{EF}$, O = grayscale vector
    ▷ We have two complex numbers for pixel p
    $C_1 = ||v|| \cos\theta + i||v|| \sin\theta$
    $C_2 = ||u|| \cos\phi + i||u|| \sin\phi$
    point E = $(\Re(C_1), \Re(C_1), \Re(C_1))$ ▷ lies on O
    plane P = plane formed perpendicular to vector O at point E
    point B = Intersection of Plane P and Red axis ▷ plane and axis are known
    vector R = $\overrightarrow{EB}$
    vector R$^{\perp}$ = lies in plane P, and has angle $+\pi/2$ with R
    ▷ find vector u in RGB space
    $u_1 = \Re(C_2)\hat{R}$ ▷ component of u along R
    $u_2 = \Im(C_2)\hat{R^{\perp}}$ ▷ component of u along $R^{\perp}$
    $u = \overrightarrow{EF} = u_1 + u_2$
    ▷ find component of vector v in RGB space
    $\overrightarrow{CE} = \Re(C_1)\hat{O}$ ▷ component of v along O
    ▷ find vector v in RGB space
    $v = \overrightarrow{CF} = \overrightarrow{CE} + \overrightarrow{EF} = [I_R(p), I_G(p), I_B(p)]$
    point $F = (I_R(p), I_G(p), I_B(p))$
  end for
  return $I_{RGB} = \{I_R, I_G, I_B\}$ ▷ Final real-valued image

B COMPLEX INPUT USING FOURIER TRANSFORM

In addition to the R2C transform, DFT (Discrete Fourier transform) can be used to generate complex-valued representations of images. However, due to a lack of spatial structure and localized context, Fourier representation does not provide additional benefits. To assert this observation, we conduct an experiment with two variations of the DCSNet encoder on the CIFAR10 dataset in Tab. 8.

Table 8: Comparison with DFT (Discrete Fourier transform) and iRGB input to DCSNet encoders.

	DCSNet-Encoder		DCSNet-Encoder small	
Input	DFT	iRGB	DFT	iRGB
Acc(%)	85.8	**94.3**	77.6	**92.8**

Table 9: Additional ablation study to compare our Fourier filter module and self-attention.

Metric	DUTS		ECSSD		HKU-IS		PASCAL-S		DUT-O	
	SA	Ours	SA	Ours	SA	Ours	SA	Ours	SA	Ours
$S_m \uparrow$	0.735	**0.740**	0.828	**0.831**	**0.748**	0.733	0.745	**0.747**	0.738	**0.739**
maxF $\uparrow$	0.651	**0.654**	0.820	**0.825**	0.688	**0.689**	0.711	**0.713**	**0.645**	**0.645**
$E_\xi^{max} \uparrow$	0.788	**0.801**	0.875	**0.884**	0.763	**0.791**	0.787	**0.801**	0.788	**0.795**
MAE $\downarrow$	**0.104**	0.107	0.091	**0.083**	0.170	**0.157**	0.146	**0.131**	0.112	**0.109**

C BACKGROUND

We start by introducing the discrete Fourier transform (DFT), which plays a vital role in signal processing. For clarity we consider 1D DFT. Give a sequence of N complex numbers $x[n], 0 \leq n \leq N-1$, the DFT of $x[n]$ will be:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i(2\pi/N)kn} = \sum_{n=0}^{N-1} x[n] W_N^{kn} \tag{7}$$

where i is the iota, and $W_N = e^{-i(2\pi/N)}$.

Since $X[k]$ repeats on intervals of length N, we can take value of $X[k]$ at N consecutive points $k = 0, 1, \ldots, N-1$. Specifically, $X[k]$ represents the spectrum of sequence $x[n]$ at the frequency $w_k = 2\pi k/N$.

It is well known that the DFT is a bijective function, i.e., the inverse of DFT function exists. Given $X[k]$, we can recover the original signal $x[n]$ by the inverse DFT also denoted as IDFT

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{i(2\pi/N)kn} \tag{8}$$

Note that, in real DFT, the input $x[n]$ is real, and its DFT is conjugate symmetric Rao et al. (2021), i.e., $X[N-k] = X^*[k]$. The reverse is true as well; if we perform IDFT to $X[k]$ which is conjugate symmetric, a real discrete signal can be covered. This is a major point that half of the DFT $\{X[k] : 0 \leq k \leq [N/2]\}$ contains full information of $x[n]$.

However, in complex DFT, the input $x[n]$ is complex. Hence, its DFT includes both positive and negative frequencies. This means that unlike real DFT, $X[k]$ is not conjugate symmetric. $X[k]$ between 0 to $N/2$ is positive and between $N/2$ and $N-1$ its negative.

The DFT described above can be extended to 2D signals. Given 2D complex signal $X[m, n], 0 \leq m \leq M-1, 0 \leq n \leq N-1$, the 2D DFT of $x[m.n]$ is given by:

$$X[u, v] = \sum_{M-1}^{m=0} \sum_{N-1}^{n=0} x[m, n] e^{-i2\pi(\frac{um}{M} + \frac{vn}{N})} \tag{9}$$

The 2D DFT can be thought of as performing 1D DFT on the two dimensions alternatively. Similar to complex 1D DFT, 2D DFT does not have properties of conjugate symmetry.

D ABLATION EXPERIMENTS

We also provide another ablation study to compare complex-valued self-attention and our proposed learnable Fourier filter. For this ablation, we replace the Fourier filter module with the complex-valued self-attention (SA) module as proposed by Eilers & Jiang (2023); Yang et al. (2020). We can notice performance-decline in most cases when the Fourier filter module is replaced. Table 9 empirically validates our proposed method for capturing global information in complex-domain.

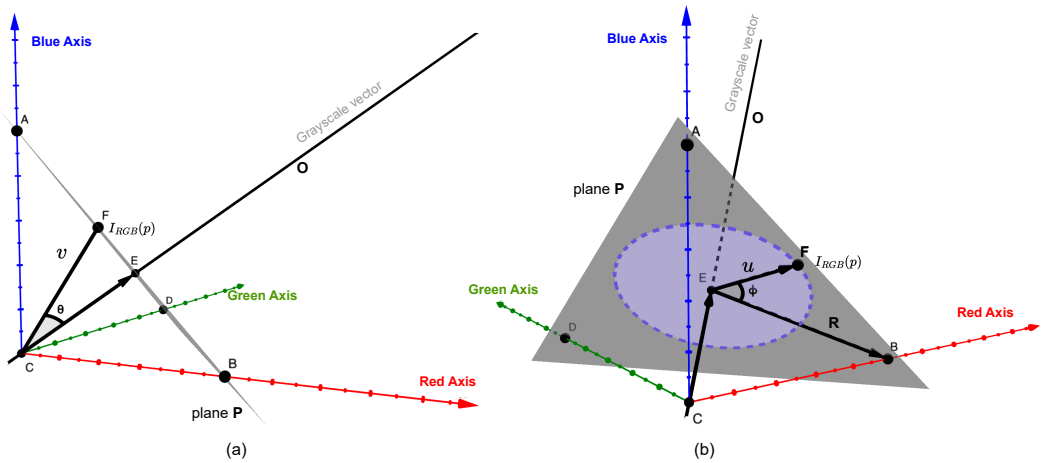


Figure 7: R2C color transformation: Given an RGB color $I_{RGB}(p) = \{I_R(p), I_G(p), I_B(p)\}$ for a pixel p, we transform p as $I_{iRGB} = \{I_\theta(p), I_\phi(p)\}$.

E Additional Results

For comparison with complex-valued method for saliency prediction, we follow Jiang et al. (2019) in our experimental setup. DCSNet is trained from scratch over the training set of SALICON Jiang et al. (2015), following Jiang et al. (2019). We test our trained model on 3 widely used image saliency datasets, i.e., MIT1003Cornia et al. (2016), CAT2000Borji & Itti (2015), and DUTYang et al. (2013) and compare using 4 metrics: area under the curve (AUC), normalized scanpath saliency (NSS), CC and KL divergence.

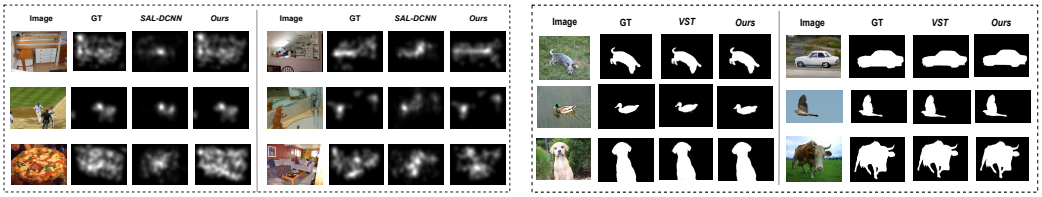
In Table 11, we present a comparison with only published complex-valued saliency prediction methodJiang et al. (2019). As shown in the table, our method performs best overall.

Table 10: Comparison with FCCNYadav & Jerripothula (2023) across diffrent number of parameters.

ImageNet	FCCN			DCSNet		
	Resnet18	ResNet50	ResNet152	Encoder		
Param (M)	11	26	60	5	15	17
Acc (%)	73.41	76.26	77.27	71.24	76.07	**78.83**

F Qualitative Results

In addition to all quantitative evaluations, we also validate our results qualitatively. We present these results in Fig. 8. The figure verifies our empirical claims when comparing with the complex-valued method as well as the real-valued method.



(a) Saliency map compared with SAL-DCNNJiang(b) Salient Objects compared with VSTLiu et al. (2021) et al. (2019)

Figure 8: Qualitative results of our proposed CSNet compared with other methods.

Table 11: Comparison with existing complex-valued method SAL-DCNN Jiang et al. (2019)(light gray). Following Jiang et al. (2019), we take 201, 400, and $1,035$ test images for MIT1003, CAT2000 and DUT respectively. Our DCSNet obtains best results across all three datasets.

	MIT1003Cornia et al. (2016)				CAT2000Borji & Itti (2015)				DUTYang et al. (2013)			
	AUC	NSS	CC	KL	AUC	NSS	CC	KL	AUC	NSS	CC	KL
SALICONJiang et al. (2015)	0.82	1.28	0.42	1.61	0.77	0.99	0.39	1.17	0.85	2.27	0.48	1.24
DVAWang & Shen (2018)	0.86	2,19	0.66	0.87	0.81	1.50	0.56	0.84	0.91	3.11	0.67	0.88
SALGANPan et al. (2017)	0.87	2.05	0.65	0.96	0.81	1.47	0.56	0.97	0.91	2.80	0.68	0.90
ML-NetCornia et al. (2016)	0.84	2.01	0.61	1.01	0.79	1.37	0.51	0.99	0.88	2.87	0.61	1.15
SAMCornia et al. (2018)	0.87	2.19	0.61	1.30	0.84	1.74	0.63	1.12	0.91	2.96	0.67	1.07
BMSZhang & Sclaroff (2016)	0.77	1.15	0.37	1.43	0.78	1.20	0.46	1.07	0.83	1.76	0.42	1.40
PQFTGuo & Zhang (2010)	0.70	0.78	0.25	1.67	0.75	0.98	0.37	1.18	0.77	1.26	0.33	1.53
SRHou & Zhang (2007)	0.70	0.80	0.25	1.69	0.72	0.87	0.32	6.05	0.67	0.70	0.15	3.54
Sal-DCNNJiang et al. (2019)	0.87	2.10	0.62	0.89	0.86	2.03	0.79	0.63	0.92	3.07	0.76	0.55
Sal-DCNN-PPJiang et al. (2019)	0.86	1.98	0.61	0.93	0.86	2.00	0.77	0.65	0.92	3.06	0.75	0.57
Sal-DCNN-PJiang et al. (2019)	0.86	1.97	0.60	0.98	0.86	1.99	0.76	0.74	0.92	3.05	0.75	0.60
Sal-DenseNetJiang et al. (2019)	0.85	1.95	0.59	1.04	0.86	1.95	0.74	0.97	0.91	3.03	0.74	0.63
DCSNet (ours)	**0.93**	**2.35**	**0.71**	**0.82**	**0.89**	**2.26**	**0.83**	**0.57**	**0.95**	**3.23**	**0.81**	**0.52**