frontiers in Integrative Neuroscience

REVIEW
published: 27 February 2020
doi: 10.3389/fnint.2020.00010

What Can Computational Models Learn From Human Selective Attention? A Review From an Audiovisual Unimodal and Crossmodal Perspective

Di Fu[1,2,3], *Cornelius Weber*[3], *Guochun Yang*[1,2], *Matthias Kerzel*[3], *Weizhi Nan*[4], *Pablo Barros*[3], *Haiyan Wu*[1,2], *Xun Liu*[1,2]* *and Stefan Wermter*[3]

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China, [2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China, [3] Department of Informatics, University of Hamburg, Hamburg, Germany, [4] Department of Psychology, Center for Brain and Cognitive Sciences, School of Education, Guangzhou University, Guangzhou, China

Selective attention plays an essential role in information acquisition and utilization from the environment. In the past 50 years, research on selective attention has been a central topic in cognitive science. Compared with unimodal studies, crossmodal studies are more complex but necessary to solve real-world challenges in both human experiments and computational modeling. Although an increasing number of findings on crossmodal selective attention have shed light on humans' behavioral patterns and neural underpinnings, a much better understanding is still necessary to yield the same benefit for intelligent computational agents. This article reviews studies of selective attention in unimodal visual and auditory and crossmodal audiovisual setups from the multidisciplinary perspectives of psychology and cognitive neuroscience, and evaluates different ways to simulate analogous mechanisms in computational models and robotics. We discuss the gaps between these fields in this interdisciplinary review and provide insights about how to use psychological findings and theories in artificial intelligence from different perspectives.

Keywords: selective attention, visual attention, auditory attention, crossmodal learning, computational modeling, deep learning

OPEN ACCESS

Edited by:
Valerio Santangelo, University of Perugia, Italy

Reviewed by:
Lihan Chen, Peking University, China
Riccardo Brunetti, Università Europea di Roma, Italy

***Correspondence:**
Xun Liu
liux@psych.ac.cn

Received: *03 September 2019*
Accepted: *11 February 2020*
Published: *27 February 2020*

Citation:
Fu D, Weber C, Yang G, Kerzel M, Nan W, Barros P, Wu H, Liu X and Wermter S (2020) What Can Computational Models Learn From Human Selective Attention? A Review From an Audiovisual Unimodal and Crossmodal Perspective. Front. Integr. Neurosci. 14:10. doi: 10.3389/fnint.2020.00010

1. INTRODUCTION

"The art of being wise is knowing what to overlook."
–William James, 1842-1910.

The real world is complex, uncertain and rich in dynamic ambiguous stimuli. Detecting sudden changes in the environment is significant for organisms to survive because these events need prompt identification and response (Todd and Van Gelder, 1979). Considering the limited capacity for processing information, selective attention is like a filter with the ability to remove unwanted or irrelevant information and thus optimizes a human's action to achieve the current goal (Desimone and Duncan, 1995). It is crucial as well for intelligent agents to integrate and utilize external and

internal information efficiently and to reach a signal-to-noise ratio as high as humans can (signal detection theory, SDT) (Green and Swets, 1966; Swets, 2014).

Selective attention is involved in the majority of mental activities, and it is used to control our awareness of the internal mind and the outside world. Selective attention also helps to integrate information from multidimensional and multimodal inputs (Talsma et al., 2010). Empirical research shows that stimuli with multimodal properties are more salient than unimodal stimuli; therefore, selective attention is more easily captured by multimodal inputs to promote further processing (Van der Burg et al., 2008, 2009). Selective attention is predominantly categorized by psychologists and neuroscientists into "endogenous" and "exogenous" attention. Endogenous attention helps to allocate limited cognitive resources to the current task (Posner and Snyder, 1975; Corbetta and Shulman, 2002; Styles, 2006). The metaphor for this process is described as directing a spotlight in a dark room. Such a process helps us, for instance, to search for one specific email only by glimpsing the crammed email box. However, the action can sometimes be interrupted by attractive advertisements or breaking news on a website. This latter kind of orienting attention is called exogenous attention which is usually caused by an unexpected change in the environment. It is considered to be instinctive and spontaneous and often results in a reflexive saccade (Smith et al., 2004; Styles, 2006). Another point of view distinguishes between "covert" and "overt" orienting attention: covert attention can attend events or objects with the absence of eyes movement, while overt attention guides the fovea to the stimulus directly with eyes or head movements (Posner, 1980). This is because covert attention requires inhibition of saccades to sustain fixation, which is not needed during overt attention (Kulke et al., 2016). Analogously, covert and overt mechanisms exist in the auditory system. Since humans cannot move ears like eyes, the difference between these two mechanisms is that covert auditory attention can govern attention without any motion, while overt auditory attention attends to sound sources with head movements (Kondo et al., 2012; Morillon and Baillet, 2017). Head movements contribute to sound localization during overt auditory attention (Wallach, 1940; Perrett and Noble, 1997).

To understand the mechanisms underlying selective attention is helpful for computational models of selective attention for different purposes and requirements (Das et al., 2017). Attention models have been proposed and applied in computer science for decades, and attention mechanisms have achieved high performance in sequence modeling (Vaswani et al., 2017; Peng et al., 2019). Bio-inspired implementations of attention in computer science address the limited computation capacity of machines through assigning computational resources by priority (Xu et al., 2015). However, gaps exist between computational models and theories of human selective attention. Some theories are metaphysical and mystifying, especially for readers that lack experience in humans' behavioral and neural studies. Frintrop et al. (2010) published a survey about computational visual systems with an extensive description of the concepts, theories and neural pathways of visual attention mechanisms. It is stated that "the interdisciplinarity of the topic holds not only benefits but also difficulties: concepts of other fields are usually hard to access due to differences in vocabulary and lack of knowledge of the relevant literature" (p. 1). These interdisciplinary challenges are still unsolved thus far. Additionally, the development and application of technical measurements and methods like functional magnetic resonance imaging (fMRI), Magnetoencephalography (MEG), and state-of-the-art artificial neural networks (ANN) and deep learning (DL) open up a new window for studies on humans, primates, and robots. Such new findings should be valuated and integrated into the current framework.

Although there are several review articles on selective attention in the field of both psychology and computer science (Shinn-Cunningham, 2008; Frintrop et al., 2010; Lee and Choo, 2013), most of them only focus either on a single modality or on general crossmodal processing (Lahat et al., 2015; Ramachandram and Taylor, 2017). However, it is essential to combine and compare selective attention mechanisms from different modalities together to provide an integrated framework with similarities and differences among various modalities. In the current review, firstly, we aim to integrate selective attention concepts, theories, behavioral, and neural mechanisms studied by the unimodal and crossmodal experiment designs. Secondly, we aim to deepen the understanding of the interdisciplinary work in multisensory integration and crossmodal learning mechanisms in psychology and computer science. Thirdly, we aim to bridge the gap between humans' behavioral and neural patterns and intelligent system simulation to provide theoretical and practical benefits to both fields.

The current review is organized into the following parts. Section 2 is about the existing mainstream attention theories and models based on human experimental findings and attention mechanisms in computer science. Section 3 summarizes human visual selective attention studies and introduces the modeling work in computer science inspired by psychology. Section 4 describes results on less studied auditory selective attention and the corresponding modeling work. Section 5 reviews mechanisms and models about crossmodal selective attention and state-of-the-art approaches in intelligent systems. Here, to provide focus, we select the most representative phenomena and effects in psychology: Pop-out Effect (visual attention), Cocktail Party Effect (auditory attention), and audiovisual crossmodal integration and conflict resolution (crossmodal attention). Since these effects are also well-established and often simulated in computer science, we highlight the classic and latest work. Finally, we discuss the current limitations and the future trends of utilization and implications of human selective attention models in artificial intelligence.

2. DIFFERENT THEORIES AND MODELS OF SELECTIVE ATTENTION

2.1. Classic Bottom-Up and Top-Down Control vs. Priority Map Theory

The mainstream view of selective attention proposes that there exist two complementary pathways in the brain cortex, the

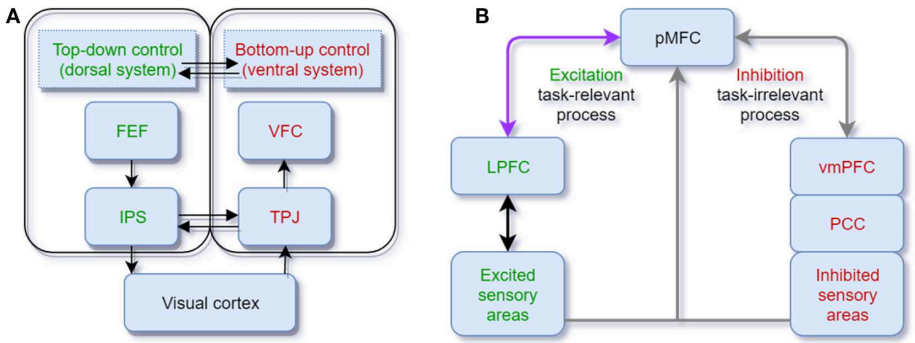

FIGURE 1 | (A) Neuroanatomical model of bottom-up and top-down attentional processing in the visual cortex. The dorsal system (green) executes the top-down attentional control. FEF, frontal eye field; IPS, intraparietal sulcus. The ventral system (red) executes the bottom-up processing. VFC, ventral frontal cortex; TPJ, temporoparietal junction (adapted from Corbetta and Shulman, 2002); **(B)** Cortical oscillation model of attentional control in visual and auditory sensory areas. The posterior medial frontal cortex (pMFC) modulates selective attention by the excitation of task-relevant processing and the inhibition of task-irrelevant processing. Theta oscillations facilitate the communication between the pMFC and lateral prefrontal cortex (LPFC) (purple arrow). Gamma oscillations and alpha oscillations are promoted in task-relevant and task-irrelevant cortical areas, respectively (gray arrows) (adapted from Clayton et al., 2015).

dorsal and ventral systems. The former, which includes parts of the intraparietal sulcus (IPS) and frontal eye field (FEF), is in charge of the top-down process guided by goals or expectations. The latter, which involves the ventral frontal cortex (VFC) and right temporoparietal junction (TPJ), is in charge of the bottom-up process triggered by sensory inputs or salient and unexpected stimuli without any high-level feedback. When novelty is perceived, the connection between the TPJ and IPS plays the role of cutting off continuous top-down control (Corbetta and Shulman, 2002) (see **Figure 1A**). The classic bottom-up and top-down control theory can explain many cases in selective attention, and a lot of computational models are based on this simple theoretical structure (Fang et al., 2011; Mahdi et al., 2019). However, in some cases, stimuli that are not relevant to the current goal, and that do not have any salient physical features can also capture attention. For instance, Anderson et al. (2011) let participants do a visual search task in the training phase to determine the direction of a line segment inside of a target. One target is associated with a high reward compared with other targets. During the test phase, that target only appears as a shape without any reward property. Participants show significantly longer reaction times doing the visual search among conditions with this foregoing high-value distractor, suggesting their attention is still captured by these goal-irrelevant stimuli. Other research finds that emotional information can also increase the salience (Vuilleumier, 2005; Pessoa and Adolphs, 2010) to capture attention. Thus, beyond the classical theoretical dichotomy, the priority map theory remedies the explanatory gap between goal-driven attentional control and stimulus-driven selection by adding the past selection history to explain selection biases (Awh et al., 2012). Here, selection history means the attention bias to stimuli that have been shown in the previous context. This bias could be irrelevant or in conflict with the current goal, so selection history should be independent of top-down or goal-driven control. In general, these two theoretical frameworks are both helpful to explain most behavioral cases of selective attention.

2.2. Functional Neural Networks Model

The Functional neural networks model separates attention into clear sub-components. Fan and Posner designed the Attentional Network Test (ANT) by combining the classic Flanker task and Posner cueing task to provide a quantitative measurement for studying the sub-components: alerting, orienting, and executive control (Fan et al., 2002, 2005; Fan and Posner, 2004). The component of the alerting network increases the focus on the potential stimuli of interest, and anatomical mechanisms of alerting are correlated with the thalamic, frontal, and parietal regions. The orienting function is responsible for selecting task-related or survival-related information from all the sensory inputs. The orienting network also determines an attention shift between exogenous attention engagement (bottom-up) and endogenous attention disengagement (top-down). Orienting is associated with the superior parietal lobe (SPL), TPJ, and frontal eye fields (FEF). The executive control component of attention plays a dominant role in planning, decision-making, conflict detection and resolution. The anterior cingulate cortex (ACC) and lateral prefrontal cortical regions are involved in the executive control component (Benes, 2000). During the ANT, participants are asked to determine the direction of the central arrow above or below the fixation. The central arrow is acommpanied by congruent or incongruent flankers. In neutral conditions, the central arrow has no flankers. There are four cue conditions: no cue, center cue, double cue, and spatial cue. Effects are calculated by subtracting participants' reaction time (RT) under two different conditions: the alerting effect = RT (no-cue) - RT (double-cue); the orienting effect = RT (center cue) - RT (spatial cue); the executive control effect = RT (incongruent flanking) - RT (congruent flanking) (Fan et al., 2002). Clinical studies using the ANT can explore specific differences of cognitive performance between patients and healthy participants (Urbanek et al., 2010; Togo et al., 2015). For example, Johnson et al. (2008) used the ANT to test children with attention deficit hyperactivity disorder (ADHD) and found that they show deficits in the alerting and executive control networks but not in the

orienting network. The model and findings arising from the ANT could serve to provide useful interventions for clinical treatment.

2.3. Neural Oscillation Model

Neural oscillations characterize the electrical activity of a population of neurons (Musall et al., 2012). Synchronization of oscillations is the coordination of firing patterns of groups of neurons from different brain areas (Varela et al., 2001). In contrast, the desynchronization of oscillations is the inhibition of neuron activities with opposite phases. Attention is correlated with synchronization and desynchronization of specific cortical neural oscillations. Clayton et al. (2015) propose a gamma-theta power-phase coupling model of attention and point out that attention is selectively adjusted via the excitation of task-relevant processes and the inhibition of task-irrelevant processes (see **Figure 1B**). The excitation of task-relevant processes is controlled by frontomedial theta (fm-theta) power (4–8 Hz) from the posterior medial frontal cortex (pMFC) to the lateral prefrontal cortex (LPFC). Among the communication between LPFC and excited sensory areas, gamma power (>30 Hz) is associated with the excitation of the task-relevant processes. The inhibition of task-irrelevant processes is linked with alpha power (8–14 Hz). The pMFC deploys the crucial inhibition processing by controlling the default mode network [posterior cingulate cortex (PCC) and ventromedial prefrontal cortex (vmPFC)] via the alpha oscillation. The limitation of the model is that the results obtained and presented across different brain regions are mainly correlations and descriptive results rather than causal relationships. Besides, most of the empirical evidence for the model was obtained by visual tasks instead of other modalities. Nevertheless, this gamma-theta power-phase coupling model shows interpretative neural pathways of the neural oscillation of selective attention.

2.4. Free-Energy Model and Information Theory

The free-energy model explains attention from a hierarchical inference perspective (Friston, 2009; Feldman and Friston, 2010). The gist of the model is that the stimuli in the living environment can be viewed as sensory inputs, surprise or uncertainty which can increase the entropy of the human brain. Our brains have a tendency to maintain the information order to minimize the energy cost caused by surprise. In doing so, perception brings about the sensory inputs, and attention infers the consequence caused by the inputs to adjust action and control the entropy growth.

Corresponding to the free-energy model, Fan's review (Fan, 2014) tries to combine the information theory and experimental neural findings to explain the top-down mechanisms of humans' cognition control (the hub of the cognition capacity) and selective attention. Inspired by the free-energy view, Fan points out that cognitive control is a high-level uncertainty or entropy reduction mechanism instead of a low-level automatic information perception. According to Shannon's information theory (Shannon, 1948), uncertainty can be quantified by entropy, and the rate of entropy is used to calculate the time density of the information transmission through different channels. Performance costs appear during cognitive channel switching. The benefits of the information theory are that attention or other cognitive processes can be quantified, and situations (like incongruent or congruent conditions in conflict processing) can be computed as bits quantitatively. Fan assimilates stimulus types, time frequency of the stimulus presentation, and human reaction time from cognitive psychology experimental tasks into entropy, surprise, and channel capacity. In this theory, if we know the probability of an event or a stimulus condition, we can calculate the surprise value of that condition and infer the information processing rate. For example, studies found that visual attention can select 30–60 bits per glimpse (Verghese and Pelli, 1992) and the upper limit of human information processing is around 50 bps. Under this framework, the anterior insula (AI) and the anterior cingulate cortex (ACC) are associated with processing the uncertain inputs and the frontoparietal cortex plays a ubiquitous role in the active control.

Research from network neuroscience takes a similar viewpoint that the brain is designed to be functioning with the lowest cost (Bullmore and Sporns, 2012; Barbey, 2018). However, the free-energy model and information theory concentrate on top-down control pathways which may fail to explain some bottom-up phenomena. For instance, why can human attention be captured by the salient external stimuli involuntarily? It can cause the rise of the information entropy and be opposite to the hypothesis that the human brain instinctively resists the disorder. Besides, experimental evidence of processing channels is still lacking.

2.5. Attention Mechanisms in Computer Science

Previous models (1980s–2014) mainly use the saliency-based winner-take-all algorithm based on human datasets to mimic humanlike visual or auditory attention (Borji and Itti, 2012; Lee and Choo, 2013). Those models aim to extract the target information from the environment or noisy background. In recent years since 2014, attention mechanisms have been applied to Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long-short Term Memory (LSTM) for sequence modeling work. Attention mechanisms were firstly used in computer vision (Ba et al., 2014) and then became widely used across different domains according to the type of input data, such as object recognition (Hara et al., 2017), image description generation (Xu et al., 2015), speech recognition (Chorowski et al., 2015), machine translation (Luong et al., 2015), video caption generation (Gao L. et al., 2017), sentiment classification (Wang et al., 2016), visual question answering (Li et al., 2018), etc.

Attention mechanisms in computer science can be distinguished as soft and hard attention (Xu et al., 2015), or as global and local attention (Luong et al., 2015). Soft attention is the expectation of selected information in the input attention distribution. For example, there is a translation task to translate one German sentence "Ich komme aus Deutschland" into an English sentence "I come from Germany." In machine translation, attention scores mean different weights assigned to words in the source sentence (German) according to each word

TABLE 1 | Main theories of visual selective attention based on various processing pathways.

Theory	Viewpoint	Processing
Stimulus-driven Theory (1992)	Singletons automatically capture visual attention	Bottom-up
Goal-driven Theory (1992)	Individuals' intentions determine attentional capture	Top-down
Contingent Capture Hypothesis (1992)	Contingent on attentional control settings induced by task demands	Top-down
Attention Selection Bias Competition (1995)	Response to distractors around the target is inhibited	Bottom-up & Top-down
Signal Suppression Hypothesis (2010)	Salience signal automatically generated by singletons can be suppressed	Bottom-up & Top-down

in the target sentence (English). In this example, corresponding to "Germany," "Deutschland" should be assigned more weights than other words in the source sentence. Soft attention focuses more broadly than hard attention. Hard attention only concentrates on information of the specific location by assigning zero weight to other information (Xu et al., 2015). The concepts of global and local attention vaguely correspond to soft and hard attention, respectively. Recently, an important application is the self-attention mechanism (Vaswani et al., 2017). Different from soft and hard attention, self-attention does not capture features of mapping between source and target but can learn the inherent structure both within the source and target text. In the above example, "from" is more likely to be followed by "Germany." Self-attention can be applied in each decoder layer of neural networks to achieve distributed processing (Bahdanau et al., 2014). In this way, self-attention shows good performance and efficiency when the input sentence is too long as in machine translation (Luong et al., 2015) or the input image is too large as in computer vision (Peng et al., 2019).

In summary, we conclude in this section that human attention is a process to allocate cognitive resources with different weights according to the priority of the events. Similarly, in computer science, attention mechanisms in models are designed to be allocating different weights to relevant input information and ignore irrelevant information with low-valued weights. However, the connection between computer science models and psychology is still loose and broad. Especially for understanding crossmodal selective attention from a functional view, it is required to explore the human cognition processing from a computational perspective, which is also beneficial for confirming psychological and biological hypotheses in computer science.

3. VISUAL SELECTIVE ATTENTION—"POP-OUT" EFFECT

3.1. Behavioral and Neural Mechanisms of Human Visual Selective Attention

Many systematic reviews in the areas of primate vision and computer vision have introduced the concepts and research findings in visual selective attention (Frintrop et al., 2010; Borji and Itti, 2012; Lee and Choo, 2013). In our current review, we further concentrate in particular on mechanisms of the "pop-out" effect and computational models based on the saliency map. In general, the "pop-out" effect describes saliency processing. Considering that an object is not salient by itself (Itti and Baldi, 2009), the "pop-out" effect usually happens when an object has more salient physical features than other objects in the context, such as location, color, shape, orientation, brightness, etc. (VanRullen, 2003). Saliency can also be extended to affective and social domains, like familiarity, threat, etc. (Fan, 2014). Humans' attention can be immediately captured by salient objects, which can explain why the warning signs on streets are always red and apparent.

Nevertheless, controversy remains about the role of top-down control when a salient stimulus captures attention. Stimulus-driven theory (bottom-up saliency hypothesis) suggests that an abrupt-onset object can automatically capture humans' attention without any intention and be processed faster than other non-onset elements (Yantis and Jonides, 1984; Theeuwes, 1991). To the contrary, the goal-driven theory (Bacon and Egeth, 1994) and the contingent capture hypothesis (Folk et al., 1992) propose that the overlap dimension between stimulus properties and task setting goals is the crucial factor, since it can determine whether the salient stimulus can be captured or not. Experiments show that if the salient stimulus has no task-relevant feature, participants adopt a feature-search mode autonomously to suppress the distraction from the salient stimulus (Bacon and Egeth, 1994).

Hybrid theories attempt to integrate components of both stimulus-driven and goal-driven theories in attention capture. Findings from monkey studies showed that attention selection through biased competition occurred when the target and the distractor were both within the receptive field. Neurons responded primarily to the target, whereas the responses to the distractor were attenuated (Desimone and Duncan, 1995). Subsequently, Mounts (2000) discovered a phenomenon named "surround inhibition." If a salient stimulus appears near the target, it can be inhibited by top-down control. Later, the signal suppression hypothesis proposed that the salient stimulus automatically generates a salience signal at first and then the signal can be subsequently suppressed, possibly resulting in no attention capture (Sawaki and Luck, 2010; Gaspelin et al., 2015, 2017) (the theories are summarized in **Table 1**).

Neural findings of humans and primates contribute a lot to understand saliency processing in the primary cortex and subcortex. The saliency map theory (Li, 1999, 2002) suggests that neurons in the primary visual cortex (V1) play a crucial role for the input feature processing during the "pop-out" effect. V1 is the neural foundation of the preattentive process during visual search, and it only responds to stimuli located in the

classical receptive fields (CRFs). In this saliency map theory, V1 is considered to define the saliency degree of visual inputs. Various features of the target and context enter into the V1 CRFs at the same time. When features of the target are more significant than the context, the target pops out. The saliency map computes the saliency value for all locations in the CRFs rather than only encoding the target location (Veale et al., 2017). In comparison to the classical feature integration model (Treisman and Gormican, 1988) and Itti's saliency model (Itti and Koch, 2000), the main property of the saliency map theory is that saliency processing is only based on a single general feature selection map rather than using a combination map to bind several individual feature maps together. Furthermore, dominant inputs from V1 convey signals to an evolutionarily old structure in the midbrain—the superior colliculus (SC). Superficial layers of the SC encode saliency representations through center-surround inhibition and transfer the inputs to deep layers to trigger priority selection mechanisms to guide attention and gaze (Stein et al., 2002; Veale et al., 2017; White et al., 2017). There is not only bottom-up processing in the primary visual cortex and SC, but also top-down processing. Within the primary visual cortex, the top-down mechanism is mediated by V2 and the interaction occurs in human V4 (Melloni et al., 2012). Moreover, deep layers of the SC represent goal-related behaviors independent of the visual stimuli (Hafed and Krauzlis, 2008; Hafed et al., 2008; Veale et al., 2017).

The large-scale human brain networks also play important roles in visual selective attention. The salience network (SN), composed of AI (anterior insula) and ACC (anterior cingulate cortex), is considered to be working as the salience filter to accept inputs from the sensory cortex and trigger cognitive control signals to the default mode network (DMN) and central-executive network (CEN). Functions of the SN are mainly about accomplishing the dynamic switch between externally and internally oriented attention (Uddin and Menon, 2009; Menon and Uddin, 2010; Uddin, 2015). Another taxonomic cingulo-opercular network shares a large overlap with the SN, containing the anterior insular/operculum, dorsal anterior cingulate cortex (dACC), and thalamus. The cingulo-opercular network has the highest cortical nicotinic acetylcholine receptor (nAChr) density, which is highly correlated with attention functions (Picard et al., 2013). However, conclusions about functions of the cingulo-opercular network are not consistent. For instance, Sadaghiani and D'Esposito (2014) revealed that the cingulo-opercular network plays a role in staying alert but not in selective attention during visual processing. In sum, the V1 and SC consist of primary cortex-subcortex pathways of saliency processing and attention orienting. The AI and ACC consist of large-scale functional networks of saliency processing, alertness and attention shifting. However, the correlation or interaction between these two pathways remains unclear.

Besides elementary physical salient features, scene regions that contain semantic meaning also proved to play a critical role in attentional guidance (Henderson and Hollingworth, 1999; Wolfe and Horowitz, 2017). Henderson and Hayes (2017) express the spatial distribution of meaning across scenes as meaning maps, which are obtained by participants' ratings of the meaningfulness of scene regions. They encode the meaning maps comparable to the image salience and operationalize the attention distribution to be duration-weighted fixation density. Their work demonstrates that both, salience and meaning, predict attention but only meaning guides attention while viewing real-world scenes. According to the cognitive-relevance theory of attentional guidance, the meaning maps contain more semantic information for the real context. Their updated findings appear to be particularly insightful and practical for artificial intelligence methods for labeling real-world images.

3.2. Computational Models Based on Human Visual Selective Attention

Based on human saccade and fixation research, a vast body of bio-inspired visual attention models has been developed and broadly applied in object segmentation (Gao G. et al., 2017), object recognition (Klein and Frintrop, 2011), image caption generation (Bai and An, 2018), and visual question answering (VQA) (Liu and Milanova, 2018). The visual attention model aims to predict the human eye fixation with minimal errors (Borji and Itti, 2012). Consistent with humans' visual processing pathways, models in visual attention are generally classified based on the bottom-up and top-down streams (Borji and Itti, 2012; Liu and Milanova, 2018). Bottom-up models are successful in modeling low-level and early processing stages (Khaligh-Razavi et al., 2017). The most classic saliency model, which uses features of color, orientation, edge, and intensity, allocates an attention weight to each pixel of the image (Itti et al., 1998; Itti and Koch, 2000) (see **Figure 2A**). The "winner-take-all" strategy is the core algorithm of saliency models. However, several criticisms on the saliency model cannot be ignored either. For instance, a salient feature is obtained by calculating the difference between input at one location and other input surrounding it so that any spatial discountinuities of features can be detected (Itti et al., 1998). This center-surround scheme is analogous to attention selection via bias competition within the visual receptive fields (Desimone and Duncan, 1995). However, the salient feature obtained by this scheme can only correspond to a small local region of an image scene with higher contrast but not to a whole object or an extended part of it (VanRullen, 2003; Lee and Choo, 2013) (also see **Figure 2A**).

In contrast, high-level task-driven attention models remain to be explored and developed further. Some research predicts human eye fixation with free-viewing scenes based on end-to-end deep learning architectures (Jetley et al., 2016; Kruthiventi et al., 2017; Kummerer et al., 2017). Deep neural networks (DNNs) have sometimes been shown to have better performance than other known models by using top-down processing mechanisms. Especially, DNNs can successfully simulate human-like attention mechanisms (Hanson et al., 2018). Here task-driven components can not only be implemented as targets but also implemented as prior knowledge, motivation, and other types of cues. Furthermore, models like DeepFeat incorporating bottom-up and top-down saliency maps by combining low- and high-level visual factors surpass other individual bottom-up and top-down approaches (Mahdi et al., 2019). Nowadays, computer vision research intends to make models learn the semantic

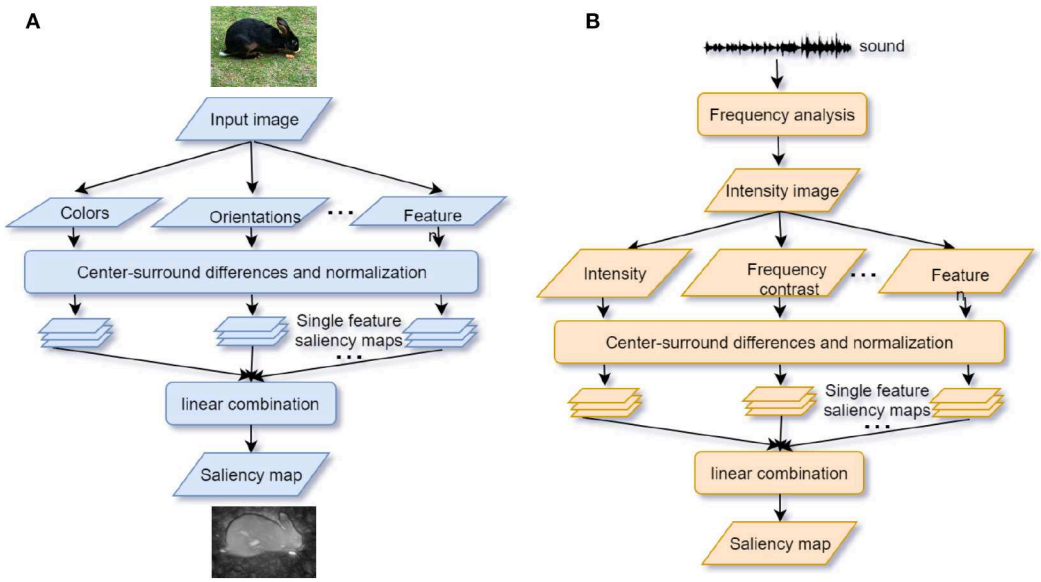

FIGURE 2 | (A) Visual saliency model. Features are extracted from the input image. The center-surround mechanism and normalization are used to generate the individual feature saliency maps. Finally, the saliency map is generated by a linear combination of different individual saliency maps (adapted from Itti et al., 1998); **(B)** Auditory saliency model. The structure of the model is similar to the visual saliency model by converting sound inputs into a frequency "intensity image" (adapted from Kayser et al., 2005).

meaning rather than simply classify objects. For instance, image captioning requires models not only to detect objects but also extract relationships between objects (Hinz et al., 2019). Co-saliency tends to be a promising preprocessing step for many high-level visual tasks such as video foreground extraction, image retrieval, and object detection. Because co-saliency implies priorities based on human visual attention, it can detect the most important information among a set of images with a reduced computational demand (Yao et al., 2017). In future research, co-saliency approaches may be combined with the meaning maps of human attention for better image interpretation accuracy.

As the number of interdisciplinary studies keeps increasing, research from psychology and artificial intelligence complement each other deepening the understanding of human visual attention mechanisms and improving the performance of computational models. On the one hand, psychologists interpret humans' behavioral or neural patterns by comparing them with the performance of DNNs. For example, Eckstein et al. (2017) found that human participants often miss giant targets in scenes during visual search but computational models such as Faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016), and YOLO (Redmon and Farhadi, 2017) do not show any similar recognizing failures. Their results suggest that humans use "missing giant targets" as the response strategy to suppress potential distractors immediately. On the other hand, computer scientists interpret features of computational models by comparing their performance with simulations of humans' behaviors. For instance, Hanson et al. (2018) found that the Deep Learning (DL) network rather than the single hidden layer backpropagation neural network can replicate human category learning. This is because hidden layers of the DL network can selectively attend to relevant category features as humans do during category learning.

4. AUDITORY SELECTIVE ATTENTION—COCKTAIL PARTY EFFECT

4.1. Behavioral and Neural Mechanisms of Human Auditory Selective Attention

At a noisy party, a person can concentrate on the target conversation (a top-down process) and easily respond to someone calling his/her name (a bottom-up process). This capability (in a real-life scenario) is named "Cocktail Party Effect" (Cherry, 1953). Auditory information conveys both temporal and spatial features of objects. For instance, we can determine whether water in a kettle is boiling by the special sounds of different heating phases. Auditory scene analysis (ASA) allows the auditory system to perceive and organize sound information from the environment (Bregman, 1994). Since humans cannot close their ears spontaneously to avoid irrelevant information, selective attention is important to segregate the forefront auditory information from a complex background and distinguish meaningful information from noise. Besides, auditory selective attention allows humans to localize sound sources and filter out irrelevant sound information effectively.

In the Cocktail Party problem, energetic masking and informational masking cause ambiguity between the auditory target and noise in the environment. Energetic masking occurs when different sound sources have overlaps in frequency spectra at the same time. The perception and recognition of the target sound can be weakened physically by noise (e.g., the target speech

overlaps with a white noise masker). Informational masking occurs when the target and masker voices sound similar (e.g., a target male is speaking while another nontarget male is speaking at the same time). The listener cannot discriminate them perceptually (Brungart, 2001; Lidestam et al., 2014). The neural mechanisms of these two causes are different. Scott et al. (2004) asked participants to listen to a target speaker with added noise (energetic masking) or added speech (informational masking). They found that informational masking was associated with the activation in the bilateral superior temporal gyri (STG) and energetic masking was associated with the activation in the frontoparietal cortex. The activation was correlated with explicit attentional mechanisms but not specifically to the auditory processing.

In accordance with the Gestalt framework, ASA is the solution to the Cocktail Party problem (Bee and Micheyl, 2008). Similar to visual processing, ASA can be separated into two components. The primitive analysis (bottom-up process) and the schema-based processing (top-down process) (Bregman, 1994). In the primitive analysis, auditory signals are separated into independent units and integrated into disparate auditory streams according to sound features and time-frequency. In the schema-based processing, prior knowledge such as language, music, other auditory memory, and endogenous attention helps to compare the auditory input signals with previous experience (Shinn-Cunningham, 2008) (see **Figure 3A**). In laboratory studies, psychologists adopt the dichotic listening paradigm to mimic the Cocktail Party problem. During the task, participants are asked to attend to the auditory materials presented to one ear and ignore the auditory materials presented to the other ear. Afterwards, participants are asked to report the information from the attended or unattended ear. Previous studies show that a higher working memory capacity (WMC) predicts a better attention focus (Conway et al., 2001; Colflesh and Conway, 2007), because a lower capacity cannot accomplish segregation and grouping of any auditory information well. Those findings are in accordance with the controlled attention theory of working memory (Baddeley et al., 1974).

Event-related potential (ERP) N1-P2 components, alpha oscillations, and frequency-following responses (FFRs) disclose how the human brain copes with the Cocktail Party problem (Du et al., 2011; Strauß et al., 2014; Lewald and Getzmann, 2015). The ERP N1 component peaks between 80 and 120 ms after the onset of a stimulus. It is sensitive to the exogenous auditory stimuli features (Michie et al., 1990). N1 (equivalent in MEG is M100) is generated from the primary auditory cortex (A1) around the superior surface of the temporal lobes (Zouridakis et al., 1998). P2 is always observed as the following component of N1. It peaks at around 200 ms after receiving the external stimulus. These early components support the early selection model of auditory attention (Woldorff et al., 1993; Broadbent, 2013; Lee et al., 2014). Alpha oscillations occur in the parietal cortex and other auditory cortical regions during spatial attention. Selective attention modulates alpha power oscillations in temporal synchrony with the sensory input and enhances the neural activity related to attended stimuli. Wöstmann et al. (2016) conducted a MEG study with a dichotic task and revealed that alpha oscillations are synchronized with speech rates and can predict the listener's speech comprehension. Scalp-recorded frequency-following responses (FFRs) are part of auditory brainstem responses (ABR). They are evoked potentials generated from the brainstem area (Mai et al., 2019). FFRs are phase-locked to the envelope or waveform of the low-frequency periodic auditory stimuli (Zhang and Gong, 2019). In the Cocktail Party problem, FFRs encode important features of speech stimuli to enhance the ability to discriminate the target stimuli from the distracting stimuli (Du et al., 2011). In summary, to exert the auditory selective attention, N1-P2 components are involved in perceiving and detecting the auditory stimuli in the early control processing; alpha oscillations and FFRs are mainly modulated by the selective control to accentuate the target and suppressing noise.

Analogous to the specialized streams of visual selective attention, there are "what" and "where" pathways in the auditory cortex (see **Figure 3B**). The ventral "what" pathway, which involves the anterolateral Heschl' gyrus, anterior superior temporal gyrus, and posterior planum temporale, is in charge of identifying auditory objects. The dorsal "where" pathway, which involves the planum temporale and posterior superior temporal gyrus (pSTG), is in charge of spatially localizing auditory objects. Within the "what" pathway, the supratemporal plane-inferior parietal lobule (STP-IPL) network dynamically modulates auditory selective attention; within the "where" pathway, the medial pSTG shows a higher-level representation of auditory localization by integrating the sound-level and timing features of auditory stimuli (Higgins et al., 2017; Häkkinen and Rinne, 2018). In addition, the "where" pathway is observed to activate around 30ms earlier than the "what" pathway implying that top-down spatial information may modulate the auditory object perception (Alain et al., 2001; Ahveninen et al., 2006). However, current studies find that functional overlaps exist in brain areas under different processing pathways, suggesting that brain areas are not function-specific (Schadwinkel and Gutschalk, 2010; Yin et al., 2014). The observed brain activities are not only stimulus-dependent but also task-dependent (Häkkinen et al., 2015). Besides, a suggested "when" pathway for temporal perception (Lu et al., 2017) deserves to be studied further because the temporal coherence is crucial for binding and segregating features into speech and speaker recognition when attention is engaged. Apart from the paralleled pathways, the distributed processing under different structures may also provide feedback to facilitate the auditory attention (Bizley and Cohen, 2013).

For the Cocktail Party problem, previous neural findings show the attentional selective mechanism occurs in different phases of information processing. Ding and Simon (2012) found that the selective mechanism exists in both top-down modulation and bottom-up adaptation during the Cocktail Party problem. When the unattended speech signals were physically stronger, attended speech could still dominate the posterior auditory cortex responses by the top-down execution. Besides, when the intensity of the target was more than 8dB louder than the background, the bottom-up neural responses only adjusted to the target speaker rather than the background

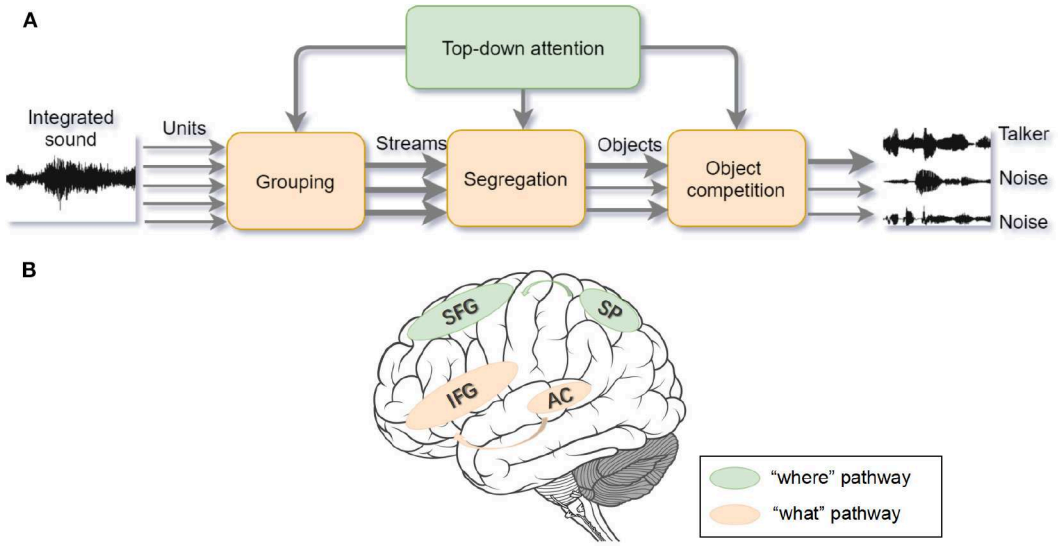

FIGURE 3 | (A) Auditory selective attention model with interaction between bottom-up processing and top-down modulation. The compound sound enters the bottom-up processing in the form of segregated units and then the units are grouped into streams. After segregation and competition, foreground sound stands out from the background noise. The wider arrow represents the salient object with higher attentional weights. Top-down attention control can modulate processing on each stage (adapted from Bregman, 1994; Shinn-Cunningham, 2008). **(B)** The "where" and "what" cortical pathways of auditory attention processing. Within the dorsal "where" pathway, the superior frontal gyrus (SFG) and superior parietal (SP) areas activate during sound localization. Within the ventral "what" pathway, inferior frontal gyrus (IFG) and auditory cortex activate to recognize the object (adapted from Alain et al., 2001).

speaker. Golumbic et al. (2013) demonstrate that the selective mechanism happens only in the high-level cortices such as the inferior frontal cortex, anterior and inferior temporal cortex, and IPL. Here, only attended speech was selectively retained. However, in the low-level auditory cortices like the STG, both attended and unattended speech were represented. In addition, one study used functional near-infrared spectroscopy (fNIRS)-hyperscanning and found that the brain-to-brain interpersonal neural synchronization (INS) selectively enhances at the left TPJ only between the listener and the attended speaker but not between the listener and the unattended speaker. The listener's brain activity overtakes the speaker's showing a faster speech prediction by the listener. Besides, the INS increased only for the noisy naturalistic conversations with competing speech but not for the two-person conversation and was only associated with the speech content. Their findings implied that the prediction of the speaker's speech content might play an important role in the Cocktail Party Effect (Dai et al., 2018). In summary, the human brain's auditory processing during the Cocktail Party problem is not hierarchical but heterarchical, which is mainly a bottom-up process aided by top-down modulation (Bregman, 1994). This includes interactions between different pathways and adaptations to the environment (Shinn-Cunningham, 2008; Bizley and Cohen, 2013).

4.2. Computational Models for the Human Cocktail Party Problem Solution

In the future, we may have moving robots offering food and drinks in noisy restaurants by precisely localizing speaking customers. Steps to solve the Cocktail Party problem in computer science can be mainly separated into: speech separation, sound localization, speaker identification, and speech recognition. The aims of acoustic models for the Cocktail Party problem are: identifying multiple speakers and disentangling each speech stream from noisy background. Numerous classical acoustic models are data-driven and based on algorithms of signal processing (Dávila-Chacón et al., 2018). Those models are robust and with good accuracy but lack the prior knowledge, biological plausibility and rely on the large datasets. Currently, models inspired by the human auditory attention system rely on smaller datasets and have shown improved adaptation. In this section, we focus on the following bio-inspired models: (1) computational auditory scene analysis (CASA): neural oscillator models as examples; (2) saliency models; (3) top-down- and bottom-up-based models.

Based on the Gestalt framework (Rock and Palmer, 1990), the goal of most CASA models is to segregate sounds with similar patterns or connections and group them into independent streams from the mixed auditory scene. Stemming from CASA models, neural oscillator models show good adaptation in auditory segregation. Neural oscillator models perform stream segregation based on the oscillatory correlation. Attention interest is modeled as a Gaussian distribution across the attended frequency. The attentional leaky integrator (ALI) consists of the connection weights between oscillators and the attentional process. The synchronized oscillators activate the ALI to separate sounds into streams like the endogenous attention mechanism (Wrigley and Brown, 2004). Furthermore, to make use of the temporal proximity of sound sources, Wang and Chang (2008) propose a two-dimensional (time and frequency) network

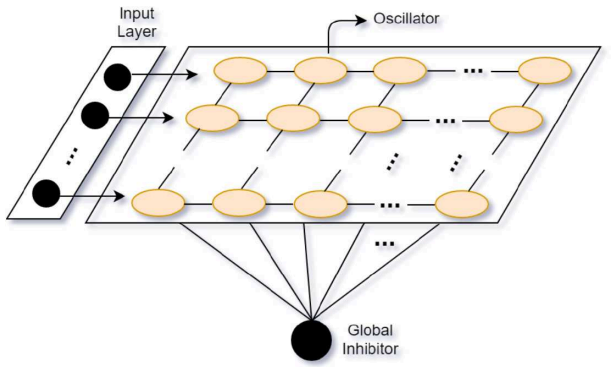

FIGURE 4 | Locally Excitatory, Globally Inhibitory Oscillator Network (LEGION) (adapted from Wang and Terman, 1995).

oscillator model with relaxation oscillators of local excitation and global inhibition (see Locally Excitatory, Globally Inhibitory Oscillator Network, LEGION; Wang and Terman, 1995) (see **Figure 4**). Analogous to humans' neural oscillations, the local excitation mechanism makes oscillators synchronize when they are stimulated by the same stimuli and the global inhibition has an effect on the whole network to make oscillators desynchronize by different stimuli (Dipoppa et al., 2016). In their model, they use alternating sequences of high- and low-frequency tones as inputs. Tones with similar patterns (e.g., close frequency, onset or offset time) tend to be grouped into the same stream. One stream corresponds to an assembly of synchronized neural oscillators. The oscillator models mimic the human selective attentional control and show good adaptation to separate the multi-tone streams.

The oscillator models try to mimic the endogenous attentional control while the saliency models try to mimic the exogenous attention. Similar to visual saliency models (see section 3.2), auditory saliency models are built by abstracting features (intensity, frequency contrast, and temporal contrast) from the sound "intensity image," which is a visual conversion of auditory time-frequency spectrograms and normalized to be an integrated saliency map (Kayser et al., 2005; Kalinli and Narayanan, 2007) (see **Figure 2B**). Considering that humans and other primate animals can process the pure auditory signals without any visual conversion, Kaya and Elhilali (2012) modify the auditory saliency model by directly extracting the multi-dimensional temporal auditory signal features (envelope, frequency, rate, bandwidth, and pitch) of the auditory scene as input. Their model relies on the selection of parameters to reduce error rates of the saliency determination by fewer features. Several limits exist for developing the auditory saliency models. Firstly, unlike visual attention, acoustic signals are distributed across different frequency bands and time windows. This makes auditory models rely much on temporal features. There is no apparent physical marker for a person to locate sounds compared with eye gaze used in visual saliency models. Secondly, in some cases differences between the saliency of sound streams are not apparent enough for the auditory saliency models to discriminate (e.g., separating one girl's voice from a group of chatting girls). Therefore, more high-level features or top-down modulation could be helpful for a model to indicate the significant sound stream. To integrate both endogenous and exogenous attention in the model, Morissette and Chartier (2015) propose a model by extracting frequency, amplitude, and position as features and connecting them with the oscillator model LEGION. Segments with consistent features are bound into the saliency map according to the temporal correlation. Notably, a module of inhibition-of-return (IOR) is inserted to inhibit attention from fixing at the most salient scene for a long time. This mechanism can achieve the attentional shifting and orientation (Klein, 2000).

Prior knowledge (e.g., memory, prediction, and expectation) also plays a crucial role in human auditory perception, therefore several top-down- and bottom-up-based models integrate the prior knowledge into the data-driven models. Some of them extract acoustic features of the target sound and store them in memory-like modules to mimic humans' long-term memory function as top-down modulation. Oldoni et al. (2013) combine a self-organized map (SOM) of the acoustic features in the bottom-up processing to continuously learn the saliency and novelty of acoustic features. After training, each SOM unit matches up with a representative sound prototype. For the top-down processing, the IOR and LEGION mechanisms are introduced to shift and select attention, respectively. Xu et al. (2015) propose an Auditory Selection framework with Attention and Memory (ASAM). In this model, there is one speech perceptor extracting the voiceprint of speakers and accumulating the voiceprint in a lifelong-learning memory module during the training phase to be the prior knowledge for the model. Later, the learned voiceprint is used to attend and filter the target speech from the sound input to achieve the top-down and bottom-up interaction. The testing performance showed good robustness and adaptation for both top-down (follow a specific conversation) and bottom-up (capture the salient sound or speech) attention tasks.

Shi et al. (2018) propose the Top-Down Auditory model (TDAA) and use the prediction of the target speaker as the top-down modulation. Their model contributes to the auditory scene analysis with multiple unknown speakers. They adopt the Short-Time Fourier Transformation (STFT) and Bidirectional Long-Short Term Memory (BiLSTM) to predict the number of the speakers. Later, the classifier recurrent neural networks (RNN) separate the most salient speaker and iterate until no more speakers can be separated to avoid repeated prediction. Finally, an attention module is used to separate each speaker's spectrum from the spectrum mixture. Besides, binaural models are apt to make use of the spatial localization information to address the Cocktail Party problem. For instance, Ma et al. (2018) train DNNs to localize acoustic features in full 360° azimuth angles. After the training phase, the binaural localization with spectral features is used as prior knowledge in the top-down modulation of the model. Their model serves to predict the speech with different localizations under noisy situations with room reverberation. In summary, those top-down and bottom-up interaction models incorporate mechanisms of processing in the human auditory

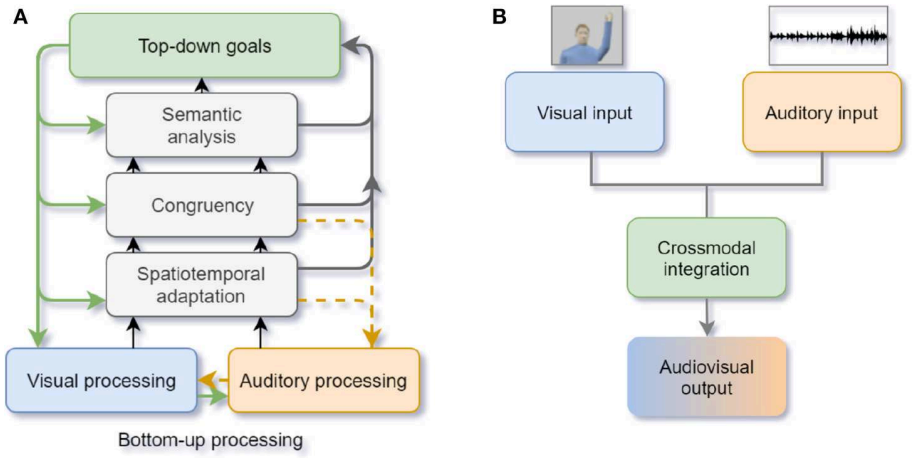

FIGURE 5 | (A) Human crossmodal integration and attentional control. The black and gray arrows denote the feed-forward bottom-up stimulus saliency processing and the green arrows denote the top-down modulation of attention. The yellow dashed arrows represent the recurrent adjustment (adapted from Talsma et al., 2010); **(B)** Artificial neural networks of crossmodal integration. The crossmodal integration mechanisms are used to realign the input from visual and auditory modalities (adapted from Parisi et al., 2017, 2018).

system. They selectively attend or shift attention to the target speech dynamically rather than only focusing on the stream separation, which can be more adaptive to those uncertain and complex auditory scenarios.

5. AUDIOVISUAL CROSSMODAL SELECTIVE ATTENTION

5.1. Behavioral and Neural Mechanisms of Human Crossmodal Selective Attention

In order to survive in an uncertain and multimodal world, humans develop the ability to integrate and discriminate simultaneous signals from multiple sensory modalities, such as vision, audition, tactile, and olfaction. For example, humans can make use of visual cues like lip movement and body gestures to recognize and localize sounds in noisy circumstances. The crossmodal integration ability is beneficial for humans to localize and perceive objects but can also cause ambiguity. Crossmodal conflicts arise when information from different modalities are incompatible with each other and can result in failures of the crossmodal integration and object recognition. To resolve conflicts, selective attention is required to focus on the task-relevant modality information and to ignore the interference from irrelevant modalities (Veen and Carter, 2006). For humans, the capacity for conflict adaptation plays a crucial role in learning and adapting to the environment. When human toddlers detect any conflict between the current environment and their prior knowledge, they will generate curiosity and be motivated to learn new knowledge or rules (Wu and Miao, 2013). Curiosity is also important for the trial and error learning of robots (Hafez et al., 2019). In this subsection, we mainly talk about behavioral and neural mechanisms of selective attention underlying audiovisual crossmodal integration and conflict resolution.

First, how and when does a crossmodal conflict occur? Previous studies proved that humans tend to integrate visual and auditory stimuli with spatial-temporal linkage into the same object (Senkowski et al., 2008). The "Unity assumption" proposes that when humans believe that the multisensory inputs they perceive are generated from the same source, crossmodal integration occurs (e.g., when students think the speech they hear in the lecture room matches the lip movements of the professor, they believe that the speech is from the professor) (Roseboom et al., 2013). Besides, prior knowledge and experience can generate expectation effects to facilitate object recognition during crossmodal integration. Therefore, when the stimuli from different modalities are spatially (e.g., ventriloquism effect; Choe et al., 1975) or temporally incongruent (e.g., double flash illusion; Roseboom et al., 2013) or contrary to our expectations (e.g., see a cat with a "bark" sound), humans perceive crossmodal conflicts. During the early integration stage, selective attention plays the role of capturing the salient visual and auditory stimuli by bottom-up processing. When conflicts are detected, selective attention executes a top-down modulation from higher-level semantic representations according to the internal goal and relevant modalities. The crossmodal information processing is not only a feed-forward process but also contains backward feedback and recurrent processes, which are important to facilitate the primary sensory processing (Talsma et al., 2010; see **Figure 5A**).

Second, which modality dominates when humans are confronted with audiovisual conflicts? Lots of studies have examined the "ventriloquism effect," which originally refers to the strong visual bias during the sound localization (Thurlow and Jack, 1973; Choe et al., 1975; Warren et al., 1981). Research findings show that this strong modality bias changes through the lifespan of a human (Sloutsky, 2003). Compared to toddlers, adults are more likely to have visual stimuli preferences (Sloutsky, 2003). Some researchers argue that the ventriloquism effect

results from an optimal or suboptimal decision-making strategy, especially when unimodal stimuli are blurred. If the auditory stimuli are more reliable than the visual stimuli, an auditory bias occurs as well (Alais and Burr, 2004; Shams and Kim, 2010; Ma, 2012; Roseboom et al., 2013). To sum up, vision in general has a higher spatial resolution than audition, whereas audition has a higher temporal resolution than vision. As the modality appropriateness hypothesis points out, the information from one modality dominates according to the temporal or spatial features of the audiovisual event and the modality with the higher accuracy (Welch and Warren, 1980).

Third, how do humans resolve crossmodal conflicts? In the conflict-monitory theory, the module of conflict monitoring (CM) is activated when conflicts are detected and passes the signal to the executive control (EC) module to accomplish the task-related conflict resolution by the top-down attentional control (Botvinick et al., 2001). From the previous findings, to perceive crossmodal signals and detect crossmodal conflicts, selective attention plays the role of gating crossmodal coupling between sensory function areas in a modality-general fashion (Eimer and Driver, 2001; Mcdonald et al., 2003; Convento et al., 2018). However, to solve crossmodal conflicts, selective attention inclines toward processing in a modality-specific fashion (Yang et al., 2017; Mengotti et al., 2018).

Except for some specific vision and audition processing brain areas, the superior colliculus (SC) is a crucial brain area with multisensory convergence zones from visual and auditory primary cortices to higher-level multisensory areas. As it is mentioned in section 3.1, the SC also implements selective attention by orienting both covert and overt attention toward the salient stimulus and triggers corresponding motor outputs (e.g., eye movements, saccades) (Wallace et al., 1998; Meredith, 2002; Krauzlis et al., 2013). Besides, the superior temporal sulcus (STS), inferior parietal sulcus (IPS), frontal cortex (including premotor and ACC), and posterior insula are involved in the crossmodal processing (for review see Calvert, 2001; Stein and Stanford, 2008). Within the crossmodal brain functional network, the STS plays the role of linking unimodal representations (Hertz and Amedi, 2014). The parietal lobe is thought to process representations of visual, auditory, and crossmodal spatial attention (Farah et al., 1989). However, when audiovisual inputs are incongruent, crossmodal attenuations or deactivations occur (Kuchinsky et al., 2012). To resolve conflicts, as human fMRI studies have shown, the dorsal anterior cingulate cortex (dACC) is responsible for dealing with conflicts between the current goal and irrelevant distractors. The dACC is positively correlated with attention orientation and interference suppression (Weissman et al., 2004). Song et al. (2017) conducted a mice experiment by using a task with audiovisual conflicts, where audition was required to dominate vision. They found that when the conflict occurred, the co-activation of the primary visual and auditory cortices suppressed the response evoked by vision but maintained the response evoked by audition in the posterior parietal cortex (PPC).

Electrophysiological studies have shown the existence of cells that respond to stimulation in more than one modality to accomplish crossmodal integration and conflict resolution. Diehl and Romanski (2014) found that neurons in the ventrolateral prefrontal cortex (VLPFC) of Macaques were bimodal and nonlinear multisensory. When incongruent faces and vocalizations were presented, those neurons showed significant changes with an early suppression and a late enhancement during the stimulus displaying period. Other experimental evidence argues that coherent oscillations across different modality cortices are the key mechanism of the crossmodal interplay (Wang, 2010). An enhancement of the phase locking for the short-latency gamma-band activity (GBA) is found for the attended multisensory stimuli. The early GBA enhancement enables the amplification and integration of crossmodal task-relevant inputs (Senkowski et al., 2008). Incongruent crossmodal inputs cause a stronger gamma-band coherence than congruent inputs suggesting the involvement of gamma oscillations decoupling under crossmodal binding (Misselhorn et al., 2019). Attentional control during the crossmodal integration and conflict resolution is associated with alpha-band effects from the frontoparietal attention network rather than primary sensory cortices. Frontal alpha oscillations are involved in the top-down perceptual regulation; parietal oscillations are involved in the intersensory reorientation (Misselhorn et al., 2019). Reversed to the gamma oscillation patterns, incongruent conditions showed weaker alpha oscillation changes compared to congruent conditions. This gamma-alpha oscillation cycle pattern is proposed to be the information gating mechanism by inhibiting task-irrelevant regions and selectively routing the task-relevant regions (Jensen and Mazaheri, 2010; Bonnefond and Jensen, 2015). In sum, cortical areas that have multimodal convergence zones accomplish crossmodal integration of projections from visual and auditory primary cortices. Neural oscillations coordinate the temporal synchronization between the visual and auditory modality.

5.2. Computational Models Simulating Human Crossmodal Selective Attention

In robotics, crossmodal research focuses mainly on multisensory binding to make robots interact with the environment with higher robustness and accuracy. Compared with unimodal information, crossmodal information is more beneficial to model complex behaviors or achieve high-level functions on artificial systems, such as object detection (Li et al., 2019), scene understanding (Aytar et al., 2017), lip reading (Mroueh et al., 2015; Chung et al., 2017), etc. In psychology, crossmodal research focuses on how crossmodal information helps humans to recognize objects or events by integrating multimodal information and eliminating the crossmodal ambiguity (Calvert, 2001). In computer science, crossmodal research focused on recognizing one modality by using a multimodal dataset or making use of the data from one single modality and retrieve relevant data of other modalities (Skocaj et al., 2012; Wang et al., 2017). However, compared with unimodal, computational modelings based on crossmodal attention remains lacking. In this section, we particularly introduce the undeveloped computational modeling work on selective attention from the audiovisual crossmodal perspective.

Many studies focus on multimodal fusion (Ramachandram and Taylor, 2017), but research about crossmodal selective attention in computer science is limited. Parisi et al. conducted a series of audiovisual crossmodal conflict experiments to explore human selective attention mechanisms in complex scenarios (Parisi et al., 2017, 2018; Fu et al., 2018). During human behavioral tasks, visual and auditory stimuli were presented in an immersive environment. Four loudspeakers were set behind the corresponding positions on a 180-degree screen, where four human-like avatars with visual cues (lip movement or arm movement) were shown. The visual cue and the sound localization could be congruent or incongruent (e.g., the left-most sound with the right-most avatar's lip movement). During each trial, human participants were asked to determine where the sound was coming from. Participants had to pay attention to the sound localization and suppress the attentional capture by any visual stimuli. Analyses of human behavior results showed that even though arm moving was visually more salient than lip moving, humans had higher error rates of the sound localization when viewing lip movement. This suggests that lip moving might contain more speech or semantic information so it is more difficult to be ignored. Besides, the magnitude of the visual bias was also significant when the incongruent AV stimuli were coming from the two avatars at the extreme right and left sides of the screen. This indicated a wider integration window than other simplified scenes. Based on the bio-inspired cortico-collicular architecture, deep and self-organizing neural networks consisting of visual and auditory neuron layers and crossmodal neuron layers were used to learn crossmodal integration and selective attention (see **Figure 5B**). In this way, human-like responses were modeled and embedded in an iCub robot.

The work above shows that computational models can simulate human selective attention on audiovisual sound localization and semantic association. Due to the limited resources and sensory modules, the future exploration of modeling and simulating the attention module is desirable in crossmodal robotics. Besides, selective attention mechanisms can boost the applicability and accuracy of robots in real human-robot interaction scenarios. Robots can select more reliable modalities and reduce distraction and errors.

6. CONCLUDING REMARKS AND OUTSTANDING QUESTIONS

The current review summarizes experimental findings, theories, and model approaches of audiovisual unimodal and crossmodal selective attention from psychology, neuroscience, and computer science perspective. Currently, psychologists and neural scientists are working toward computational modeling, standardizing, and replication. In parallel, computer scientists are trying to design and make agent systems more intelligent with higher-level cognitive functions, meta-learning abilities, and lower learning costs. Some advantages, unresolved problems, and future directions of collaborative research in psychology, neuroscience, and computer science are summarized as follows:

6.1. How Psychology, Neuroscience, and Computer Science Benefit From Each Other

One the one hand, findings and methods from psychology and neuroscience can interpret and improve models' performance (Hohman et al., 2018). For instance, representational similarity analysis (RSA) is nowadays also used to compare the responses recorded in fMRIs and artificial systems like deep learning CNNs. RSA analyzes the similarity of fMRI responses and brain representations by a set of stimuli (Kriegeskorte et al., 2008). Dwivedi and Roig (2019) found that RSA shows good performance on transfer learning and task taxonomy by computing correlations between the models on certain tasks. On the other hand, the-state-of-the-art approaches offer tools to analyze big data of neural findings. For example, the SyConn framework used deep CNNs and random forest classifiers to accelerate data analyses on animal brains to compute the synaptic wiring of brain areas (Dorkenwald et al., 2017). Another application of computational modeling is examining theories and interpreting mechanisms in human behaviors or neural responses (O'Reilly, 2006). The key idea is to examine crucial cognitive function in hidden layers of the modal. Models can be built to simulate normal behaviors and then mimic the "damage" by changing parameters of sub-units. If the "damage" causes similar abnormal behaviors as psychiatric patients do, the changed units may be the corresponding mechanisms to the behaviors. For instance, Wang and Fan (2007) collected human behavioral data by the ANT and used leabra (local, error-driven, and associative, biologically realistic algorithm) model (O'Reilly, 1998) to explore the potential interaction between each functional network (alerting, orienting, and executive control). Their model successfully simulated healthy human behavior. After changing one parameter of the executive control module, their model could simulate the behavior of schizophrenic patients, suggesting the crucial role of executive control.

6.2. Limits Remain in Current Interdisciplinary Research

Even though we have reviewed and summarized a number of findings from psychology and computer science, lots of unsolved issues of attention processing remain to be disclosed. The simulation work of crossmodal attention and conflict processing is insufficient on robots. Besides, the problem of perceptual constancy has not been deeply addressed in computer science. For humans, it is easy to recognize one object from different perspectives, such as finding an open door in a dim room. Moreover, humans can transfer the intrinsic knowledge to learn and infer new objects or concepts with a small number of learning samples. However, artificial intelligent systems cannot reach humans' performance yet. For example, even though the scale-invariant feature transform (SIFT) algorithm (Lowe, 1999) can extract features from variant shapes of the same object, it cannot recognize the variant objects when only colors exist without any structural patterns. Current deep learning approaches like the VGG net (Simonyan and Zisserman, 2015) has shown better

performance on object recognition than traditional approaches. However, such deep networks rely on the training dataset and need substantial computational resources.

6.3. Future Directions for Interdisciplinary Research

There is a lot of potential for psychologists and computer scientists to work together to investigate both human cognition and intelligent systems. On the one hand, psychologists can focus on designing paradigms to diagnose and remedy shortages of current models to improve the model accuracy. Besides, neural studies are still needed to understand human brain mechanisms better. It will be insightful to develop bio-inspired computational models with a better interpretability. On the other hand, for computer science, enhancing the complexity of models to increase the adaptivity and flexibility to the environment is required. At last, to balance the computational complexity and biological plausibility is also crucial, because humans' behavioral patterns are limited by their capacity and energy load, even though the properties of machines will keep improving. In summary, deepening the understanding of each processing mechanism rather than only describing phenomena is the direction for research from both sides to endeavor.

AUTHOR CONTRIBUTIONS

SW and XL contributed to the conception and organization of the manuscript. DF wrote the first draft of the manuscript. CW, GY, MK, WN, PB, HW, XL, and SW authors contributed to manuscript reading and revision and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (NSFC, No. 61621136008), the German Research Foundation (DFG) under project Transregio Crossmodal Learning (TRR 169), and CAS-DAAD joint fellowship.

ACKNOWLEDGMENTS

We especially appreciate feedback and support from the Knowledge Technology Group in Hamburg and the Carelab in Beijing. We thank Katja Kösters for proofreading the manuscript and improving the language clarity. We also particularly thank Dr. German I. Parisi, Ge Gao, Zhenghan Li, Honghui Xu, and Antonio Andriella for the fruitful discussions.

REFERENCES

Ahveninen, J., Jääskeläinen, I. P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., et al. (2006). Task-modulated "what" and "where" pathways in human auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14608–14613. doi: 10.1073/pnas.0510480103

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (2001). "What" and "where" in the human auditory system. *Proc. Natl. Acad. Sci. U.S.A.* 98, 12301–12306. doi: 10.1073/pnas.211209098

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Anderson, B. A., Laurent, P. A., and Yantis, S. (2011). Value-driven attentional capture. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10367–10371. doi: 10.1073/pnas.1104047108

Awh, E., Belopolsky, A. V., and Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends Cogn. Sci.* 16, 437–443. doi: 10.1016/j.tics.2012.06.010

Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., and Torralba, A. (2017). Cross-modal scene networks. *IEEE Trans. Patt. Anal. Mach. Intell.* 40, 2303–2314. doi: 10.1109/TPAMI.2017.2753232

Ba, J., Mnih, V., and Kavukcuoglu, K. (2014). "Multiple object recognition with visual attention," in *International Conference on Learning Representations* (Banff, AB).

Bacon, W. F., and Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Percept. Psychophys.* 55, 485–496. doi: 10.3758/BF03205306

Baddeley, A., Hitch, G., and Bower, G. (1974). Recent advances in learning and motivation. *Work. Mem.* 8, 647–667.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations* (Banff, AB).

Bai, S., and An, S. (2018). A survey on automatic image caption generation. *Neurocomputing* 311, 291–304. doi: 10.1016/j.neucom.2018.05.080

Barbey, A. K. (2018). Network neuroscience theory of human intelligence. *Trends Cogn. Sci.* 22, 8–20. doi: 10.1016/j.tics.2017.10.001

Bee, M. A., and Micheyl, C. (2008). The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? *J. Compar. Psychol.* 122, 235–251. doi: 10.1037/0735-7036.122.3.235

Benes, F. M. (2000). Emerging principles of altered neural circuitry in schizophrenia. *Brain Res. Rev.* 31, 251–269. doi: 10.1016/S0165-0173(99)00041-7

Bizley, J. K., and Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nat. Rev. Neurosci.* 14, 693–707. doi: 10.1038/nrn3565

Bonnefond, M., and Jensen, O. (2015). Gamma activity coupled to alpha phase as a mechanism for top-down controlled gating. *PLoS ONE* 10:e0128667. doi: 10.1371/journal.pone.0128667

Borji, A., and Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Trans. Patt. Anal. Mach. Intell.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652. doi: 10.1037/0033-295X.108.3.624

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press.

Broadbent, D. E. (2013). *Perception and Communication*. Amsterdam: Elsevier.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* 109, 1101–1109. doi: 10.1121/1.1345696

Bullmore, E., and Sporns, O. (2012). The economy of brain network organization. *Nat. Rev. Neurosci.* 13, 336–349. doi: 10.1038/nrn3214

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123. doi: 10.1093/cercor/11.12.1110

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Choe, C. S., Welch, R. B., Gilford, R. M., and Juola, J. F. (1975). The "ventriloquist effect": visual dominance or response bias? *Percept. Psychophys.* 18, 55–60. doi: 10.3758/BF03199367

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems* (Montreal, QC), 577–585.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE), 3444–3453.

Clayton, M. S., Yeung, N., and Kadosh, R. C. (2015). The roles of cortical oscillations in sustained attention. *Trends Cogn. Sci.* 19, 188–195. doi: 10.1016/j.tics.2015.02.004

Colflesh, G. J., and Conway, A. R. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. *Psychon. Bull. Rev.* 14, 699–703. doi: 10.3758/BF03196824

Convento, S., Rahman, M. S., and Yau, J. M. (2018). Selective attention gates the interactive crossmodal coupling between perceptual systems. *Curr. Biol.* 28, 746–752. doi: 10.1016/j.cub.2018.01.021

Conway, A. R., Cowan, N., and Bunting, M. F. (2001). The cocktail party phenomenon revisited: the importance of working memory capacity. *Psychon. Bull. Rev.* 8, 331–335. doi: 10.3758/BF03196169

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3:201. doi: 10.1038/nrn755

Dai, B., Chen, C., Long, Y., Zheng, L., Zhao, H., Bai, X., et al. (2018). Neural mechanisms for selectively tuning in to the target speaker in a naturalistic noisy situation. *Nat. Commun.* 9:2405. doi: 10.1038/s41467-018-04819-z

Dai, J., Li, Y., He, K., and Sun, J. (2016). "R-fcn: object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems* (Barcelona), 379–387.

Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. (2017). Human attention in visual question answering: do humans and deep networks look at the same regions? *Comput. Vis. Image Understand.* 163, 90–100. doi: 10.1016/j.cviu.2017.10.001

Dávila-Chacón, J., Liu, J., and Wermter, S. (2018). Enhanced robot speech recognition using biomimetic binaural sound source localization. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 138–150. doi: 10.1109/TNNLS.2018.2830119

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Diehl, M. M., and Romanski, L. M. (2014). Responses of prefrontal multisensory neurons to mismatching faces and vocalizations. *J. Neurosci.* 34, 11233–11243. doi: 10.1523/JNEUROSCI.5168-13.2014

Ding, N., and Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11854–11859. doi: 10.1073/pnas.1205381109

Dipoppa, M., Szwed, M., and Gutkin, B. S. (2016). Controlling working memory operations by selective gating: the roles of oscillations and synchrony. *Adv. Cogn. Psychol.* 12, 209–232. doi: 10.5709/acp-0199-x

Dorkenwald, S., Schubert, P. J., Killinger, M. F., Urban, G., Mikula, S., Svara, F., et al. (2017). Automated synaptic connectivity inference for volume electron microscopy. *Nat. Methods* 14, 435–442. doi: 10.1038/nmeth.4206

Du, Y., Kong, L., Wang, Q., Wu, X., and Li, L. (2011). Auditory frequency-following response: a neurophysiological measure for studying the "cocktail-party problem". *Neurosci. Biobehav. Rev.* 35, 2046–2057. doi: 10.1016/j.neubiorev.2011.05.008

Dwivedi, K., and Roig, G. (2019). "Representation similarity analysis for efficient task taxonomy & transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 12387–12396. doi: 10.1109/CVPR.2019.01267

Eckstein, M. P., Koehler, K., Welbourne, L. E., and Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Curr. Biol.* 27, 2827–2832. doi: 10.1016/j.cub.2017.07.068

Eimer, M., and Driver, J. (2001). Crossmodal links in endogenous and exogenous spatial attention: evidence from event-related brain potential studies. *Neurosci. Biobehav. Rev.* 25, 497–511. doi: 10.1016/S0149-7634(01)00029-X

Fan, J. (2014). An information theory account of cognitive control. *Front. Hum. Neurosci.* 8:680. doi: 10.3389/fnhum.2014.00680

Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I., and Posner, M. I. (2005). The activation of attentional networks. *Neuroimage* 26, 471–479. doi: 10.1016/j.neuroimage.2005.02.004

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., and Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *J. Cogn. Neurosci.* 14, 340–347. doi: 10.1162/089892902317361886

Fan, J., and Posner, M. (2004). Human attentional networks. *Psychiatr. Prax.* 31(Suppl. 2):210–214. doi: 10.1055/s-2004-828484

Fang, Y., Lin, W., Lau, C. T., and Lee, B.-S. (2011). "A visual attention model combining top-down and bottom-up mechanisms for salient object detection," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Prague: IEEE), 1293–1296.

Farah, M. J., Wong, A. B., Monheit, M. A., and Morrow, L. A. (1989). Parietal lobe mechanisms of spatial attention: modality-specific or supramodal? *Neuropsychologia* 27, 461–470. doi: 10.1016/0028-3932(89)90051-1

Feldman, H., and Friston, K. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 4:215. doi: 10.3389/fnhum.2010.00215

Folk, C. L., Remington, R. W., and Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 1030–1044. doi: 10.1037/0096-1523.18.4.1030

Frintrop, S., Rome, E., and Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: a survey. *ACM Trans. Appl. Percept.* 7:6. doi: 10.1145/1658349.1658355

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Fu, D., Barros, P., Parisi, G. I., Wu, H., Magg, S., Liu, X., et al. (2018). "Assessing the contribution of semantic congruency to multisensory integration and conflict resolution," in *IROS 2018 Workshop on Crossmodal Learning for Intelligent Robotics* (Madrid).

Gao, G., Lauri, M., Zhang, J., and Frintrop, S. (2017). "Saliency-guided adaptive seeding for supervoxel segmentation," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 4938–4943.

Gao, L., Guo, Z., Zhang, H., Xu, X., and Shen, H. T. (2017). Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimedia* 19, 2045–2055. doi: 10.1109/TMM.2017.2729019

Gaspelin, N., Leonard, C. J., and Luck, S. J. (2015). Direct evidence for active suppression of salient-but-irrelevant sensory inputs. *Psychol. Sci.* 26, 1740–1750. doi: 10.1177/0956797615597913

Gaspelin, N., Leonard, C. J., and Luck, S. J. (2017). Suppression of overt attentional capture by salient-but-irrelevant color singletons. *Attent. Percept. Psychophys.* 79, 45–62. doi: 10.3758/s13414-016-1209-1

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* 77, 980–991. doi: 10.1016/j.neuron.2012.12.037

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*, Vol. 1. New York, NY: Wiley.

Hafed, Z. M., Goffart, L., and Krauzlis, R. J. (2008). Superior colliculus inactivation causes stable offsets in eye position during tracking. *J. Neurosci.* 28, 8124–8137. doi: 10.1523/JNEUROSCI.1317-08.2008

Hafed, Z. M., and Krauzlis, R. J. (2008). Goal representations dominate superior colliculus activity during extrafoveal tracking. *J. Neurosci.* 28, 9426–9439. doi: 10.1523/JNEUROSCI.1313-08.2008

Hafez, M. B., Weber, C., Kerzel, M., and Wermter, S. (2019). Deep intrinsically motivated continuous actor-critic for efficient robotic visuomotor skill learning. *Paladyn J. Behav. Robot.* 10, 14–29. doi: 10.1515/pjbr-2019-0005

Häkkinen, S., Ovaska, N., and Rinne, T. (2015). Processing of pitch and location in human auditory cortex during visual and auditory tasks. *Front. Psychol.* 6:1678. doi: 10.3389/fpsyg.2015.01678

Häkkinen, S., and Rinne, T. (2018). Intrinsic, stimulus-driven and task-dependent connectivity in human auditory cortex. *Brain Struct. Funct.* 223, 2113–2127. doi: 10.1007/s00429-018-1612-6

Hanson, C., Caglar, L. R., and Hanson, S. J. (2018). Attentional bias in human category learning: the case of deep learning. *Front. Psychol.* 9:374. doi: 10.3389/fpsyg.2018.00374

Hara, K., Liu, M.-Y., Tuzel, O., and Farahmand, A.-M. (2017). Attentional network for visual object detection. *arXiv: 1702.01478*.

Henderson, J. M., and Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* 1, 743–747. doi: 10.1038/s41562-017-0208-0

Henderson, J. M., and Hollingworth, A. (1999). High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271. doi: 10.1146/annurev.psych.50.1.243

Hertz, U., and Amedi, A. (2014). Flexibility and stability in sensory processing revealed using visual-to-auditory sensory substitution. *Cereb. Cortex* 25, 2049–2064. doi: 10.1093/cercor/bhu010

Higgins, N. C., McLaughlin, S. A., Rinne, T., and Stecker, G. C. (2017). Evidence for cue-independent spatial representation in the human auditory cortex during active listening. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7602–E7611. doi: 10.1073/pnas.1707522114

Hinz, T., Heinrich, S., and Wermter, S. (2019). "Generating multiple objects at spatially distinct locations," in *International Conference on Learning Representations (ICLR)* (New Orleans, LA).

Hohman, F. M., Kahng, M., Pienta, R., and Chau, D. H. (2018). Visual analytics in deep learning: an interrogative survey for the next frontiers. *IEEE Trans. Visualizat. Comput. Graph.* 25, 2674–2693. doi: 10.1109/TVCG.2018.2843369

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295–1306.

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vis. Res.* 40, 1489–1506. doi: 10.1016/S0042-6989(99)00163-7

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* 11, 1254–1259. doi: 10.1109/34.730558

Jensen, O., and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front. Hum. Neurosci.* 4:186. doi: 10.3389/fnhum.2010.00186

Jetley, S., Murray, N., and Vig, E. (2016). "End-to-end saliency mapping via probability distribution prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 5753–5761. doi: 10.1109/CVPR.2016.620

Johnson, K. A., Robertson, I. H., Barry, E., Mulligan, A., Daibhis, A., Daly, M., et al. (2008). Impaired conflict resolution and alerting in children with ADHD: evidence from the attention network task (ANT). *J. Child Psychol. Psychiatry* 49, 1339–1347. doi: 10.1111/j.1469-7610.2008.01936.x

Kalinli, O., and Narayanan, S. S. (2007). "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Eighth Annual Conference of the International Speech Communication Association* (Antwerp).

Kaya, E. M., and Elhilali, M. (2012). "A temporal saliency map for modeling auditory attention," in *2012 46th Annual Conference on Information Sciences and Systems (CISS)* (New Jersey, NJ: IEEE), 1–6.

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Curr. Biol.* 15, 1943–1947. doi: 10.1016/j.cub.2005.09.040

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., and Kriegeskorte, N. (2017). Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* 76, 184–197. doi: 10.1016/j.jmp.2016.10.007

Klein, D. A., and Frintrop, S. (2011). "Center-surround divergence of feature statistics for salient object detection," in *2011 International Conference on Computer Vision* (Barcelona: IEEE), 2214–2219. doi: 10.1109/ICCV.2011.6126499

Klein, R. M. (2000). Inhibition of return. *Trends Cogn. Sci.* 4, 138–147. doi: 10.1016/S1364-6613(00)01452-2

Kondo, H. M., Pressnitzer, D., Toshima, I., and Kashino, M. (2012). Effects of self-motion on auditory scene analysis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6775–6780. doi: 10.1073/pnas.1112852109

Krauzlis, R. J., Lovejoy, L. P., and Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annu. Rev. Neurosci.* 36, 165–182. doi: 10.1146/annurev-neuro-062012-170249

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Kruthiventi, S. S., Ayush, K., and Babu, R. V. (2017). Deepfix: a fully convolutional neural network for predicting human eye fixations. *IEEE Trans. Image Process.* 26, 4446–4456. doi: 10.1109/TIP.2017.2710620

Kuchinsky, S. E., Vaden, K. I. Jr., Keren, N. I., Harris, K. C., Ahlstrom, J. B., Dubno, J. R., et al. (2012). Word intelligibility and age predict visual cortex activity during word listening. *Cereb. Cortex* 22, 1360–1371. doi: 10.1093/cercor/bhr211

Kulke, L. V., Atkinson, J., and Braddick, O. (2016). Neural differences between covert and overt attention studied using EEG with simultaneous remote eye tracking. *Front. Hum. Neurosci.* 10:592. doi: 10.3389/fnhum.2016.00592

Kummerer, M., Wallis, T. S., Gatys, L. A., and Bethge, M. (2017). "Understanding low-and high-level contributions to fixation prediction," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 4789–4798. doi: 10.1109/ICCV.2017.513

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE* 103, 1449–1477. doi: 10.1109/JPROC.2015.2460697

Lee, A. K., Larson, E., Maddox, R. K., and Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hear. Res.* 307, 111–120. doi: 10.1016/j.heares.2013.06.010

Lee, K., and Choo, H. (2013). A critical review of selective attention: an interdisciplinary perspective. *Artif. Intell. Rev.* 40, 27–50. doi: 10.1007/s10462-011-9278-y

Lewald, J., and Getzmann, S. (2015). Electrophysiological correlates of cocktail-party listening. *Behav. Brain Res.* 292, 157–166. doi: 10.1016/j.bbr.2015.06.025

Li, G., Gan, Y., Wu, H., Xiao, N., and Lin, L. (2019). Cross-modal attentional context learning for rgb-d object detection. *IEEE Trans. Image Process.* 28, 1591–1601. doi: 10.1109/TIP.2018.2878956

Li, W., Yuan, Z., Fang, X., and Wang, C. (2018). "Knowing where to look? Analysis on attention of visual question answering system," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 1–8.

Li, Z. (1999). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. Natl. Acad. Sci. U.S.A.* 96, 10530–10535. doi: 10.1073/pnas.96.18.10530

Li, Z. (2002). A saliency map in primary visual cortex. *Trends Cogn. Sci.* 6, 9–16. doi: 10.1016/S1364-6613(00)01817-9

Lidestam, B., Holgersson, J., and Moradi, S. (2014). Comparison of informational vs. energetic masking effects on speechreading performance. *Front. Psychol.* 5:639. doi: 10.3389/fpsyg.2014.00639

Liu, X., and Milanova, M. (2018). Visual attention in deep learning: a review. *Int. Robot. Automat. J.* 4, 154–155. doi: 10.15406/iratj.2018.04.00113

Lowe, D. G. (1999). "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, Vol. 99 (Corfu), 1150–1157. doi: 10.1109/ICCV.1999.790410

Lu, K., Xu, Y., Yin, P., Oxenham, A. J., Fritz, J. B., and Shamma, S. A. (2017). Temporal coherence structure rapidly shapes neuronal interactions. *Nat. Commun.* 8:13900. doi: 10.1038/ncomms13900

Luong, M.-T., Pham, H., and Manning, C. D. (2015). "Effective approaches to attention-based neural machine translation," in *Conference on Empirical Methods in Natural Language Processing* (Lisbon).

Ma, N., Gonzalez, J. A., and Brown, G. J. (2018). Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 2122–2131.

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends Cogn. Sci.* 16, 511–518. doi: 10.1016/j.tics.2012.08.010

Mahdi, A., Qin, J., and Crosby, G. (2019). Deepfeat: a bottom-up and top-down saliency model based on deep features of convolutional neural nets. *IEEE Trans. Cogn. Dev. Syst.* doi: 10.1109/TCDS.2019.2894561. [Epub ahead of print].

Mai, G., Schoof, T., and Howell, P. (2019). Modulation of phase-locked neural responses to speech during different arousal states is age-dependent. *NeuroImage* 189, 734–744. doi: 10.1016/j.neuroimage.2019.01.049

Mcdonald, J. J., Teder-Sälejärvi, W. A., Russo, F. D., and Hillyard, S. A. (2003). Neural substrates of perceptual enhancement by cross-modal spatial attention. *J. Cogn. Neurosci.* 15, 10–19. doi: 10.1162/089892903321107783

Melloni, L., van Leeuwen, S., Alink, A., and Müller, N. G. (2012). Interaction between bottom-up saliency and top-down control: how saliency maps are created in the human brain. *Cereb. Cortex* 22, 2943–2952. doi: 10.1093/cercor/bhr384

Mengotti, P., Boers, F., Dombert, P. L., Fink, G. R., and Vossel, S. (2018). Integrating modality-specific expectancies for the deployment of spatial attention. *Sci. Rep.* 8:1210. doi: 10.1038/s41598-018-19593-7

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667. doi: 10.1007/s00429-010-0262-0

Meredith, M. A. (2002). On the neuronal basis for multisensory convergence: a brief overview. *Cogn. Brain Res.* 14, 31–40. doi: 10.1016/S0926-6410(02)00059-9

Michie, P. T., Bearpark, H. M., Crawford, J. M., and Glue, L. C. (1990). The nature of selective attention effects on auditory event-related potentials. *Biol. Psychol.* 30, 219–250. doi: 10.1016/0301-0511(90)90141-I

Misselhorn, J., Friese, U., and Engel, A. K. (2019). Frontal and parietal alpha oscillations reflect attentional modulation of cross-modal matching. *Sci. Rep.* 9:5030. doi: 10.1038/s41598-019-41636-w

Morillon, B., and Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proc. Natl. Acad. Sci. U.S.A.* 114, E8913–E8921. doi: 10.1073/pnas.1705373114

Morissette, L., and Chartier, S. (2015). "Saliency model of auditory attention based on frequency, amplitude and spatial location," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)* (Budapest: IEEE), 1–5.

Mounts, J. R. (2000). Attentional capture by abrupt onsets and feature singletons produces inhibitory surrounds. *Percept. Psychophys.* 62, 1485–1493. doi: 10.3758/BF03212148

Mroueh, Y., Marcheret, E., and Goel, V. (2015). "Deep multimodal learning for audio-visual speech recognition," In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brisbane, QLD: IEEE), 2130–2134.

Musall, S., von Pföstl, V., Rauch, A., Logothetis, N. K., and Whittingstall, K. (2012). Effects of neural synchrony on surface EEG. *Cereb. Cortex* 24, 1045–1053. doi: 10.1093/cercor/bhs389

Oldoni, D., De Coensel, B., Boes, M., Rademaker, M., De Baets, B., Van Renterghem, T., et al. (2013). A computational model of auditory attention for use in soundscape research. *J. Acoust. Soc. Am.* 134, 852–861. doi: 10.1121/1.4807798

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends Cogn. Sci.* 2, 455–462. doi: 10.1016/S1364-6613(98)01241-8

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science* 314, 91–94. doi: 10.1126/science.1127242

Parisi, G. I., Barros, P., Fu, D., Magg, S., Wu, H., Liu, X., et al. (2018). "A neurorobotic experiment for crossmodal conflict resolution in complex environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 2330–2335.

Parisi, G. I., Barros, P., Kerzel, M., Wu, H., Yang, G., Li, Z., et al. (2017). "A computational model of crossmodal processing for conflict resolution," in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Manchester: IEEE), 33–38.

Peng, Y., Yan, K., Sandfort, V., Summers, R. M., and Lu, Z. (2019). A self-attention based deep learning method for lesion attribute detection from CT reports. *arXiv: 1904.13018*. doi: 10.1109/ICHI.2019.8904668

Perrett, S., and Noble, W. (1997). The contribution of head motion cues to localization of low-pass noise. *Percept. Psychophys.* 59, 1018–1026. doi: 10.3758/BF03205517

Pessoa, L., and Adolphs, R. (2010). Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat. Rev. Neurosci.* 11:773. doi: 10.1038/nrn2920

Picard, F., Sadaghiani, S., Leroy, C., Courvoisier, D. S., Maroy, R., and Bottlaender, M. (2013). High density of nicotinic receptors in the cingulo-insular network. *Neuroimage* 79, 42–51. doi: 10.1016/j.neuroimage.2013.04.074

Posner, M., and Snyder, C. (1975). "Attention and cognitive control," in *Information Processing and Cognition*, ed R. L. Solso (Hillsdale, MI: NJ Erlbaum), 55–85.

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25. doi: 10.1080/00335558008248231

Ramachandram, D., and Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Sig. Process. Mag.* 34, 96–108. doi: 10.1109/MSP.2017.2738401

Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 7263–7271.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* (Montreal, QC), 91–99.

Rock, I., and Palmer, S. (1990). The legacy of gestalt psychology. *Sci. Am.* 263, 84–91. doi: 10.1038/scientificamerican1290-84

Roseboom, W., Kawabe, T., and Nishida, S. (2013). The cross-modal double flash illusion depends on featural similarity between cross-modal inducers. *Sci. Rep.* 3:3437. doi: 10.1038/srep03437

Sadaghiani, S., and D'Esposito, M. (2014). Functional characterization of the cingulo-opercular network in the maintenance of tonic alertness. *Cereb. Cortex* 25, 2763–2773. doi: 10.1093/cercor/bhu072

Sawaki, R., and Luck, S. J. (2010). Capture versus suppression of attention by salient singletons: electrophysiological evidence for an automatic attend-to-me signal. *Attent. Percept. Psychophys.* 72, 1455–1470. doi: 10.3758/APP.72.6.1455

Schadwinkel, S., and Gutschalk, A. (2010). Activity associated with stream segregation in human auditory cortex is similar for spatial and pitch cues. *Cereb. Cortex* 20, 2863–2873. doi: 10.1093/cercor/bhq037

Scott, S. K., Rosen, S., Wickham, L., and Wise, R. J. (2004). A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *J. Acoust. Soc. Am.* 115, 813–821. doi: 10.1121/1.1639336

Senkowski, D., Schneider, T. R., Foxe, J. J., and Engel, A. K. (2008). Crossmodal binding through neural coherence: implications for multisensory processing. *Trends Neurosci.* 31, 401–409. doi: 10.1016/j.tins.2008.05.002

Shams, L., and Kim, R. (2010). Crossmodal influences on visual perception. *Phys. Life Rev.* 7, 269–284. doi: 10.1016/j.plrev.2010.04.006

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shi, J., Xu, J., Liu, G., and Xu, B. (2018). "Listen, think and listen again: capturing top-down auditory attention for speaker-independent speech separation," in *Proceedings of the International Joint Conference on Artificial Intelligence* (Rio), 4353–4360. doi: 10.24963/ijcai.2018/605

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182–186. doi: 10.1016/j.tics.2008.02.003

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)* (Vancouver, BC).

Skocaj, D., Leonardis, A., and Kruijff, G.-J. M. (2012). *Cross-Modal Learning*. Boston, MA: Springer.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends Cogn. Sci.* 7, 246–251. doi: 10.1016/S1364-6613(03)00109-8

Smith, D. T., Rorden, C., and Jackson, S. R. (2004). Exogenous orienting of attention depends upon the ability to execute eye movements. *Curr. Biol.* 14, 792–795. doi: 10.1016/j.cub.2004.04.035

Song, Y.-H., Kim, J.-H., Jeong, H.-W., Choi, I., Jeong, D., Kim, K., et al. (2017). A neural circuit for auditory dominance over visual perception. *Neuron* 93, 940–954. doi: 10.1016/j.neuron.2017.01.006

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9:255. doi: 10.1038/nrn2331

Stein, B. E., Wallace, M. W., Stanford, T. R., and Jiang, W. (2002). Book review: cortex governs multisensory integration in the midbrain. *Neuroscientist* 8, 306–314. doi: 10.1177/107385840200800406

Strauß, A., Wöstmann, M., and Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Front. Hum. Neurosci.* 8:350. doi: 10.3389/fnhum.2014.00350

Styles, E. (2006). *The Psychology of Attention*. New York, NY: Psychology Press.

Swets, J. A. (2014). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. New York, NY: Psychology Press.

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Theeuwes, J. (1991). Exogenous and endogenous control of attention: the effect of visual onsets and offsets. *Percept. Psychophys.* 49, 83–90. doi: 10.3758/BF03211619

Thurlow, W. R., and Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". *Percept. Mot. Skills* 36(3 Suppl.):1171–1184. doi: 10.2466/pms.1973.36.3c.1171

Todd, J. T., and Van Gelder, P. (1979). Implications of a transient–sustained dichotomy for the measurement of human performance. *J. Exp. Psychol. Hum. Percept. Perform.* 5, 625–638. doi: 10.1037/0096-1523.5.4.625

Togo, F., Lange, G., Natelson, B. H., and Quigley, K. S. (2015). Attention network test: assessment of cognitive function in chronic fatigue syndrome. *J. Neuropsychol.* 9, 1–9. doi: 10.1111/jnp.12030

Treisman, A., and Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychol. Rev.* 95, 15–48. doi: 10.1037/0033-295X.95.1.15

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* 16:55. doi: 10.1038/nrn3857

Uddin, L. Q., and Menon, V. (2009). The anterior insula in autism: under-connected and under-examined. *Neurosci. Biobehav. Rev.* 33, 1198–1203. doi: 10.1016/j.neubiorev.2009.06.002

Urbanek, C., Weinges-Evers, N., Bellmann-Strobl, J., Bock, M., Dörr, J., Hahn, E., et al. (2010). Attention network test reveals alerting network dysfunction in multiple sclerosis. *Multiple Scler. J.* 16, 93–99. doi: 10.1177/1352458509350308

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol.-Hum. Percept. Perform.* 34:1053. doi: 10.1037/0096-1523.34.5.1053

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. (2009). Poke and pop: tactile-visual synchrony increases visual saliency. *Neurosci. Lett.* 450, 60–64. doi: 10.1016/j.neulet.2008.11.002

VanRullen, R. (2003). Visual saliency and spike timing in the ventral visual pathway. *J. Physiol. Paris* 97, 365–377. doi: 10.1016/j.jphysparis.2003.09.010

Varela, F., Lachaux, J.-P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2:229. doi: 10.1038/35067550

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Veale, R., Hafed, Z. M., and Yoshida, M. (2017). How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling. *Philos. Trans. R. Soc. B Biol. Sci.* 372:20160113. doi: 10.1098/rstb.2016.0113

Veen, V. V., and Carter, C. S. (2006). Conflict and cognitive control in the brain. *Curr. Dir. Psychol. Sci.* 15, 237–240. doi: 10.1111/j.1467-8721.2006.00443.x

Verghese, P., and Pelli, D. G. (1992). The information capacity of visual attention. *Vis. Res.* 32, 983–995. doi: 10.1016/0042-6989(92)90040-P

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends Cogn. Sci.* 9, 585–594. doi: 10.1016/j.tics.2005.10.011

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *J. Neurophysiol.* 80, 1006–1010. doi: 10.1152/jn.1998.80.2.1006

Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *J. Exp. Psychol.* 27:339. doi: 10.1037/h0054629

Wang, B., Yang, Y., Xu, X., Hanjalic, A., and Shen, H. T. (2017). "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, CA: ACM), 154–162.

Wang, D., and Chang, P. (2008). An oscillatory correlation model of auditory streaming. *Cogn. Neurodyn.* 2, 7–19. doi: 10.1007/s11571-007-9035-8

Wang, D., and Terman, D. (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Trans. Neural Netw.* 6, 283–286. doi: 10.1109/72.363423

Wang, H., and Fan, J. (2007). Human attentional networks: a connectionist model. *J. Cogn. Neurosci.* 19, 1678–1689. doi: 10.1162/jocn.2007.19.10.1678

Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiol. Rev.* 90, 1195–1268. doi: 10.1152/physrev.00035.2008

Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, TX), 606–615. doi: 10.18653/v1/D16-1058

Warren, D. H., Welch, R. B., and McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. *Percept. Psychophys.* 30, 557–564. doi: 10.3758/BF03202010

Weissman, D. H., Gopalakrishnan, A., Hazlett, C., and Woldorff, M. (2004). Dorsal anterior cingulate cortex resolves conflict from distracting stimuli by boosting attention toward relevant events. *Cereb. Cortex* 15, 229–237. doi: 10.1093/cercor/bhh125

Welch, R. B., and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* 88, 638–667. doi: 10.1037/0033-2909.88.3.638

White, B. J., Berg, D. J., Kan, J. Y., Marino, R. A., Itti, L., and Munoz, D. P. (2017). Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nat. Commun.* 8:14263. doi: 10.1038/ncomms14263

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., Pantev, C., Sobel, D., et al. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proc. Natl. Acad. Sci. U.S.A.* 90, 8722–8726. doi: 10.1073/pnas.90.18.8722

Wolfe, J. M., and Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nat. Hum. Behav.* 1:0058. doi: 10.1038/s41562-017-0058

Wöstmann, M., Herrmann, B., Maess, B., and Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3873–3878. doi: 10.1073/pnas.1523357113

Wrigley, S. N., and Brown, G. J. (2004). A computational model of auditory selective attention. *IEEE Trans. Neural Netw.* 15, 1151–1163. doi: 10.1109/TNN.2004.832710

Wu, Q., and Miao, C. (2013). Curiosity: from psychology to computation. *ACM Comput. Surv.* 46, 1–26. doi: 10.1145/2543581.2543585

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., et al. (2015). "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of International Conference on Machine Learning* (Lille), 2048–2057.

Yang, G., Nan, W., Zheng, Y., Wu, H., Li, Q., and Liu, X. (2017). Distinct cognitive control mechanisms as revealed by modality-specific conflict adaptation effects. *J. Exp. Psychol. Hum. Percept. Perform.* 43:807. doi: 10.1037/xhp0000351

Yantis, S., and Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 10:601. doi: 10.1037/0096-1523.10.5.601

Yao, X., Han, J., Zhang, D., and Nie, F. (2017). Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. *IEEE Trans. Image Process.* 26, 3196–3209. doi: 10.1109/TIP.2017.2694222

Yin, P., Fritz, J. B., and Shamma, S. A. (2014). Rapid spectrotemporal plasticity in primary auditory cortex during behavior. *J. Neurosci.* 34, 4396–4408. doi: 10.1523/JNEUROSCI.2799-13.2014

Zhang, X., and Gong, Q. (2019). Frequency-following responses to complex tones at different frequencies reflect different source configurations. *Front. Neurosci.* 13:130. doi: 10.3389/fnins.2019.00130

Zouridakis, G., Simos, P. G., and Papanicolaou, A. C. (1998). Multiple bilaterally asymmetric cortical sources account for the auditory N1m component. *Brain Topogr.* 10, 183–189. doi: 10.1023/A:1022246825461

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.