Prédiction du parcours visuel en utilisant un apprentissage adversairial inter-observateurs cohérent

Mohamed Amine KERKOURI[1] Marouane TLIBA[1] Aladine CHETOUANI[1] Alessandro BRUNO[2]

[1]Laboratoire PRISME, 12 rue de Blois, 45100 Orleans, France

[2]IULM University, Milan, Italie

Résumé – Le trajet oculaire visuel est un concept essentiel pour l'attention visuelle, et sa prédiction est une tâche cruciale. Dans cet article, nous proposons une méthode d'apprentissage adversairial inter-observateurs pour prédire les parcours visuels à l'aide d'un réseau de neurones générateur léger et d'un réseau discriminatif utilisé comme fonction de pertes dynamiques. Notre méthode maintient la cohérence entre les distributions liées à la nature subjective des parcours visuels. Nous montrons la compétitivité de notre approche par rapport aux méthodes de l'état-de-l'art lors d'une phase de test.

Abstract – Visual scanpath is an essential concept for visual attention, and its prediction is a crucial task. We propose an inter-observer adversairial learning method to predict scanpaths using a lightweight generative neural network and a discriminative network used as a dynamic loss. Our method maintains coherence between distributions related to the subjective nature of scanpaths. We demonstrate the competitiveness of our approach compared to state-of-the-art methods in a testing phase.

1 Introduction

La rétine humaine reçoit environ 10^{10} bits/sec d'informations visuelles. La plupart de ces informations représentent des récepteurs haute définition situés dans la fovéa, qui couvre environ $1°$ du champ visuel. Cette quantité énorme d'informations est ensuite réduite à 3×10^6 bits/sec avant de passer par le nerf optique, puis réduite davantage lorsqu'elle traverse le cortex visuel [5]. Le mécanisme appelé "Attention visuelle" est régi par les contraintes anatomiques mentionnées précédemment ainsi que par d'autres contraintes neurologiques et psychologiques. L'observateur est incité à ne prêter attention qu'à certaines régions spécifiques de la scène. Ce phénomène se manifeste par des mouvements oculaires saccadiques, représentant le déplacement du regard d'une région à une autre pour un stimulus visuel. Lorsque les mouvements oculaires se concentrent sur une zone, le regard se fixe sur des points spécifiques, appelés "points de fixation". Ces derniers peuvent être collectés avec des oculomètres (i.e. eye-tracker) permettant la projection des points de fixation de plusieurs observateurs sur une carte binaire, mieux connue sous le nom de "carte de fixation". En outre, une "carte de saillance" est généralement obtenue avec des filtres de lissage pour donner une distribution spatiale en forme de "blob". Chaque valeur de pixel représentant la probabilité que le pixel attire l'attention des spectateurs. Le mécanisme décrit ci-dessus confère une efficacité remarquable au système visuel humain.

L'intérêt de la communauté scientifique pour la prédiction des trajets oculaires a récemment augmenté. Par exemple, le principe "winner-take-all" (WTA) a été utilisé par Itti et al. [4] dans leur première étude, où le trajet oculaire est extrait des régions les plus saillantes. Dans [12], les auteurs ont généré des trajets oculaires à partir d'une carte de saillance en utilisant des caractéristiques statistiques dérivées de plusieurs ensembles de données. Dans [1], des couches LSTM et le modèle VGG ont été utilisés avec un apprentissage adversairial. Dans [15], la carte de saillance a été modélisée comme un champ de gravité où la masse du regard se déplace en utilisant les lois physiques. Une carte de saillance fovéale a été utilisée conjointement avec des cartes d'inhibition de retour pour prédire les trajets oculaires dans [2]. Les auteurs de [9] ont présenté un modèle de bout en bout pour prédire simultanément le trajet oculaire et la carte de saillance d'une image [9], plus tard généralisé pour les images à $360°$ [10].

Les réseaux de neurones pour prédire les trajets oculaires doivent modéliser la distribution inter-observateurs et émuler les propriétés qualitatives des données réelles, tout en maintenant la cohérence entre les subjectivités de plusieurs observateurs.

À travers ces travaux précédents, nous avons constaté que cette tâche présente encore des défis fondamentaux et intéressants. La nature stochastique des trajets oculaires est fonction de la subjectivité des observateurs, et modéliser cette distribution inter-observateurs de manière cohérente s'avère être une tâche non évidente. En même temps, la modélisation de plusieurs observateurs rend difficile pour les réseaux de neurones de générer des résultats qui émulent les propriétés qualitatives des données réelles. Ainsi, la principale préoccupation se manifeste dans la façon de former un réseau de neurones pour prédire les trajets oculaires tout en maintenant la cohérence entre les subjectivités de plusieurs observateurs.

La méthode proposée apporte les contributions suivantes pour résoudre les défis mentionnés précédemment :

- Nous utilisons une approche d'apprentissage adversairial avec un jeu min-max. Cette méthode dynamique permet de mieux mettre en évidence la nature complexe des trajets oculaires.
- Nous conditionnons l'apprentissage sur la distribution probabiliste de tous les utilisateurs, obligeant le réseau à extraire les propriétés subjectives de la population d'observateurs.
- Nous prouvons la validité et la compétitivité de la méthode proposée en testant notre modèle sur deux grands ensembles de données.

Dans le reste de cet article, la section 2 décrit en détail la méthode proposée ainsi que les détails de l'apprentissage. Dans la section 3, nous présentons le protocole expérimental ainsi que les résultats quantitatifs et qualitatifs obtenus. La section 4 conclut l'article avec les conclusions.

2 Méthode Proposée

Pour résoudre les défis mentionnés dans la section 1, nous avons conçu une architecture d'apprentissage adversairial avec un mo-

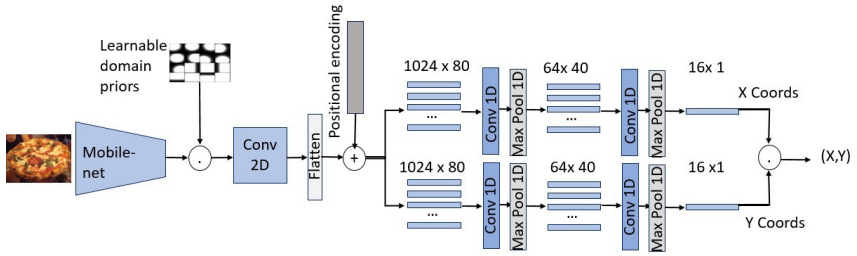

FIGURE 1 : Architecture du modèle Generateur.

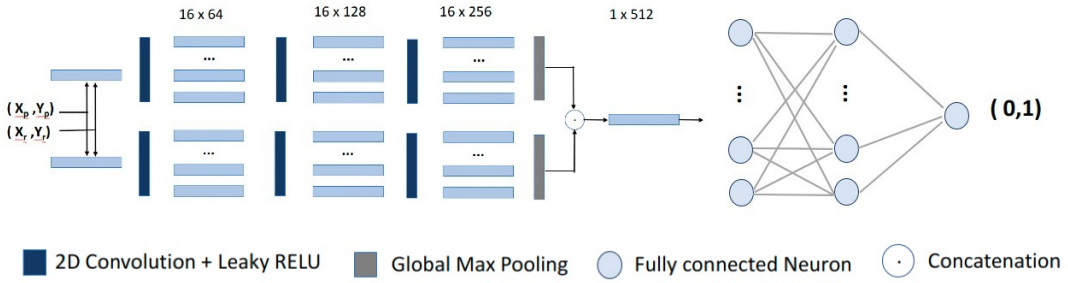

FIGURE 2 : Architecture du modèle Discriminateur.

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
PathGan[1]	0.9608	0.5698	0.9530	0.8172	0.8252	-0.2904	0.0825
Le Meur[12]	0.9505	0.6231	0.9488	0.8605	0.8457	0.8780	0.4784
G-Eymol[15]	0.9338	0.6271	0.9521	0.8967	0,8524	0.8727	0.3449
SALYPATH [9]	0.9659	**0.6275**	0.9521	0.8965	0,8605	0.3472	0.4572
our model (adversairial)	**0.9745**	0.6246	**0.9642**	**0.8892**	**0.8631**	**0.9762**	**0.5226**

TABLE 1 : Résultats de la prédiction de trajet oculaire sur Salicon.

dèle générateur entièrement convolutionnel et un modèle discriminateur utilisé comme une perte progressive dynamique. Ce dernier affine la capacité prédictive du générateur pendant l'apprentissage en améliorant sa propre capacité discriminative. Cette section présente les modèles proposés (c'est-à-dire le générateur et le discriminateur) ainsi que les stratégies d'apprentissage appliquées.

2.1 Architecture du générateur

Le modèle proposé utilise des composants légers pour prédire des trajets oculaires de longueurs variables. La figure 1 illustre l'architecture globale de notre modèle qui prend une image en entrée et génère un trajet oculaire. Pour encoder l'entrée dans un espace de représentation différent, nous utilisons un réseau MobileNet pré-entraîné comme extracteur de caractéristiques léger.

Pour améliorer la capacité représentative de notre modèle liée à notre tâche, nous introduisons l'utilisation de priors spécifiques au domaine à travers un ensemble apprenable de distributions gaussiennes spatiales, qui est une généralisation de la théorie de "biais central" pour l'attention visuelle [13]. Nous modélisons ces priors en utilisant les équations 1 et 2, où $\mu_{x,y}$, $\sigma_{x,y}$ et S représentent la moyenne de la distribution, l'écart-type et l'ensemble de priors gaussiens, respectivement.

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp^{-((\frac{x-\mu_x}{2\sigma_x})+(\frac{y-\mu_y}{2\sigma_y}))} \quad (1)$$

$$S = G_1(x,y), G_2(x,y), ..., G_{16}(x,y) \quad (2)$$

Dans cette étude, nous avons modélisé 16 priors gaussiens différents, chacun avec deux paramètres. Les informations contenues dans l'ensemble S sont ensuite intégrées aux caractéristiques résultant de MobileNet par concaténation suivi d'une convolution 2D. Comme nous pouvons considérer les trajets oculaires comme une séquence ordinale, nous avons ajouté une fonction d'encodage de position et utilisé une architecture basée sur la convolution 1D pour prédire la succession des fixations. Plus précisément, nous avons utilisé 2 branches de convolutions 1D afin de dissocier les représentations de la séquence multi-variable (c'est-à-dire les deux dimensions spatiales).

2.2 Architecture du discriminateur

Le réseau discriminateur utilisé est illustré par la figure 2. Son but est de discriminer entre les distributions des les trajets oculaires réels et générés. Pendant l'apprentissage, ce modèle améliore progressivement sa capacité à représenter la distribution des trajets oculaires réels, agissant ainsi comme une fonction de perte dynamique améliorant les performances du modèle générateur de trajet oculaire.

Inspirés par la performance du générateur, nous avons séparé les séquences représentant les coordonnées en deux dimensions spatiales différentes, permettant ainsi la désentrelacement des caractéristiques des deux dimensions. Chaque branche du modèle se compose d'une succession de convolutions 1D activées par une fonction de Leaky ReLU avec une pente de 0,2. Les caractéristiques extraites sont progressivement augmentées en proportion de la profondeur du réseau. À la fin de chaque branche, une couche de Max Pooling globale est utilisée sur chacun des vecteurs de caractéristiques. Les vecteurs résultants sont ensuite concaténés pour construire une représentation globale du trajet oculaire. Enfin, nous avons utilisé trois couches entièrement connectées pour discriminer les caractéristiques et ainsi classer les trajets oculaires.

2.3 Entraînement adversairial

Afin de modéliser avec une plus grande précision et maintenir la cohérence de la prédiction des trajets oculaires avec plusieurs utilisateurs, nous avons opté pour l'utilisation du jeu min-max entre le classificateur-discriminateur et le générateur prédictif de réseaux, représenté par l'équation suivante :

$$\min_G \max_D V(G,D) = \mathbb{E}\hat{y} \sim p\text{data}(y)[\log D(\hat{y}|p(y))] + \mathbb{E}x \sim p\text{x}(x)[1 - \log D(G(x|p(y))] \quad (3)$$

où x représente l'image d'entrée, $p(y)$ est la distribution de différents trajets oculaires aléatoires de vérité terrain de plusieurs utilisateurs changée périodiquement pendant l'entraînement, et $\hat{y}$ représente le trajet oculaire prédit. G et D sont les modèles de générateur et de discriminateur, respectivement.

Cette approche d'entraînement vise à réduire la distance entre le trajet oculaire prédit et l'ensemble des utilisateurs, permettant l'intégration des biais cognitifs de plusieurs spectateurs tout en maintenant de bonnes formes qualitatives pour le trajet oculaire.

Ainsi, cela oblige le réseau à apprendre une représentation non spécifique à l'utilisateur de la fonction perceptive. Le modèle a été entraîné pendant 246 époques avec un taux d'apprentissage égal à 10^{-5}

3 Évaluation

3.1 Jeux de données

Nous avons évalué notre méthode sur deux jeux de données largement utilisés, à savoir **Salicon** [6] et **MIT1003** [8]. **Salicon** [6] se compose de 9000 images pour l'apprentissage, 1000 images pour la validation et 5000 images pour les tests avec les cartes de saillance correspondantes et les données de trajet oculaire pour tous les utilisateurs. **MIT1003** [8] est généralement présenté avec la référence MIT300 dataset [7]. Il se compose de 1003 images de scènes naturelles avec les cartes de saillance correspondantes et les points de fixation recueillis lors de sessions de suivi oculaire. Chaque image dispose de 15 observations (trajets oculaires), ce qui implique une totalité de 15045 trajets oculaires pour toute la base de donnée.

3.2 Protocole expérimental

Dans notre travail, nous avons testé notre modèle sur les 5000 images du dataset Salicon avec environ 250 000 trajets oculaires, ce qui garantit la fiabilité empirique des résultats. Il convient de noter que dans cette étude notre modèle a été entraîné uniquement sur l'ensemble d'apprentissage de ce dataset. Nous avons ensuite utilisé l'ensemble de données MIT1003 complet pour tester notre modèle de manière croisée sans affiner notre modèle sur cet ensemble de données. De même que pour le premier dataset, le nombre important de trajets oculaires assure des résultats empiriquement fiables.

Pour évaluer les performances de notre méthode, nous avons utilisé trois mesures couramment utilisées : MultiMatch, NSS et Congruence. La métrique $MultiMatch$ (MM) [3] compare la similarité de deux vecteurs en utilisant cinq caractéristiques (forme, direction, longueur, position et durée). Étant donné que le modèle ne prédit que les coordonnées spatiales, nous n'utilisons que les quatre premières caractéristiques et mesurons les performances globales avec leur valeur moyenne. Deux métriques hybrides qui comparent les trajets oculaires prédits avec une carte de saillance générale pour une image donnée sont également utilisées : NSS et $Congruence$. NSS calcule la valeur moyenne de la saillance des emplacements de fixation du trajet oculaire sur la carte de saillance réelle, tandis que la $Congruence$ calcule le ratio des points de fixation prédits qui se trouvent dans les régions saillantes après avoir seuillé et binarisé la carte de saillance réelle. Les métriques hybrides permettent de mesurer l'accord et la cohérence entre le trajet oculaire prédit et les utilisateurs.

3.3 Résultats quantitatifs

Les tableaux 1 et 2 présentent les résultats obtenus après le test de notre modèle selon le protocole décrit dans la section 3.2 sur les ensembles de données Salicon et MIT1003, respectivement. Les performances de notre méthode sont comparées à celles des méthodes de l'état-de-lart.

Les résultats obtenus sur Salicon (voir le tableau 1) montrent que notre modèle a surpassé les méthodes de pointe pour les composantes forme, longueur et position de la métrique $multimatch$. Plus précisément, nous constatons une amélioration significative dans les composantes forme et longueur, tandis que la valeur moyenne globale de la métrique $multimatch$ montre une amélioration par rapport aux autres modèles. Nous avons également obtenu les meilleurs résultats pour la métrique $congruency$, ce qui indique que les fixations prédites sont principalement situées dans les régions saillantes, maintenant ainsi une certaine cohérence avec la distribution des utilisateurs sur l'ensemble de données Salicon. Ceci est d'autant plus souligné et soutenu par les résultats des méthodes de l'état-de-lart.

Comme nous avons testé le modèle sur l'ensemble de données MIT1003 [8] sans aucun type de fine-tuning, les résultats obtenus dans le tableau 2 montrent une diminution naturelle par rapport à ceux obtenus sur Salicon. Néanmoins, les résultats obtenus restent assez élevés et compétitifs avec les résultats de l'état-de-l'art. Cela montre la capacité de généralisation de notre approche aux distributions de données provenant de sources différentes. C'est particulièrement vrai, sachant que certains des modèles comparatifs ont été entraînés sur l'ensemble de données MIT1003 tels que DCSM [2] et Le Meur [12]. Les résultats montrent une grande amélioration dans les composantes forme et longueur de $multimatch$ par rapport aux autres modèles, tout en maintenant des résultats compétitifs pour la direction. Le résultat global est compétitif par rapport aux autres modèles. Les performances sur les métriques hybrides (c'est-à-dire NSS et $congruency$) maintiennent des marges proches de l'état de l'art, car notre modèle n'a été entraîné sur aucun sous-ensemble de MIT1003, qui a une distribution différente des observateurs. Nous pouvons également remarquer que les modèles Le Meur [12] et G-Eymol [15] ont été capables de maintenir une performance légèrement meilleure car ils reposent sur une étape de génération de carte de saillance avant l'échantillonnage du trajet oculaire.

3.4 Résultats qualitatifs

La figure 3 présente quelques résultats qualitatifs des trajets oculaires prédits (au centre) comparés aux trajets oculaires de référence (sur les deux côtés). Les prédictions de notre modèle montrent une grande fidélité aux trajets oculaires d'origine tout en maintenant une cohérence entre les trajets oculaires provenant

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
PathGan[1]	0.9237	0.5630	0.8929	0.8124	0.7561	-0.2750	0.0209
DCSM (VGG)[2]	0.8720	0.6420	0.8730	0.8160	0,8007	-	-
DCSM (ResNet)[2]	0.8780	0.5890	0.8580	**0.8220**	0,7868	-	-
Le Meur[12]	0.9241	0.6378	0.9171	0.7749	0,8135	0.8508	**0.1974**
G-Eymol[15]	0.8885	0.5954	0.8580	0.7800	0,7805	**0.8700**	0.1105
SALYPATH [9]	0.9363	0.6507	0.9046	0.7983	0,8225	0.1595	0.0916
our model	**0.9614**	**0.6529**	**0.9423**	0.7862	**0.8357**	0.7523	0.1797

TABLE 2 : Évaluation inter-dataset : Résultats de la prédiction de trajet oculaire sur MIT1003.

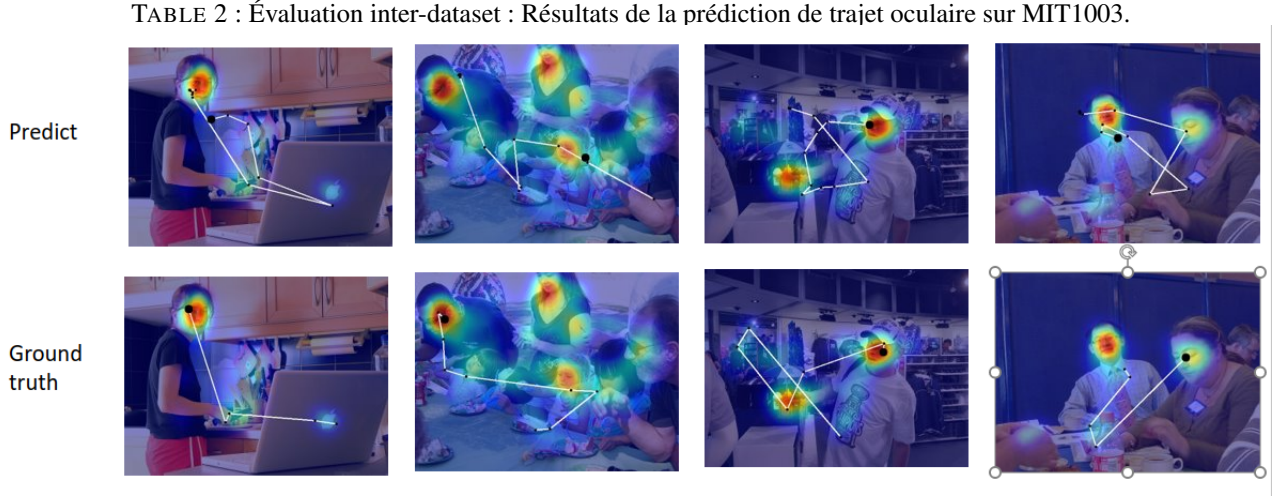

FIGURE 3 : Résultat qualitatif de scanpaths prédit.

de différents utilisateurs, rendant chaque trajet oculaire prédit hautement plausible.

4 Conclusion

Dans cet article, nous avons introduit une méthode d'apprentissage adversaire qui utilise un réseau discriminatif en tant que perte dynamique pour améliorer progressivement la capacité de représentation de notre modèle, tout en maintenant la cohérence inter-observateurs résultant de la nature subjective des trajets oculaires. Nous avons testé notre modèle sur les deux ensembles de données les plus utilisés pour la modélisation de l'attention visuelle et avons obtenu des résultats compétitifs sur plusieurs métriques hybrides et celles basées sur des vecteurs. Les résultats qualitatifs ont montré que notre méthode a réussi à prédire les trajets oculaires obtenus dans le monde réel. Cela confirme que le remplacement des fonctions de perte traditionnelles par des méthodes d'apprentissage adversaire donnerait de meilleurs résultats pour des tâches complexes de perception et d'attention.

Références

[1] Marc ASSENS, Xavier Giro-i NIETO, Kevin MCGUINNESS et Noel E O'CONNOR : Pathgan : visual scanpath prediction with generative adversarial networks. *In Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[2] Wentao BAO et Zhenzhong CHEN : Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*, 2020.

[3] Richard DEWHURST, Marcus NYSTRÖM, Halszka JARODZKA, Tom FOULSHAM, Roger JOHANSSON et Kenneth HOLMQVIST : It depends on how you look at it : Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012.

[4] Laurent ITTI et Christof KOCH : Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[5] Laurent ITTI, Geraint REES et John K TSOTSOS : *Neurobiology of attention*. Elsevier, 2005.

[6] Ming JIANG, Shengsheng HUANG, Juanyong DUAN et Qi ZHAO : Salicon : Saliency in context. *In CVPR*, pages 1072–1080. IEEE Computer Society, 2015.

[7] Tilke JUDD, Frédo DURAND et Antonio TORRALBA : A benchmark of computational models of saliency to predict human fixations. *In MIT Technical Report*, 2012.

[8] Tilke JUDD, Krista EHINGER, Frédo DURAND et Antonio TORRALBA : Learning to predict where humans look. *In IEEE International Conference on Computer Vision (ICCV)*, 2009.

[9] Mohamed A. KERKOURI, Marouane TLIBA, Aladine CHETOUANI et Rachid HARBA : Salypath : A deep-based architecture for visual attention prediction. *In 2021 IEEE International Conference on Image Processing (ICIP)*, pages 1464–1468, 2021.

[10] Mohamed Amine KERKOURI, Marouane TLIBA, Aladine CHETOUANI et Mohamed SAYEH : Salypath360 : Saliency and scanpath prediction framework for omnidirectional images, 2022.

[11] Olivier LE MEUR, Thierry BACCINO et Aline ROUMY : Prediction of the inter-observer visual congruency (iovc) and application to image ranking. *In Proceedings of the 19th ACM international conference on Multimedia*, pages 373–382, 2011.

[12] Olivier LE MEUR et Zhi LIU : Saccadic model of eye movements for free-viewing condition. *Vision research*, 116:152–164, 2015.

[13] Sabira K MANNAN, Keith H RUDDOCK et David S WOODING : The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial vision*, 1996.

[14] Robert J PETERS, Asha IYER, Laurent ITTI et Christof KOCH : Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005.

[15] Dario ZANCA, Stefano MELACCI et Marco GORI : Gravitational laws of focus of attention. *IEEE transactions on pattern analysis and machine intelligence*, 2019.