Finding the Secret of Image Saliency in the Frequency Domain

Jia Li, *Member, IEEE*, Ling-Yu Duan, *Member, IEEE*, Xiaowu Chen, *Member, IEEE*, Tiejun Huang, *Senior Member, IEEE*, and Yonghong Tian, *Senior Member, IEEE*

Abstract—There are two sides to every story of visual saliency modeling in the frequency domain. On the one hand, image saliency can be effectively estimated by applying simple operations to the frequency spectrum. On the other hand, it is still unclear which part of the frequency spectrum contributes the most to popping-out targets and suppressing distractors. Toward this end, this paper tentatively explores the secret of image saliency in the frequency domain. From the results obtained in several qualitative and quantitative experiments, we find that the secret of visual saliency may mainly hide in the phases of intermediate frequencies. To explain this finding, we reinterpret the concept of discrete Fourier transform from the perspective of template-based contrast computation and thus develop several principles for designing the saliency detector in the frequency domain. Following these principles, we propose a novel approach to design the saliency detector under the assistance of prior knowledge obtained through both unsupervised and supervised learning processes. Experimental results on a public image benchmark show that the learned saliency detector outperforms 18 state-of-the-art approaches in predicting human fixations.

Index Terms—Image saliency, Fourier transform, spectral analysis, fixation prediction, learning-based, experimental study

1 Introduction

The history of visual saliency is an extremely long story. The concept of computational visual saliency modeling, however, is still very young in the field of computer vision and image processing. In a word, visual saliency modeling aims to simulate the selective mechanism in human vision system by detecting the most conspicuous content in images and videos. With this tool, applications such as image/video retargeting [1], [2], smart advertising [3], [4] and image analysis in remote sensing [5], [6] can achieve impressive performance by focusing on the same visual content as human being does.

The booming of visual saliency modeling is usually considered to originate from the work in [7]. After the rapid development in the past two decades, three tightly correlated branches emerge in this field, including objectness proposal generation, fixation prediction and salient object segmentation. Among these branches, *objectness proposal generation* focuses on locating "objects," including both targets and distractors, in the input scene with rectangular windows [8], [9]. *Fixation prediction* aims to roughly pop-out only targets and inhibit probable distractors [10], [11], [12], while *salient object segmentation* proposes to exactly segment only the closed contours of salient targets [13], [14]. Among these three branches, fixation prediction can be intuitively viewed as a preparatory step of salient object segmentation and a special case of objectness proposal generation, which is also the major concern of this study.

In existing works, saliency is often defined, explicitly or implicitly, as the visual irregularity measured by various contexts (e.g., local/global), features (e.g., intensity/color/orientation), domains (e.g., spatial/spatiotemporal/spectral) and attentional mechanisms (e.g., bottom-up/top-down). In this study, we roughly categorize existing saliency models into four groups. The first group, denoted as the *BS* group, contains bottom-up models that measure visual saliency by heuristically computing and combining a set of contrast-like features in the spatial or spatiotemporal domain (e.g., [7], [10], [15]). However, these models may fail to distinguish targets from distractors that share similar visual attributes. It is argued that the prior knowledge obtained through similar scenes viewed before plays an important role in separating targets and distractors [16].

Following this idea, some saliency models are proposed to incorporate prior knowledge into visual saliency estimation, while such prior knowledge can be obtained by unsupervised learning (i.e., the *UL* group) or supervised learning (i.e., the *SL* group). In particular, models in the *UL* group focus on refining the extraction of saliency cues. These models (e.g., [12], [17], [18], [19]) often sample massive image patches for training sparse codes (visual words, basis functions, principle/independent components, etc.). With these sparse codes, a less redundant representation of the input image can be obtained, which is believed to be more suitable for extracting saliency cues. Similar to models in the *BS* group, such saliency cues are often heuristically combined to measure the degree of saliency. On the contrary, models

- *J. Li and X. Chen are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. E-mail: {jiali, chen}@buaa.edu.cn.*
- *J. Li is also with the International Research Institute for Multidisciplinary Science at Beihang University, Beijing 100191, China.*
- *L. Duan, T. Huang, and Y. Tian are with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China. E-mail: {lingyu, tjhuang, yhtian}@pku.edu.cn.*

Manuscript received 8 Aug. 2014; revised 23 Jan. 2015; accepted 10 Apr. 2015. Date of publication 19 Apr. 2015; date of current version 6 Nov. 2015. Recommended for acceptance by L. Zelnik-Manor.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2424870

in the *SL* group often emphasize the optimal integration of existing spatial/spatiotemporal saliency cues. In these models (e.g., [20], [21], [22]), it is believed that an optimal feature integration strategy can be learned from user annotations (e.g., fixation density maps or labeled salient objects) by using the supervised learning algorithms. As reported in [16], models in both *UL* and *SL* groups have some advantages, and the learned prior knowledge can be helpful in predicting fixations.

Most models discussed above fall into spatial or spatiotemporal domain. On the contrary, some models (e.g., [23], [24], [25], [26], [27]) are proposed to detect saliency in the frequency domain (i.e., the *FQ* group). In these models, saliency is often detected with three major steps: 1) applying discrete Fourier transform (DFT) or discrete Cosine transform (DCT) to input feature channels (e.g., intensity, color opponencies or motion field); 2) modulating the frequency spectrum (e.g., subtracting the average amplitude spectrum [23], setting the spectral magnitude to unity [24], or keeping only the signs of DCT coefficients [26]); 3) generating saliency map through inverse DFT/DCT. Since each DFT/DCT coefficient represents a kind of statistical information of all input stimuli, saliency models in the *FQ* group can efficiently locate the most salient targets from a "global" perspective. However, it is still unclear which part of the frequency spectrum contributes the most to popping-out targets and suppressing distractors.

In this study, we propose to find the secret of image saliency in the frequency domain by integrating the advantages of various saliency models discussed above. Toward this end, we first conduct a set of experiments to qualitatively and quantitatively measure the contribution of various spectral information to saliency computation. From these results, we find that the secret of saliency may mainly hide in the phases of intermediate frequencies. To explain this, we reinterpret Fourier transform from the perspective of template-based contrast computation, which in turn reveals several principles for designing the saliency detector in the frequency domain. Following these principles, we propose an approach to design the saliency detector under the assistance of prior knowledge obtained by unsupervised and supervised learning. Experimental results show that the learned saliency detector outperforms 18 state-of-the-art saliency models on a public image benchmark.

Our main contributions are summarized as follows:

1) Through extensive experimental studies, we find that the secret of visual saliency may mainly hide in the phases of intermediate frequencies obtained by Fourier transform. In particular, the signs of the real and imaginary parts, once correctly estimated, have remarkable contribution to locating salient targets;
2) We reinterpret discrete Fourier transform from the perspective of template-based contrast computation, followed by five principles to design the saliency detector in the frequency domain;
3) We propose a novel approach to design the image saliency detector under the assistance of prior knowledge obtained by both unsupervised and supervised learning. To the best of our knowledge, it is the first time that visual saliency estimation was formulated as a machine learning problem in the frequency domain.

The rest of this paper is organized as follows: Section 2 reviews existing saliency models. Section 3 conducts qualitative and quantitative studies to find the secret of image saliency in the frequency domain. In Section 4, we propose a learning-based approach to design the saliency detector. Experimental results are presented in Section 5, and the entire paper is concluded in Section 6.

2 Related Work

In this Section, we first briefly review representative saliency models in the *BS*, *UL* and *SL* groups, followed by discussions on models in the *FQ* group with more technical details.

Models in the *BS* group often compute visual saliency by measuring the visual "irregularity" in the spatial or spatiotemporal domain. Usually, such irregularity can be measured by multi-scale center-surround contrasts [7], visiting time in random walk [28], pattern/color distinctness [29], or the existence of high-level features [10]. Moreover, some approaches incorporated global contrast (e.g., [30], [31]) or uniqueness (e.g., [32]) to measure visual rarity from a global perspective, which achieved breakthrough results in detecting salient objects. Furthermore, such irregularity can also take temporal information (e.g., flicker [15] and motion [33], [34]) into account for video saliency estimation. In recent studies, Zhang and Sclaroff [35] proposed to use simple features (e.g., Lab color channels) and then randomly threshold them to generate a set of Boolean maps so as to measure visual saliency. On the contrary, Xu et al. [36] extracted a large set of pixel-level, object-level and semantic-level attributes and combined them for saliency estimation. These two approaches represent two feasible directions in developing bottom-up models in the spatial/spatiotemporal domain.

For models in the *BS* group, one problem is the existence of visual redundancy across feature channels. With such redundancy, targets and distractors may share some common visual attributes, making it difficult to separate them. To address this problem, two feasible solutions are proposed: encoding raw features to remove redundancy before saliency estimation (i.e., models in the *UL* group) or identifying feature channels in which targets and distractors have the least attributes in common (i.e., models in the *SL* group). Following these two solutions, models in the *UL* group utilize the prior knowledge obtained by unsupervised learning (e.g., basis functions [19], independent components [17], [18], and visual words [12]). Such prior knowledge can be used to obtain a compact representation of the input image, which contains low redundancy. On the contrary, models in the *SL* group are proposed to derive such prior knowledge from supervised learning. That is, they adopt various learning algorithms to derive the optimal "feature-saliency" mapping models from user data (e.g., human fixations and salient object masks). Usually, such models can take the form of linear weights [20], [22], Support Vector Machine [21], [37], [38], boosting classifier [11], [39], ranking model [40], and Markov Random Field [41]. Except some outliers, most of these models work by emphasizing the

feature channels that perform the best in separating targets (fixated locations or salient objects [42]) from distractors.

Beyond the models discussed above, models in the *FQ* group were proposed to estimate visual saliency in the frequency domain. In these models, a typical flowchart is to transform images or video frames into the frequency domain by using DFT/DCT (or hypercomplex Fourier transform, HFT). After that, the frequency spectrum is heuristically modulated and transformed back to the spatial domain so as to generate a saliency map with simple post-processing steps (e.g., Gaussian blurring). Surprisingly, these models can produce impressive results with high efficiency. In [43], such frequency-based models were proved to be biologically plausible.

As a pioneer work in the *FQ* group, Hou and Zhang [23] proposed to extract spectral residual from the frequency spectrum of image intensity, which was equivalent to subtracting the locally averaged amplitude from the original one. Similarly, Cui et al. [44], [45] proposed to compute temporal spectral residual so as to detect salient motion in a video sequence. Inspired by the work in [23], Guo and Zhang proposed two innovative improvements in [24], [46]. The first improvement, denoted as PFT, proposed to use only the phase spectrum of image intensity (and unity magnitude) for saliency detection.[1] In a later work, Hou et al. [26] further validated that the signs of DCT coefficients performed impressively in detecting salient image locations. In this case, such signs could be treated as a kind of image signature. The second improvement, denoted as PQFT, represented an image as a quaternion comprising of four feature channels (i.e., intensity, red/green and blue/yellow color opponencies, and motion). In the quaternion-based representation, different feature channels can encode different cues so as to measure the degree of saliency from multiple perspectives. Thus image (or video) saliency can be simultaneously estimated through the frequency spectrum obtained by applying HFT to multiple features.

Among the two improvements in [24], the latter one has much stronger impact than the former one. As the phase-only framework has been followed in [48], [49], it becomes very popular to represent an image as a quaternion in recent studies (e.g., [27], [50], [51]). For instance, Li et al. [52] combined the approaches of [23] and [24] by computing the spectral difference of amplitude and phase between an input image and its blurred version in HSV color channels. This work was further extended to compute video saliency in [27]. In [53], image intensity was assigned with a higher weight than the red/green and blue/yellow color opponencies. In this manner, image intensity was better emphasized in amplitude modulation (e.g., smoothing the spectral amplitude obtained by HFT). In [2], Fang wt al. proposed to use discrete Cosine transform for image saliency analysis, while Schauerte and Stiefelhagen [51] proposed to use the quaternion DCT and incorporated the face conspicuity map to improve the overall performance of the saliency model. Later on, Schauerte and Stiefelhagen [25] tested the performances of various frequency-based saliency models on several public image benchmarks. In this study, they conducted extensive experiments to see the influence of color spaces, weights of feature channels and number of scales. It was reported that the performance of a saliency model in the *FQ* group may differ remarkably when different color spaces or different numbers of scales were used. This implies that the redundancy in an image still exists in its frequency spectrum, which may degrade the overall performance of saliency estimation.

By inspecting all models discussed above, we can see that frequency spectrum may contain invaluable cues for visual saliency computation. Although saliency can be efficiently mined through simple spectral modulation, it is still unclear which part of frequency spectrum corresponds to the real secret for saliency estimation. As a result, in this study we propose to find the secret of image saliency in the frequency domain through qualitative and quantitative experimental studies. Furthermore, as the usage of prior knowledge in the saliency models from *UL* and *SL* groups have been proved to be helpful for separating targets and distractors, it is worth discussing how to mine the saliency secret from the frequency spectrum under the assistance of prior knowledge obtained by unsupervised and supervised learning.

3 Finding the Secret of Image Saliency in the Frequency Domain

In this Section, we conduct qualitative and quantitative experiments to explore the secret of image saliency in the frequency domain. We also reinterpret the concept of DFT from the perspective of template-based contrast computation. Finally, several principles are proposed to design the saliency detector.

3.1 A Qualitative Study on the Secret of Saliency

As discussed above, both spectral amplitude and phase can contribute to the detection of salient target. Thus two major concerns may arise: what are the roles of these two spectra in popping-out targets and suppressing distractors and which spectrum has more contribution? To address these two concerns, we design a small experiment for qualitative analysis.

Typically, the objective of saliency prediction is to generate a saliency map S for an input image I that perfectly approximates its fixation density map G:

$$I \Rightarrow S \rightarrow G, \tag{1}$$

where "$\Rightarrow$" and "$\rightarrow$" mean "generation" and "approximation," respectively. As a consequence, we can express the same correlations between I, S and G in the frequency domain:

$$\mathscr{F}[I] \Rightarrow \mathscr{F}[S] \rightarrow \mathscr{F}[G], \tag{2}$$

where $\mathscr{F}[\cdot]$ denotes discrete Fourier transform. In other words, the amplitude and phase spectra of S, which are generated from the frequency spectrum of I, are expected to approximate those spectra of G:

$$\mathcal{A}(\mathscr{F}[S]) \rightarrow \mathcal{A}(\mathscr{F}[G]),\ \mathcal{P}(\mathscr{F}[S]) \rightarrow \mathcal{P}(\mathscr{F}[G]), \tag{3}$$

where $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ denote spectral amplitude and phase, respectively. In this manner, the problem of image saliency

1. Actually, the phase-only image reconstruction have already been well studied decades ago (e.g., in [47]). However, in [24] it is the first time that the reconstructed images were used to address the problem of fixation prediction.

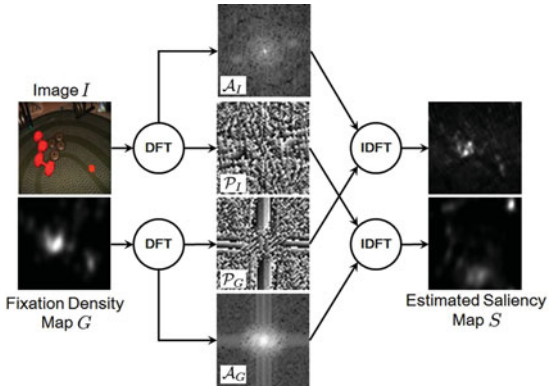


Fig. 1. Spectral phase contains the secret for locating the salient targets, and spectral amplitude helps to determine the saliency strength. $\mathcal{A}_I$ and $\mathcal{A}_G$: amplitude spectra; $\mathcal{P}_I$ and $\mathcal{P}_G$: phase spectra. IDFT: inverse DFT.

estimation can be described as *modulating the amplitude and phase spectra of an input image to approximate the corresponding spectra of its fixation density map*. For the sake of simplification, we use $\mathcal{A}_I$ (or $\mathcal{A}_G$) and $\mathcal{P}_I$ (or $\mathcal{P}_G$) to represent the amplitude and phase spectra of I (or G).

Since the final objective of visual saliency estimation is to approximate G with S, we can safely assume that $\mathcal{A}_G$ and $\mathcal{P}_G$ in (3) are the "perfect" amplitude and phase spectra, respectively. Thus an interesting concern may arise: which spectrum, $\mathcal{A}_G$ or $\mathcal{P}_G$, contributes the most to locating salient image content? Toward this end, we propose to see what saliency maps can be generated by combining $\mathcal{A}_G$ with $\mathcal{P}_I$ and $\mathcal{A}_I$ with $\mathcal{P}_G$. Here we generate a saliency map by transforming each combination of amplitude and phase spectra back to the spatial domain and then squaring the modulus of every complex-valued pixel. Here we simply normalize the saliency map to [0, 1] without any post-smoothing as in [23], [24], [53].

From Fig. 1, we find that combining $\mathcal{A}_G$ and $\mathcal{P}_I$ generates a "clean" saliency map, but most energy is assigned to unexpected locations. On the contrary, combining $\mathcal{A}_I$ and $\mathcal{P}_G$ succeeds in popping-out salient targets but fails to suppress noise in most locations. As a result, we tentatively assume that *the secret of salient location mainly hides in spectral phase, and the secret of saliency strength can be found in spectral amplitude.*

3.2 A Quantitative Study on the Secret of Saliency

Beyond qualitative analysis, we also wish to quantitatively measure the contributions of spectral amplitude and phase.

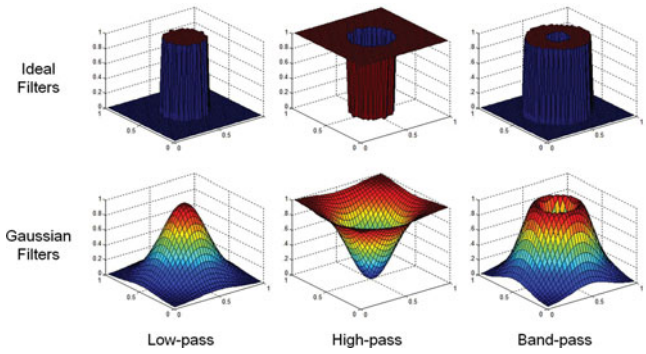


Fig. 2. Candidate filters used in the quantitative study, including the ideal/Gaussian low-pass, high-pass and band-pass filters.

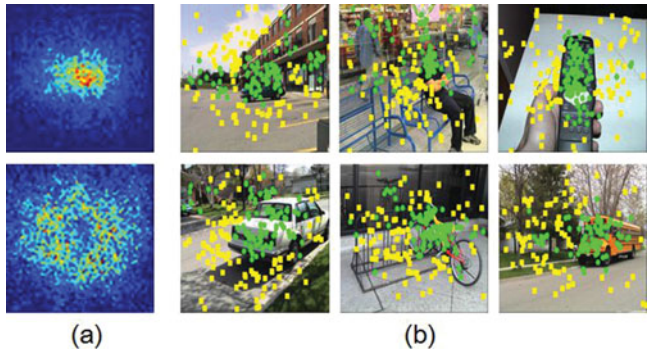


Fig. 3. Sampled instances for evaluation. (a) density maps for fixated and non-fixated pixels; (b) Sampled fixated pixels (green) and non-fixated pixels (yellow).

Therefore, we conduct several experiments on **Toronto** [17], which contains 120 color images and their fixation density maps. In the experiments, we convert all the 120 images to gray and then test the performance of saliency maps computed by combining various types of amplitude and phase spectra. In total, two types of phase spectra, $\mathcal{P}_I$ and $\mathcal{P}_G$, are combined with five types of amplitude spectra, including $\mathcal{A}_I$, $\mathcal{A}_G$, $\mathcal{A}_I \otimes H_l$, $\mathcal{A}_I \otimes H_h$ and $\mathcal{A}_I \otimes H_b$. Here H_l, H_h, H_b are low-pass, high-pass and band-pass filters, respectively. The operator $\otimes$ indicates element-wise multiplication of two matrices. For each pair of spectral amplitude and phase, we generate saliency maps in the same way as the qualitative experiment does. Moreover, we conduct the experiment three times at 128×128, 64×64 and 32×32 to explore the influence of resolution.

In the experiment, we have tested the ideal form and the Gaussian form of each filter (i.e., H_l, H_h, H_b). Suppose the frequency spectrum spans from 0 (at center) to 1 (at corner), we control the shape of a low-pass or high-pass filter by its cut-off frequency $\sigma_e \in [0, 1]$, and a band-pass filter can be controlled by the pass-band center $\sigma_c \in [0, 1]$ with bandwidth $\sigma_w \in [0, \min(2\sigma_c, 2 - 2\sigma_c)]$. The typical shapes of these filters are shown in Fig. 2. Note that we enumerate all feasible values of σ_e, σ_c and σ_w from 0 to 1 with a step of 0.1 to find the best parameters.

For quantitative evaluation, we select all fixated pixels and the same number of non-fixated pixels from each image, which are used as positive and negative instances, respectively. To alleviate the influence of dataset bias (e.g., center-bias), non-fixated pixels are selected with respect to the distribution of fixated pixels in all the 120 images. As shown in Fig. 3, non-fixated pixels are also center-biased and have a minimum distance from fixated pixels to avoid ambiguity. Let $\mathbb{P}$ and $\mathbb{N}$ be the sets of positive and negative instances, we use $S(p) \in [0, 1]$ to represent the estimated saliency value of $p \in \mathbb{P} \cup \mathbb{N}$.

In the experiment, we adopt Area under the ROC curve (AUC) to be an evaluation metric [54]. To compute AUC, we apply all thresholds in $\{0, 0.01, 0.02, \ldots, 1\}$ to binarize $S(p)$ for $p \in \mathbb{P} \cup \mathbb{N}$. At each threshold, a pair of false positive rate and true positive rate are computed to interpolate the ROC curve. In this manner, saliency maps that always assign targets with saliency values higher than distractors will gain an AUC of 1.0, while random assignment leads to an AUC of 0.5.

TABLE 1
Performance of Saliency Maps on **Toronto** When Various Amplitude and Phase Spectra Are Combined at Three Resolutions

	$\mathcal{P}_I$ (32×32)			$\mathcal{P}_G$ (32×32)			$\mathcal{P}_I$ (64×64)			$\mathcal{P}_G$ (64×64)			$\mathcal{P}_I$ (128×128)			$\mathcal{P}_G$ (128×128)		
	AUC	EOF	FS	AUC	EOF	FS	AUC	EOF	FS	AUC	EOF	FS	AUC	EOF	FS	AUC	EOF	FS
$\mathcal{A}_I$	-	-	-	**0.91**	0.83	0.87	-	-	-	0.90	0.82	0.86	-	-	-	0.90	0.82	0.86
$\mathcal{A}_G$	0.49	0.49	0.49	-	-	-	0.49	0.50	0.49	-	-	-	0.49	0.50	0.49	-	-	-
$\mathcal{A}_I \otimes H_l$	0.40	0.43	0.41	**0.91**	0.83	0.87	0.39	0.42	0.40	**0.91**	0.83	0.87	0.39	0.42	0.41	**0.91**	0.83	0.87
$\mathcal{A}_I \otimes H_h$	0.64	0.67	0.66	0.84	**0.95**	0.89	0.65	0.70	0.68	0.81	0.93	0.87	**0.66**	0.68	**0.67**	0.71	0.85	0.78
$\mathcal{A}_I \otimes H_b$	**0.66**	**0.70**	**0.68**	0.90	0.90	0.90	**0.66**	**0.71**	**0.69**	0.84	**0.95**	**0.89**	0.65	**0.69**	**0.67**	0.84	**0.95**	**0.89**

AUC has been proved to be useful in many existing studies. However, it may be flawed for relying solely on the interpolated ROC curve without considering the distribution of thresholding points along the curve (i.e., the interpolation flaw [55]). In other words, AUC focuses on only the ordering of saliency [56], while the "gap" between fixated and non-fixated pixels are ignored. Thus a fuzzy saliency map may also reach a high AUC if it simply assigns slightly higher saliency values to fixated pixels than to non-fixated pixels.

To address this problem, we propose to measure the saliency gap between fixated and non-fixated pixels with Energy-On-Fixations (EOF). Unlike AUC that focuses on the ordering of saliency, EOF is computed as the ratio of energy assigned to targets:

$$\mathrm{EOF} = \frac{\sum_{p \in \mathbb{P}} \mathcal{S}(p)}{\sum_{p \in \mathbb{PUN}} \mathcal{S}(p)}. \tag{4}$$

A saliency model that assigns all energy to targets has an EOF of 1.0. We also compute the F-measure (denoted as FS) to equally consider both AUC (i.e., saliency ordering) and EOF (i.e., saliency gap):

$$\mathrm{FS} = \frac{2 \times \mathrm{AUC} \times \mathrm{EOF}}{\mathrm{AUC} + \mathrm{EOF}}. \tag{5}$$

Ideally, an optimal saliency map (e.g., the fixation density map) has $\mathrm{AUC} = \mathrm{EOF} = \mathrm{FS} = 1.0$, while a saliency map with random predictions has $\mathrm{AUC} = \mathrm{EOF} = \mathrm{FS} = 0.5$. Note that the evaluation is conducted by resizing the estimated saliency maps to the sizes of original images.

Given the evaluation metrics, we have shown in Table 1 the performance of the saliency maps that are estimated by combining various amplitude and phase spectra at three resolutions. From Table 1, we find three interesting phenomena:

- *Amplitude vs. phase*. Combining any amplitude spectrum with $\mathcal{P}_G$ results in much better performance of saliency maps than using $\mathcal{P}_I$ (on average, an increase of 0.25 ± 0.14 can be expected in FS). In particular, combining $\mathcal{A}_G$ with $\mathcal{P}_I$ is equivalent to random prediction, and combining $\mathcal{A}_I$ and $\mathcal{P}_G$ reaches a much higher FS.
- *The best filter*. Band-pass filters always perform among the best. When applied on $\mathcal{A}_I$ and combined with $\mathcal{P}_I$, the best band-pass filter acts as a high-pass filter with suppression of the highest frequency (e.g., $\sigma_c = 0.9, \sigma_w = 0.2$ at 32×32 and 64×64). When applied on $\mathcal{A}_I$ and combined with $\mathcal{P}_G$, the best band-pass filter is a kind of low-pass filter with a notch at frequencies around DC (e.g., $\sigma_c = 0.1$, $\sigma_w = 0.2$ at 32×32 and 64×64).
- *Influence of resolution*. In most cases, the resolution only slightly changes the overall performance. In particular, 32×32 is much more computationally efficient than 128×128.

From these phenomena, we can conclude that both spectral amplitude and phase contribute to saliency estimation. In particular, spectral phase, whose importance has already been proved in [47], contributes more than spectral amplitude. To further validate this conclusion, we conduct a small experiment at the resolution of 32×32. In the experiment, we keep only the signs of the real and the imaginary parts of $\mathcal{F}[G]$, while all the other cues are ignored. That is, we have only $\pm 1 \pm i$ in the frequency spectrum (and few cases of ± 1 and $\pm i$). Surprisingly, the estimated saliency maps can reach a FS score of 0.88 ($\mathrm{AUC} = 0.83$, $\mathrm{EOF} = 0.92$). This finding is consistent with the conclusion of [26] that saliency can be detected from the signs of DCT coefficients. However, we can see that saliency maps estimated only from the signs of DFT coefficients are still far from perfect (i.e., 0.17 in AUC, 0.08 in EOF and 0.12 in FS). This further validates the assumption that *the secret of saliency may mainly hide in the phase spectrum, and the spectral amplitude also contributes to the generation of perfect saliency maps, even at very low resolutions*.

3.3 A Template-Based Reinterpretation of DFT

To explain the experimental results obtained so far, we reinterpret the concept of discrete Fourier transform from the perspective of template-based contrast computation. As visual contrast, either local or global, plays an important role in measuring visual saliency, we aim to seek a direct link from Fourier coefficients to such contrast cues.

Given a gray image I with resolution $N \times N$, its complex-valued Fourier coefficient at (u, v) can be computed as:

$$F(u,v) = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I(x,y) e^{i\theta}, \theta = \frac{-2\pi(ux+vy)}{N}, \tag{6}$$

where $u, v \in \{0, 1, \ldots, N-1\}$. We can see that the real and the imaginary parts of $F(u, v)$ can be rewritten as:

$$\begin{aligned} \Re(u,v) &= \sum_{\cos\theta \geq 0} \cos\theta I(x,y) + \sum_{\cos\theta < 0} \cos\theta I(x,y), \\ \Im(u,v) &= \sum_{\sin\theta \geq 0} \sin\theta I(x,y) + \sum_{\sin\theta < 0} \sin\theta I(x,y). \end{aligned} \tag{7}$$

From (7), we can see that Fourier transform can be reinterpreted as computing a set of template-based contrasts.

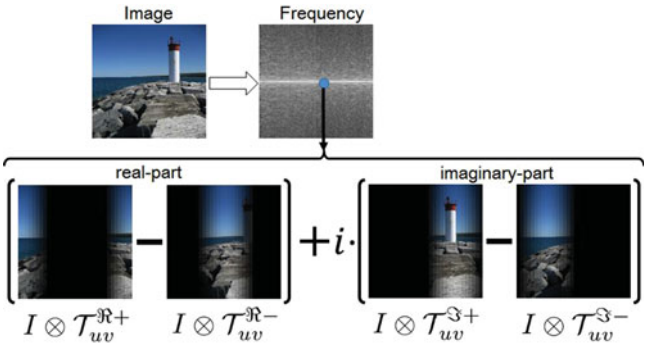


Fig. 4. Each Fourier coefficient is computed by dividing an image into two pairs of regions and computing the weighted contrast between each pair of regions.

For $\Re(u,v)$ and $\Im(u,v)$, we can define four $N \times N$ templates with non-negative coefficients:

$$\begin{aligned} \mathcal{T}_{uv}^{\Re+}(x,y) &= \max(\cos\theta, 0), \\ \mathcal{T}_{uv}^{\Im+}(x,y) &= \max(\sin\theta, 0), \\ \mathcal{T}_{uv}^{\Re-}(x,y) &= \max(-\cos\theta, 0), \\ \mathcal{T}_{uv}^{\Im-}(x,y) &= \max(-\sin\theta, 0). \end{aligned} \tag{8}$$

With these templates, we can rewrite the computation of $\Re(u,v)$ and $\Im(u,v)$ as

$$\begin{aligned} \Re(u,v) &= \left\langle I \otimes \mathcal{T}_{uv}^{\Re+} \right\rangle - \left\langle I \otimes \mathcal{T}_{uv}^{\Re-} \right\rangle, \\ \Im(u,v) &= \left\langle I \otimes \mathcal{T}_{uv}^{\Im+} \right\rangle - \left\langle I \otimes \mathcal{T}_{uv}^{\Im-} \right\rangle, \end{aligned} \tag{9}$$

where $\langle \cdot \rangle$ denotes the sum of all elements in a matrix. As shown in Fig. 4, these four templates (i.e., $\mathcal{T}_{uv}^{\Re+}$ and $\mathcal{T}_{uv}^{\Re-}$, $\mathcal{T}_{uv}^{\Im+}$ and $\mathcal{T}_{uv}^{\Im-}$) actually divide the input image into two pairs of regions. As a consequence, $\Re(u,v)$ and $\Im(u,v)$ actually represent the weighted contrasts between each pair of regions.

When DFT is applied on a gray image, there are totally $4N^2$ templates that are used to compute $2N^2$ contrast scores (as shown in Fig. 5). Consequently, the frequency spectrum stores contrast values obtained at multiple scales and directions. With these contrasts, saliency detection in the frequency domain can be viewed as finding the templates that (statistically) perform the best in capturing the difference between targets and distractors. Templates for the lowest frequencies divide images into large regions, but such partitions are often too "coarse" to accurately locate salient targets. On the contrary, templates for the highest frequencies provide "fine" partitions that achieve only high responses to noise and textures (also stated in [13]). This explains the reason that *band-pass filters perform the best in fixation prediction by discarding the lowest and highest frequencies*.

Moreover, the template-based contrast also explains the reason that the signs of the real and the imaginary parts, once correctly estimated, are highly effective in locating salient targets. Suppose there are a set of targets and distractors in the image. Given a pair of templates, in most cases one template covers more targets and the other one covers more distractors. For the real-part or the imaginary-part of each Fourier coefficient, its sign determines which template covers more targets than distractors. By assigning more energy to the region covered by the correct template in the inverse transform, we can pop-out more targets and suppress more distractors. Since the signs of the real-part and the imaginary-part are encoded in spectral phase, phase modulation becomes much more effective in locating salient objects than amplitude modulation (e.g., $\mathcal{P}_G$ is more effective than $\mathcal{A}_G$ in the quantitative experiments).

3.4 Principles for Designing Saliency Detector

From the experimental results obtained so far, we can conclude that the secret of saliency mainly hides in the phases of intermediate frequencies. Meanwhile, both spectral amplitude and phase contribute to the detection of saliency, and phase contributes more than amplitude. Inspired by the template-based reinterpretation as well as the solutions in existing frequency-domain models, we have concluded how to modulate the frequency spectrum so as to find the secret of saliency. For the sake of simplification, we use the term "saliency detection" to describe the modulation process and develops several principles for designing the saliency detector in the frequency domain:

1) Multiple complementary feature channels are preferred to fully utilize the input visual stimuli.
2) Both spectral amplitude and phase should be modulated to reach the best performance.

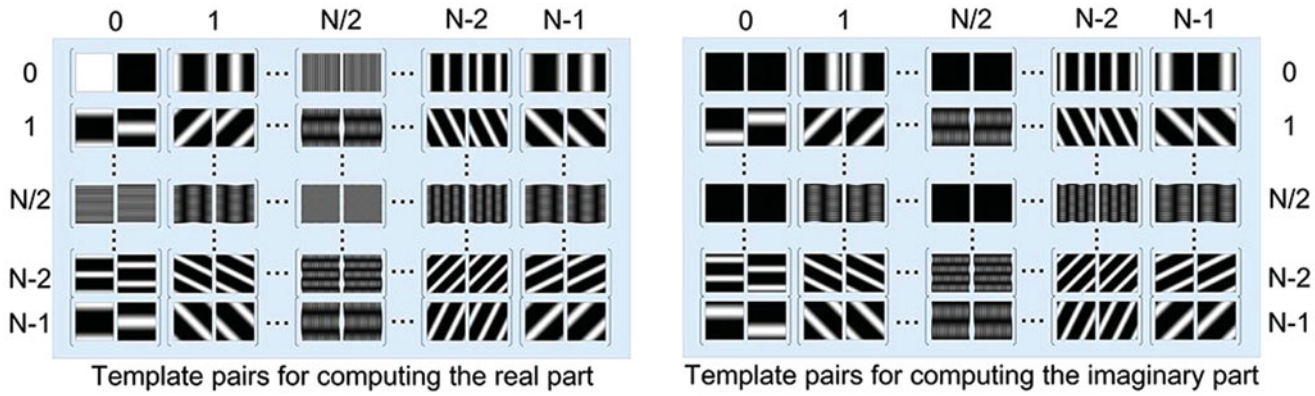


Fig. 5. Fourier transform can be interpreted as dividing an image with each of the $2N^2$ template pairs for contrast computation. As a consequence, saliency detection can be described as highlighting the most discriminative templates so as to correctly assign the energy only to salient targets in the inverse Fourier transform.

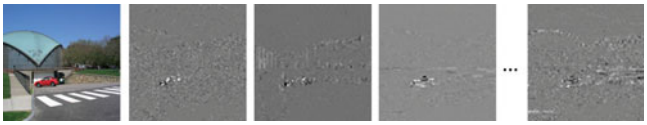


Fig. 6. Coefficient maps obtained by ICA have different visual characteristics from the original color channels.

3) Phase modulation helps to locate the salient targets, and amplitude modulation helps to clean up probable noise.
4) Intermediate frequencies should be emphasized, and the lowest and highest frequencies should be suppressed.
5) Fourier coefficients can be adjusted with respect to their neighbors for encoding template-based contrasts at similar orientations and scales.

Among these principles, the first principle is inspired by the fact that many saliency models have achieved impressive performance by using complementary features (e.g., red-green and blue-yellow color opponencies [7], [24], [53], YUV or Lab colors [25], [57] and ICA components [17], [19]). The second, third and fourth principles are directly motivated from the analysis of the experimental results discussed above. From the perspective of template-based contrast computation, we notice that there often exist some templates that perform better in capturing figure-ground contrasts than their neighboring templates (i.e., templates with similar orientations and scales). As a consequence, the corresponding Fourier coefficients can be viewed as singularities in the frequency spectrum. Inspired by this fact, the fifth principle is proposed, which aims to modulate Fourier coefficients with respect to their neighbors so as to enhance such singularities. In this manner, the corresponding templates will be emphasized in the inverse DFT, making the energy converge to salient targets other than background regions in the estimated saliency map.

4 Learning to Design Image Saliency Detector in the Frequency Domain

Following the principles proposed above, we will present how to design the saliency detector in the frequency domain. Instead of using only the heuristic spectral filters, we also put the prior knowledge obtained by unsupervised and supervised learning into consideration during the design process. With these filters, the secret of saliency can be effectively and efficiently discovered from the frequency spectrum.

4.1 Extracting Complementary Feature Channels

To design the saliency detector, several complementary feature channels should be extracted from the input image first (i.e., the first principle). Here we propose to extract complementary feature channels by using Independent Component Analysis (ICA), which can greatly remove the redundancy in adjacent pixels through unsupervised statistics. Moreover, the independence property allows us to detect visual saliency separately from each channel with simple DFT instead of using HFT [24], [25], [53]. Toward this end, we first gather 1,000 indoor/outdoor images from Flickr and sample 500 non-overlapping 8×8 patches from each image. Each patch is represented with a 192D RGB color vector, on which ICA is conducted to obtain C independent components. By jointly considering the efficiency and performance, we empirically set $C = 11$. These independent components can be viewed as some kinds of prior knowledge that help us to remove redundancy from input visual stimuli.

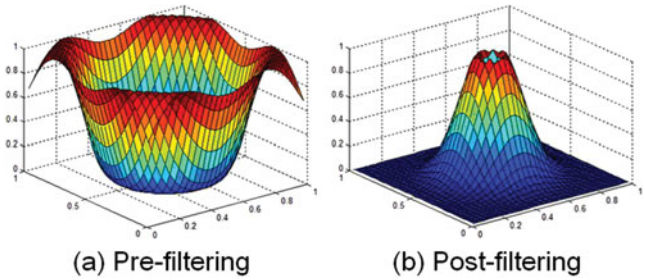


Fig. 7. Band-pass filters used in pre-filtering and post-filtering. In pre-filtering, the band-pass filter acts as a high-pass filter which also suppresses the highest frequency ($\sigma_c = 0.8$, $\sigma_w = 0.4$). In post-filtering, the band-pass filter acts as a low-pass filter that contains a notch at the zero frequency ($\sigma_c = 0.1$, $\sigma_w = 0.2$).

Usually, the coefficient maps obtained by ICA have remarkably different visual characteristics from the original color channels (as shown in Fig. 6), making it difficult to manually design the saliency detector in the corresponding spectra. Thus we propose to *train* the saliency detector for both amplitude and phase modulation (i.e., the second and third principles). Here we select 903 training images from the public image benchmark **MIT1003** [21] and leave the rest 100 images for testing purpose. Note that all the training images are resized to the same resolution of 256×256. By projecting every non-overlapping 8×8 patches from the kth image onto the cth independent component, we obtain a 32×32 coefficient map, denoted as I_{ck}. Meanwhile, we also down-sample the fixation density map of the kth image to the size of 32×32, denoted as G_k. In this manner, we can obtain C training sets with all the learned independent components, denoted as $\mathbb{T}_c = \{(I_{ck}, G_k)\}_{k=1}^{903}$ for $c = 1, \ldots, C$.

4.2 Pre-Filtering the Amplitude Spectrum

To measure image saliency on these coefficient maps, we first compute their frequency spectra through Fourier transform.[2] Inspired by the experimental results in Table 1, we adopt a Gaussian band-pass filter, denoted as H_{bh} ($\sigma_c = 0.8$, $\sigma_w = 0.4$, see Fig. 7a for the shape of this band-pass filter) to filter out the lowest frequencies and suppress the highest frequencies (i.e., the fourth principle). Thus for a training instance $(I, G) \in \mathbb{T}_c$, we have

$$F_I = \mathcal{N}(\mathcal{F}[I] \otimes H_{bh}), \quad F_G = \mathcal{F}[G], \tag{10}$$

where $\mathcal{N}(\cdot)$ denotes an operation that adjusts the frequency spectrum so as to generate only real-valued non-negative responses during inverse Fourier transform. Suppose F is a frequency spectrum and $\hat{F} = \mathcal{N}(F)$, we have

$$\hat{F}(u,v) = \frac{1}{N^2} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} e^{i\theta} \left| \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} F(u,v) e^{-i\theta} \right|. \tag{11}$$

2. Without specification, in this study the DC frequency of every spectrum is shifted to $[N/2, N/2]$ after Fourier transform and to $[0, 0]$ before inverse Fourier transform.

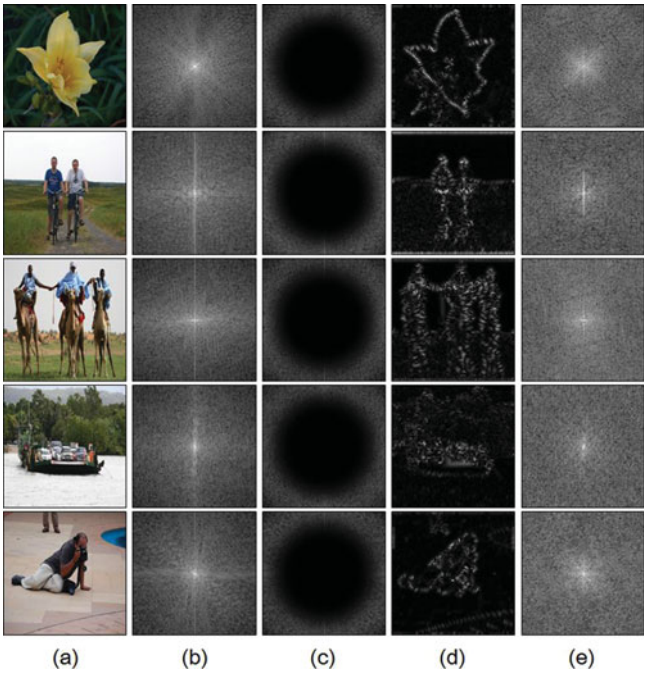


Fig. 8. Frequency spectrum should be re-adjusted after applying any operation so as to remove the probable imaginary parts generated in inverse DFT. (a)-(b) input images and amplitude spectra. (c) amplitude spectra filtered with a Gaussian band-pass filter ($\sigma_c = 0.8$,$\sigma_w = 0.4$). (d)-(e) maps obtained by inverse DFT (modulus calculated) and the new amplitude spectra.

We can see that $\mathcal{N}(\cdot)$ is equivalent to transforming the signal back to the spatial domain, computing the modulus of the complex-valued response at every pixel, and then transforming the new coefficient map back to the frequency domain (see Fig. 8 for some examples).

As stated above, Fourier coefficient $F_I(u, v)$ encodes two template-based contrasts in its real and imaginary parts. As a result, the operation in (11) is equivalent to the modulation of such contrasts. In other words, *amplitude filtering will also adjust the spectral phase*, which may be the main reason that certain amplitude modulations can also help to locate salient targets.

4.3 Learning to Design a Phase Filter

Inspired by the fact that neighboring coefficients in F_I are generated by the templates that compute contrasts at similar scales and directions, here we tentatively explore the feasibility to directly modulate the phase of each coefficient by considering the influences of its neighbours. In other words, we aim to design a convolution template with the size $M \times M$ (we empirically set $M = 3$ in this study). By convolving F_I with this template (i.e., phase filter), the phases of all coefficients are expected to approximate the corresponding phases in F_G (i.e., the fifth principle).

Different from the amplitude filter, however, it is difficult to manually design a phase filter. Therefore, we try to learn a template H_p^c from $\mathbb{T}_c$ by minimizing

$$\min_{H_p^c} \sum_{(I,G)\in\mathbb{T}_c} \mathcal{L}\Big(\mathcal{P}(F_I * H_p^c), \mathcal{P}(F_G)\Big), \tag{12}$$

where $*$ denotes convolution and $\mathcal{L}(\cdot)$ is a loss function that measures the difference between the phase spectra of

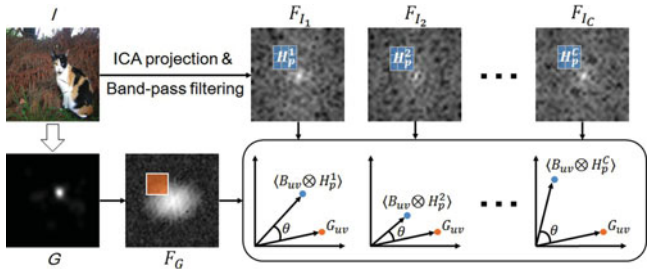


Fig. 9. A feature-specific convolution template can be learned for modulating the phase spectrum from each of the C feature channels.

$F_I * H_p^c$ and F_G. Here we first normalize F_I to let all its spectral amplitudes sum up to 1 and then sample a set of $M \times M$ coefficient patches (100 patches in this study) from F_I. Let B_{uv} be the patch centered at (u, v), we can define the loss function as

$$\begin{aligned}&\mathcal{L}\Big(\mathcal{P}(F_I * H_p^c), \mathcal{P}(F_G)\Big)\\ &\quad= \sum_{B_{uv}\in F_I} w_{uv} \cdot \ell\Big(\langle B_{uv} \otimes H_p^c \rangle, F_G(u, v)\Big),\end{aligned} \tag{13}$$

where $\ell(\cdot)$ is the Cosine distance of two complex numbers. Here $w_{uv} = |F_I(u, v)|$ is a non-negative weight to emphasize a training instance with larger amplitude since it will play a more important role in the inverse Fourier transform. By incorporating the loss function (13) into the optimization objective (12), we find that the optimization objective consists of a set of weighted Cosine distances, and the variables in H_p^c (linearly weighted and combined in each term) are the only parameters needed to be optimized. Therefore, we can solve this problem with gradient descent algorithm.

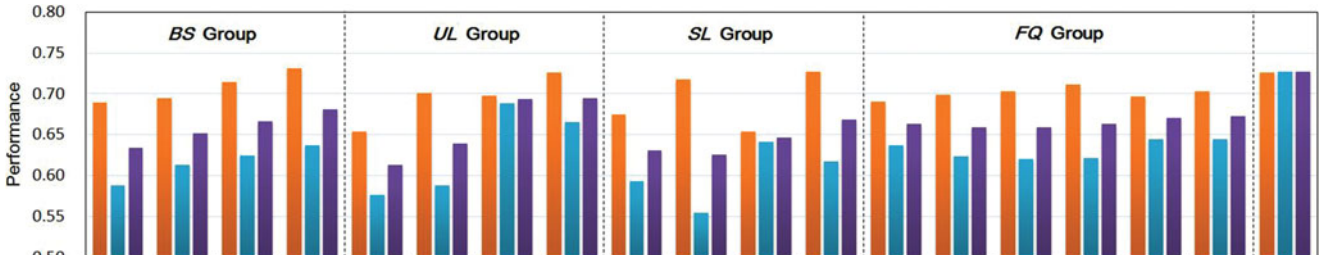
As shown in Fig. 9, the learning process aims to minimize the difference between the phase spectra of an input image and its fixation density map. Note that we learn a phase filter for each of the C feature channels (i.e., coefficient maps obtained by projecting all training images onto each of the C independent components). As stated above, the independence of these feature channels ensures that the secret of saliency is independently encoded in their frequency spectra. Thus the saliency secret in each feature channel can be best mined through the feature-specific phase filter.

4.4 Visual Saliency Estimation

Given the phase filters learned on C feature channels, we use it to mine the secret of saliency from F_I:

$$F_I^* = \mathcal{N}\big(F_I * H_p^c\big). \tag{14}$$

Note that such phase modulation will also alter the spectral amplitude, making it less accurate. Inspired by the results in Table 1, band-pass filter can be used to improve the performance when a poor amplitude spectrum (e.g., $\mathcal{A}_I$) is combined with a good phase spectrum (e.g., $\mathcal{P}_G$). Inspired by that, we apply a band-pass filter H_{bl} ($\sigma_c = 0.1, \sigma_w = 0.2$) after the phase modulation to further refine the spectral amplitude. As shown in Fig. 7b, this filter is equivalent to a Gaussian low-pass filter with a notch in frequencies around DC. Note that we apply H_{bl} *twice* on the frequency spectrum to achieve the best performance. Finally, the modulated frequency spectrum is transformed back to the spatial domain



	IT	GB	CA	BMS	SUN	AIM	ICL	SP	SNR	JUD	PMT	BST	PFT	HFT	QDCT	SIG	SR	PQFT	OUR
AUC	0.69	0.69	0.72	0.73	0.65	0.70	0.70	0.73	0.68	0.72	0.65	0.73	0.69	0.70	0.70	0.71	0.70	0.70	0.73
EOF	0.59	0.61	0.62	0.64	0.58	0.59	0.69	0.67	0.59	0.55	0.64	0.62	0.64	0.62	0.62	0.62	0.64	0.64	0.73
FS	0.63	0.65	0.67	0.68	0.61	0.64	0.69	0.69	0.63	0.63	0.65	0.67	0.66	0.66	0.66	0.66	0.67	0.67	0.73

Fig. 10. Performance of 18 saliency models on the 100 testing images of **MIT1003**. Note that the models in each group are listed from left to right with increasing FS scores.

with the moduli of complex-valued pixels being squared to obtain a conspicuous map. Note that we have C independent feature channels and all the C conspicuous maps should be combined with equal weight to generate the final saliency map.

To sum up, the saliency map S of a testing image I can be derived from the following steps,

$$\begin{aligned} F_{I_c} &= \mathcal{N}(\mathscr{F}[I_c] \otimes H_{bh}),\ \forall c \in \{1,\ldots,C\}, \\ F^*_{I_c} &= \mathcal{N}\left(F_{I_c} * H^c_p\right),\ \forall c \in \{1,\ldots,C\}, \\ S &= \sum_{c=1}^{C} \left| \mathscr{F}^{-1}\left[\mathcal{N}\left(F^*_{I_c} \otimes H_{bl}\right) \otimes H_{bl}\right]\right|^2, \end{aligned} \tag{15}$$

where I_c is the coefficient map obtained by projecting I onto the cth independent component. Note that we conduct no post-processing operation in the spatial domain, e.g., border cut, center-biased re-weighting or Gaussian smoothing. Everything is done in the frequency domain, and we simply normalize the saliency map S to the dynamic range of [0,1].

5 Experiments

To validate the effectiveness of the proposed approach, we conduct several experiments in this Section. First, we conduct extensive comparisons with state-of-the-art approaches on the 100 testing images of **MIT1003**, which is the same testing images used in [21]. To analyze the performance of various models, we conduct several small experiments on these 100 testing images as well as the 120 images in **Toronto**. In all these experiments, we adopt the same strategies as in Section 3 for sampling fixated and non-fixated pixels as well as computing the three evaluation metrics (AUC, EOF and FS).

In the experiments, we adopt 18 state-of-the-art models for quantitative comparison. The source codes or executables for these approaches are publicly available on the Internet. These approaches are selected from the four major groups introduced in Section 2, including:

- *BS* group: bottom-up models in the spatial domain, including IT [7], GB [28], CA [10] and BMS [35];
- *UL* group: models that utilize prior knowledge learned in an unsupervised manner, including AIM [17], SUN [18], ICL [19] and SP [12];
- *SL* group: models that utilize prior knowledge learned in a supervised manner, SNR [20], JUD [21], PMT [22] and BST [11];
- *FQ* group: bottom-up models in the frequency domain, including: SR [23], PFT and PQFT [24], QDCT [25], SIG [26] and HFT [53].

The performances of these 18 approaches and our approach are illustrated in Fig. 10 in terms of AUC, EOF and FS. In particular, the approaches in each group are ordered from left to right with increasing FS scores. Due to the space limitation, we only illustrate in Fig. 11 several representative examples for our approach and the best approach in each group.

From Fig. 10, we find that our approach outperforms all the other 18 approaches in terms of EOF and FS, and our AUC score is comparable with the best model in each group. Note that we conduct only simple filtering operations on the frequency spectrum at the resolution 32×32 (four filtering steps are used in total), leading to an extremely efficient algorithm. With the Matlab implementation, our approach takes only 2.01 s to process all the 100 testing images of **MIT1003** (preloaded into memory and down-sampled to the resolution 256×256) on a platform with a 3.40 GHz CPU. To sum up, our approach is very efficient and achieves the best EOF and state-of-the-art AUC.

By further investigating the quantitative results in Fig. 10 and the representative examples in Fig. 11, we believe that the success of our approach originates from the combination of prior knowledge obtained by unsupervised and supervised learning (i.e., ICA and phase filters learned from data). In other words, we assume that *unsupervised prior knowledge helps to split the secret of saliency hidden in the input signals into less-redundant feature channels, and supervised prior knowledge assists to design feature-specific detector to find the secret of saliency from each channel*. To further validate this conclusion, in the following parts of this Section we compare our model with the models from the four groups for performance analysis in detail.

5.1 Comparison with Bottom-Up Models

First, we compare our model with the bottom-up models in the *BS* and *FQ* groups. These models utilize only the bottom-up framework to process visual stimuli in the spatial/frequency domain and no prior knowledge is involved. From Fig. 10, we can see that these models achieve

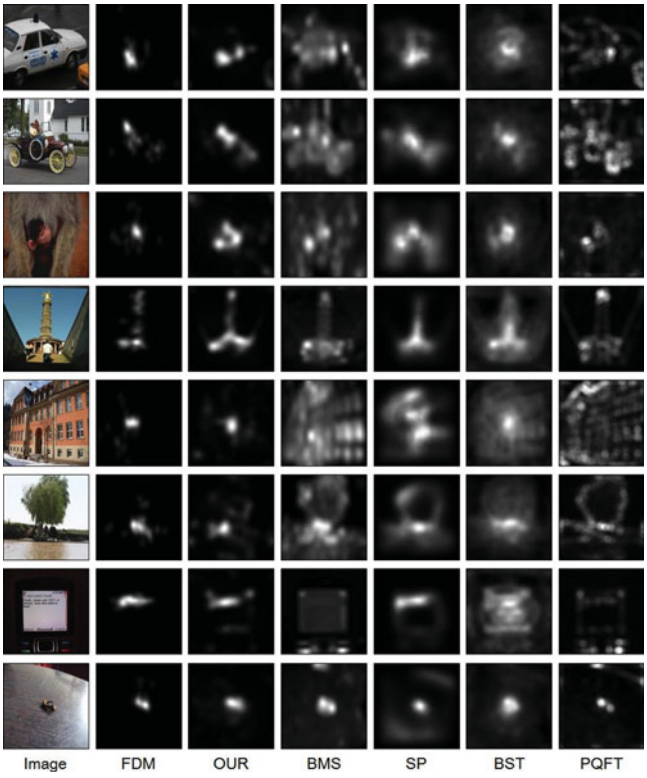


Fig. 11. Representative saliency maps generated by our model and the best model from each of the four groups. The images are from the 100 testing images of **MIT1003** and FDM indicates the fixation density map.

comparable AUC scores in [0.69, 0.73] with a standard deviation of 0.013, but their EOF scores are much lower than the EOF score of our approach. This implies that the bottom-up approaches may pop-out both targets and distractors that share common visual attributes due to the redundancy in the input visual stimuli. Thus we can assume that the independent ICA feature channels contribute a lot to the impressive performance of our approach for containing less redundancy. To validate that, we also test the performance of our approach when the following features are used: 1) Intensity, 2) Lab color channels, 3) PCA with the first 11, 54 and 192 components, and 4) ICA with 2, 5 and 20 components. The experimental results are shown in Table 2, from which we find that the number of feature channels are tightly correlated with the overall performance. In particular, the overall performance increases steadily when more

TABLE 2
Performance of Our Approach on **MIT1003** When Various Features Are Used

Feature Channels	AUC	EOF	FS
Intensity only	0.63	0.71	0.67
Lab color	0.67	0.71	0.69
PCA (11 components, 97.3 percent)	0.68	0.69	0.69
PCA (54 components, 99.7 percent)	0.73	0.71	0.72
PCA (192 components)	0.73	0.71	0.72
ICA (2 components)	0.69	0.72	0.70
ICA (5 components)	0.72	0.72	0.72
ICA (20 components)	0.72	0.71	0.72
OUR (11 ICA components)	0.73	0.73	0.73

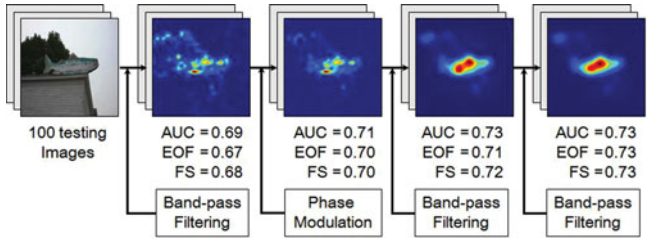


Fig. 12. Performance of our approach on 100 testing images from **MIT1003** after various filters are applied.

PCA components are used. However, the performance using 20 ICA components is slightly worse than that using 11 components. This indicates that the secret of saliency may be separately encoded in a large number of complementary feature channels, which cannot be expressed as a quaternion (e.g., PQFT, HFT and QDCT). This finding also validates the first principle for designing the saliency detector.

In particular, we also find that the six models in the *FQ* group perform unsatisfactory in AUC, which may be caused by the fact that these models (except SIG) mainly focus on the amplitude modulation. Since we have proved that the spectral phase contributes the most to locating salient targets (i.e., the third principle), such amplitude modulation is often insufficient to accurately locate salient targets. On the contrary, SIG proposes to detect saliency from the signs of DCT coefficients and outperforms the other five models in AUC. However, the secret of saliency may mainly hide in the phases of *intermediate* frequencies. Without filtering out the lowest and highest frequencies, "noise" will appear, leading to a lower EOF. Compared with these models, the success of our approach mainly arises from the phase filters learned from data. Thus an interesting concern may arise: whether models in the *FQ* group can benefit from the same ICA feature channels? To answer this question, we re-test SR, PFT and SIG on the ICA feature channels used in our approach and have their parameters manually fine-tuned to reach their best performance. We find that the FS scores of SR and PFT decrease by 2.3 and 1.2 percent, respectively. Meanwhile, the FS score of SIG stays almost unchanged (AUC decreases by 2.0 percent and EOF increases by 1.5 percent). This result may be explained by the fact that the visual characteristics of ICA features are quite different from the original color stimuli (see Fig. 6 for an example). On the contrary, our approach can adapt to such data with a learning-based framework, leading to better performance.

As shown in Fig. 12, the filters used in our approach can gradually improve the overall performance on the 100 testing images from **MIT1003** with multiple filtering operations. From Fig. 12, we find that if only the first band-pass filter is applied on the spectral amplitude, the saliency maps of all testing images in **MIT1003** reach only an AUC of 0.69 and an EOF of 0.67. After that, we have tried all the low-pass, high-pass and band-pass filters used in Section 3 and find that *none of them* can further improve the performance of our model after the first band-bass filter is applied unless we perform phase modulation. The contributions of phase modulation are two-folds: First, the phase modulation itself can improve the performance (i.e., 0.02 in

TABLE 3
Performance of Our Approach on MIT1003 When Center-Biased Gaussian Re-Weihting Is Applied

Center-biased Re-weighting	AUC	EOF	FS
Gaussian $\sigma = 0.1$	0.64	0.77	0.70
Gaussian $\sigma = 0.3$	0.71	0.74	0.73
Gaussian $\sigma = 0.5$	0.72	0.73	0.73
No re-weighting	0.73	0.73	0.73

FS on 100 testing images). Second, it becomes possible to further improve the performance with other amplitude filters after phase modulation (i.e., 0.03 in FS on 100 testing images). In other words, we can iteratively use amplitude and phase filters to refine the frequency spectrum so as to gradually improve the overall performance and avoid *local optimum*.

In addition, we conduct an experiment to validate the robustness of the proposed evaluation methodology. In the experiment, we re-weight our estimated saliency maps with different Gaussian blobs. As shown in Table 3, AUC even decreases when center-biased Gaussian re-weighting is applied. This result may be mainly caused by the non-uniform sampling strategy of non-fixated pixels. As shown in Fig. 3, both fixated and non-fixated pixels are center-biased, and a minimum distance is enforced between fixated and non-fixated pixels to avoid ambiguity. Although fixated pixels distribute nearer to image centers than non-fixated pixels, the influence of center-bias effect can be already alleviated remarkably. As shown in Table 3, it becomes very difficult to gain a remarkable improvement in AUC with simple center-biased re-weighting. Moreover, we can see from Table 3 that AUC and EOF perform as two complementary evaluation metrics when center-biased re-weighting is applied. EOF increases along with center-biased re-weighting while AUC gradually decreases, leading to stable FS. This indicates that the proposed evaluation methodology can avoid the influence of center-bias, which guarantees fair comparisons.

5.2 Comparison with Knowledge-Based Models

Beyond the bottom-up models, we also compare our model with the models that make use of the prior knowledge obtained by supervised and unsupervised learning. Typically, models in the *UL* group follow the sparse coding theory and propose to remove the redundancy in the input images with pre-trained sparse codes (or independent components, sparse functions, visual words). On the contrary, models in the *SL* group usually have to extract a large number of low-level features (e.g., local energy and center-surround contrast) and/or top-down factors such as horizontal/vertical lines and face/person/car detectors. For these features, the optimal feature combination strategy is learned to measure the contribution of each feature to saliency.

From Fig. 10, we can see that the best models from these two groups, SP and BST, have comparable AUC scores. However, models in the *UL* group usually have higher EOF than the models in the *SL* group (e.g., ICL has an EOF up to 0.69). The reason is that models in the *SL* group often extract a large number of features and many of them are redundant, leading to a lower EOF. Another problem for using such a large feature pool (e.g., 30 low-level features and four high-level features in BST) is that the computational complexity may be very high. As visual saliency estimation often acts as a preprocessing step for other applications, the efficiency should also be considered in designing saliency models.

Moreover, an important problem for models in the *SL* group is the generalization ability as they are often trained only on hundreds of images. Thus severe over-fitting risk may arise in optimizing massive parameters. On the contrary, we train only $3 \times 3 = 9$ parameters in designing each phase filter, which may alleviate the over-fitting risk to some extent (e.g., the performance of our approach reaches only $\mathrm{AUC} = 0.71$, $\mathrm{EOF} = 0.73$ and $\mathrm{FS} = 0.72$ if we train 5×5 phase filters). To further prove this, we re-train all 3×3 phase filters on the 120 images from **Toronto**. By applying these phase filters on the 100 images of **MIT1003**, we have $\mathrm{AUC} = 0.72$, $\mathrm{EOF} = 0.73$ and $\mathrm{FS} = 0.72$. The minor changes in performance indicate that our approach has lower over-fitting risk since we train only small convolution templates.

Furthermore, we find that the learned 3×3 phase filters often have strong positive responses at centers and weak negative responses at surrounding locations. This indicates that phase filtering is similar to computing center-surround contrasts in the frequency domain. From the perspective of template-based contrast computation, some templates perform better in capturing figure-ground differences than their neighbors. Therefore, convolving the frequency spectrum with the learned phase filters actually enhances the response of these templates and ensures that energy converges to salient targets in inverse DFT. This finding also validates the fifth principle for designing the saliency detector.

6 Conclusion

In this paper, we have explored the secret of image saliency in the frequency domain. Through extensive experimental studies, we find that the secret of saliency may mainly hide in the phases of intermediate frequencies. In particular, the signs of real-part and imaginary-part, once correctly estimated, have remarkable contribution to separating salient targets from distractors. By re-interpreting Fourier transform as template-based contrast computation, we present a set of principles for designing the saliency detector. Under the guidance of these principles, we have designed a saliency detector under the assistance of prior knowledge obtained by both unsupervised and supervised learning. From massive experiments, we find that the statistical information encoded in spectral amplitude and phase can be an effective cue for highly efficient visual saliency estimation.

From the results obtained so far, we find that combining supervised and unsupervised prior knowledge can be a possible research direction for visual saliency estimation. In our future work, we will explore the way to mine the prior knowledge from millions of images and to adaptively combine these two types of priories in a unique saliency model to process different types of visual scenes. Moreover, we will try to formulate the problem of visual saliency estimation as a binary classification task in the frequency domain.

ACKNOWLEDGMENTS

This work was supported in part by grants from National Natural Science Foundation of China (61370113, 61325011, 61390515 and 61421003), National Hightech R&D Program of China (2015AA016302), Supervisor Award Funding for Excellent Doctoral Dissertation of Beijing (20128000103), and Fundamental Research Funds for the Central Universities. L. Duan is the corresponding author.

REFERENCES

[1] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *Proc. ACM SIGGRAPH*, 2007, article 10.
[2] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.
[3] T. Mei, X.-S. Hua, L. Yang, and S. Li, "Videosense: Towards effective online video advertising," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 1075–1084.
[4] T. Mei, X.-S. Hua, and S. Li, "Contextual in-image advertising," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 439–448.
[5] B. Chalmond, B. Francesconi, and S. Herbin, "Using hidden scale for salient object detection," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2644–2656, Sep. 2006.
[6] Z. Li and L. Itti, "Saliency and gist features for target detection in satellite images," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2017–2029, Jul. 2011.
[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[8] Z. Zhang, J. Warrell, and P. H. Torr, "Proposal generation for object detection using cascaded ranking SVMs," in Proc. *IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 1497–1504, 2011.
[9] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3286–3293.
[10] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.
[11] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 438–445.
[12] J. Li, Y. Tian, and T. Huang, "Visual saliency with statistical priors," *Int. J. Comput. Vis.*, vol. 107, no. 3, pp. 239–253, 2014.
[13] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.
[14] J. Li, Y. Tian, L. Duan, and T. Huang, "Estimating visual saliency through single image optimization," *IEEE Signal Process. Lett.*, vol. 20, no. 9, pp. 845–848, 2013.
[15] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 631–637.
[16] J. Li and W. Gao, *Visual Saliency Computation: A Machine Learning Perspective*, 1st ed. New York, NY, USA: Springer, 2014.
[17] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 155–162.
[18] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 32, 1–20, 2008.
[19] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 681–688.
[20] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally," *Neuron*, vol. 53, pp. 605–617, 2007.
[21] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.
[22] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 150–165, 2010.
[23] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[24] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[25] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 116–129.
[26] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.
[27] C. Li, J. Xue, N. Zheng, X. Lan, and Z. Tian, "Spatio-temporal saliency perception via hypercomplex frequency spectral contrast," *Sensors*, vol. 13, no. 3, pp. 3409–3431, 2013.
[28] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 545–552.
[29] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1139–1146.
[30] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.
[31] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 733–740.
[32] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1529–1536.
[33] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.
[34] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sep. 2014.
[35] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 153–160.
[36] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–20, 2014.
[37] W. Kienzle, F. A. Wichmann, B. Scholkopf, and M. O. Franz, "A nonparametric approach to bottom-up visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 689–696.
[38] A. Borji, D. Sihite, and L. Itti, "What/where to look next? modeling top-down visual attention in complex interactive environments," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 44, no. 5, pp. 523–538, May 2014.
[39] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost," *J. Vis.*, vol. 12, no. 6, pp. 22, 1–15, 2012.
[40] J. Li, Y. Tian, T. Huang, and W. Gao, "Multi-task rank learning for visual saliency estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 623–636, 2011.
[41] G. Zhu, Q. Wang, Y. Yuan, and P. Yan, "Learning saliency by MRF and differential threshold," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2032–2043, Dec. 2013.
[42] Y. Tian, J. Li, S. Yu, and T. Huang, "Learning complementary saliency priors for foreground object segmentation in complex scenes," *Int. J. Comput. Vis.*, vol. 111, pp. 153–170, 2014.
[43] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency," in *Proc. Adv. Neuro-Inform. Process.*, 2008, pp. 251–258.
[44] X. Cui, Q. Liu, and D. Metaxas, "Temporal spectral residual: Fast motion saliency detection," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 617–620.
[45] X. Cui, Q. Liu, S. Zhang, F. Yang, and D. N. Metaxas, "Temporal spectral residual for fast salient motion detection," *Neurocomputing*, vol. 86, pp. 24–32, 2012.
[46] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.
[47] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[48] R. Peters and L. Itti, "The role of Fourier phase information in predicting saliency," *J. Vis.*, vol. 8, no. 6, p. 879, 2008.
[49] O. Buzatu and A. Savin, "Saliency based on human visual sensitivity and phase spectrum of the quaternion Fourier transform," in *Proc. Int. Symp. Signals, Circuits Syst.*, Jul. 2013, pp. 1–4.
[50] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 187–198, Feb. 2012.
[51] B. Schauerte and R. Stiefelhagen, "Predicting human gaze using quaternion dct image signature saliency and face detection," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2012, pp. 137–144.
[52] C. Li, J. Xue, N. Zheng, and Z. Tian, "Nonparametric bottom-up saliency detection using hypercomplex spectral contrast," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 1157–1160.
[53] J. Li, M. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.
[54] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
[55] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.
[56] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 9, 1–15, 2011.
[57] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 478–485.

Jia Li received the BE degree from Tsinghua University in 2005 and the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2011. He is currently an associate professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include computer vision and image/video processing. He is a member of the IEEE.

Ling-Yu Duan received the PhD degree in information technology from the University of Newcastle, Australia, in 2007. He is currently an associate professor with the School of Electrical Engineering and Computer Science, Peking University. His research interests include the areas of visual search and augmented reality, multimedia content analysis, and mobile media computing. He is a member of the IEEE.

Xiaowu Chen received the PhD degree in computer science from Beihang University, Beijing, China, in 2001. He is currently a professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His current research interests include virtual reality, computer graphics, and computer vision. He is a member of the IEEE.

Tiejun Huang received the PhD degree from the Huazhong University of Science and Technology in 1998. He is a professor with the School of Electronic Engineering and Computer Science, and the head of the Department of Computer Science, Peking University. His research areas include video coding and image understanding. He is a senior member of the IEEE.

Yonghong Tian received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2005. He is currently a professor with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing, China. His research interests include computer vision, multimedia analysis, and coding. He is a senior member of the IEEE.