Few-shot Personalized Scanpath Prediction

Ruoyu Xu[1], Jingyi Xu[1], Sounak Mondal[1], Hieu Le[2],

Gregory Zelinsky[1], Minh Hoai[3], Dimitris Samaras[1]

[1]Stony Brook University, USA   [2]EPFL, Switzerland   [3]The University of Adelaide, Australia

Abstract

A personalized model for scanpath prediction provides insights into the visual preferences and attention patterns of individual subjects. However, existing methods for training scanpath prediction models are data-intensive and cannot be effectively personalized to new individuals with only a few available examples. In this paper, we propose few-shot personalized scanpath prediction task (FS-PSP) and a novel method to address it, which aims to predict scanpaths for an unseen subject using minimal support data of that subject's scanpath behavior. The key to our method's adaptability is the Subject-Embedding Network (SE-Net), specifically designed to capture unique, individualized representations for each subject's scanpaths. SE-Net generates subject embeddings that effectively distinguish between subjects while minimizing variability among scanpaths from the same individual. The personalized scanpath prediction model is then conditioned on these subject embeddings to produce accurate, personalized results. Experiments on multiple eye-tracking datasets demonstrate that our method excels in FS-PSP settings and does not require any fine-tuning steps at test time. Code is available at:
https://github.com/cvlab-stonybrook/few-shot-scanpath

1. Introduction

Recent models of scanpath prediction have excelled at predicting human attention [9, 10, 12, 30, 31, 51, 52, 54], which is important for applications such as autonomous driving [13, 32], virtual and augmented reality [22, 42], healthcare diagnostics [5, 40], and information visualization [45]. However, these methods are trained using attention data from multiple subjects and therefore learn population-level "average" attention patterns that will fail to reflect individual differences shaped by culture, memory, and experience [21]. To capture these individualized attention patterns, and to avoid biases that can result from group averaging, models of personalized scanpath prediction (PSP) attempt to learn individual subject embeddings that are used to predict an individual's scanpath [10, 20, 56]. PSP is particu-
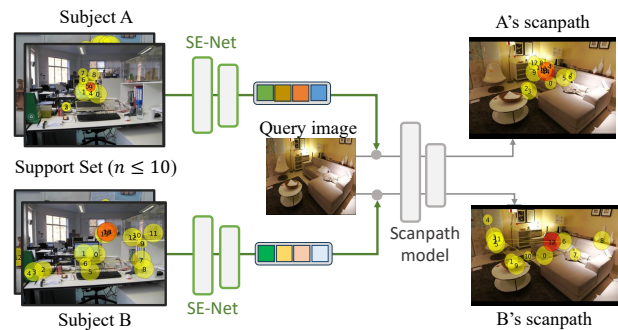

Figure 1. **Few-shot Personalized Scanpath Prediction (FS-PSP)**. Given a new subject with only a few support examples of their gaze behavior, can we adapt a base scanpath prediction model to this subject? We propose a subject-embedding extracting network, SE-Net, to achieve this personalized adaption.

larly useful for applications such as recommendation systems [7, 37, 39] and advertisements [35] because it allows subject personality to be decoded from an individual's attention.

A limitation of existing PSP models is that they require extensive data to accurately capture individual attention patterns. For example, to train models to predict the scanpaths made by people as they are searching for objects [51], it was necessary to collect the COCO-Search18 dataset of 269,760 search fixations [9], an effort requiring 10 subjects to each come to a laboratory and have their eye movements recorded for 10–12 hours. To be useful in practice, PSP models must be trainable on orders of magnitude less data. We therefore introduce the task of *few-shot personalized scanpath prediction* (FS-PSP) where an individual's attention must be predicted using fixations from only a few behavioral observations, defined here as the scanpaths of people viewing ≤ 10 images. We refer to the image-scanpath pairs collected for each subject as *support samples*.

FS-PSP is challenging because individuals' patterns of attention must be captured in just a few support samples. Existing PSP methods [10, 20, 56] often overfit to limited image content due to their reliance on large training datasets

and a lack of efficient adaptation mechanisms. Typically, the primary goal of these methods is to jointly learn a subject embedding along with scanpath patterns to improve performance on "*seen*" subjects, *i.e.*, ones with sufficient training scanpath data. However, subject embeddings are treated as a byproduct of scanpath prediction, limiting these models' ability to leverage knowledge from previously seen embeddings to adapt to new subjects. Consequently, overfitting on a few scenes and insufficient representation of novel individual visual patterns both cause PSP model performance to drop substantially in few-shot settings, as in the case of ISP [10] with ten support samples. EyeFormer[20] also requires a support set with at least 50 scanpaths to achieve stable personalized embeddings.

In this paper, we propose a flexible scanpath model that can adapt to new subject without requiring retraining or fine-tuning. This is achieved by decoupling subject embedding learning from the scanpath prediction process. First, we learn a subject embedding space that encodes personalized attention traits, then condition the scanpath prediction model on these embeddings. This separation enables robust performance in few-shot scenarios by avoiding the complexities of joint learning: the subject embedding extractor focuses exclusively on capturing the unique features of each subject's scanpath, while the scanpath prediction model only needs to learn a conditional mapping based on the corresponding subject embedding.

More specifically, we propose a Subject Embedding Network (SE-Net) to extract subject embeddings from image-scanpath pairs. During training, we use a base dataset containing a large number of images and scanpaths collected from the seen subjects. SE-Net is trained with a classification loss to distinguish between these subjects, focusing on extracting distinctive, personalized traits that support effective personalization. Additionally, we apply a contrastive loss to ensure embeddings from different scanpaths of the same subject remain similar while being distinct across different subjects. This training strategy emphasizes extracting robust, unbiased, and representative embeddings for each subject, akin to previous representation-learning-based few-shot approaches [19, 46, 48, 53]. We then train a personalized scanpath prediction network [10] on the base dataset using these learned subject embeddings, enhancing its capacity to infer scanpaths based on the given subject embeddings. At the inference stage, we extract embeddings for unseen subjects from a few image-scanpath exemplars (support set), average them to obtain a single subject embedding, and condition the scanpath prediction model on this embedding to enable generalization to new subjects.

To demonstrate the performance of our method on FS-PSP, we predict personalized scanpaths of unseen subjects under three n-shot settings ($n = 1, 5, 10$) and on three datasets: OSIE [47], COCO-Freeview [12], and COCO-

Search18 [51]. Our method outperforms the second-best model on the ScanMatch metric[14] on these datasets by 5.9%, 7.9% and 6.0%.

2. Related Works

Personalied Scanpath Prediction. Scanpath prediction aims to model both bottom-up and top-down human attention by predicting sequences of fixations as observers either freely explore a scene without specific instructions[2, 3, 23, 42, 43, 45], or engage in task-specific activities, such as visual question answering[9], webpage browsing, searching[11, 30, 51, 52, 54], and object referral[31]. Personalized scanpath prediction (PSP) aims to predict the scanpath of an individual subject rather than the group average. Although there are multiple works on personalized saliency prediction[6, 8, 24, 41, 50], PSP is a relatively new and underexplored task, as it requires building upon robust population-level scanpath prediction models. The key challenge in personalized scanpath prediction is enabling the model to recognize which subject's scanpath it is predicting. Current approaches [10, 20, 56] incorporate subject embeddings to represent different individuals. ISP [10] adds three modules to existing scanpath prediction models [9, 30] to encode subject embeddings at different stages within the model. EyeFormer[20] designs a new scanpath prediction model that leverages reinforcement learning and uses a viewer encoder to indicate different subjects. However, ISP suffers significant performance drops when working with very limited data ($N \leq 10$) for a new subject, while EyeFormer has only demonstrated its effectiveness with no fewer than 50 samples. This issue arises because over-parameterized models tend to overfit on small datasets, and must relearn subject embedding for the new subject, limiting their ability to fully utilize prior experience with previously seen subjects. To address these issues, we propose SE-Net, designed to mitigate overfitting and effectively leverage model experience.

Few-shot Learning Leveraging User Emebedding. Few-shot learning (FSL) has been widely studied in various topics, and a commonly used approach in FSL is to learn a function that maps the input to an embedding to effectively represent the prototype of different classes[16, 25, 38, 44, 46, 48, 49, 55]. User embedding, *i.e.* embeddings to represent person-specific patterns, encodes a user's unique behavioral patterns by mapping their actions through a network, capturing both intra-personal and inter-personal similarities and differences. This not only enables downstream tasks to effectively distinguish between individuals but, more importantly, allows for an understanding of unique user traits from the user embedding, enabling predictions that align with each person's distinctive patterns. Many topics address the few-shot problem by learning user embeddings, including recommendation systems [26] and gaze

estimation models [17, 34]. Although several works utilize scanpath for classification[33, 57], to the best of our knowledge, this is the first work that aims to extract user embeddings from scanpaths—a non-trivial task requiring the disentanglement of visual patterns driven by both bottom-up and top-down attention from fixation locations, temporal sequences, and durations.

3. Proposed FS-PSP Framework

This section describes our framework for FS-PSP, which consists of two major components: (1) a personalized scanpath predictor, called ISP-SENet, that predicts the scanpath for a subject, conditioned on the subject's embedding, and (2) a subject-embedding network, called SE-Net, that computes the embedding for a subject based on a small set of their gaze behaviors. In this section, we first provide a formal definition of the task and then describe the different components of our framework. The overview is shown in Fig. 2.

3.1. Problem Formulation

Let $\mathbf{d}$ represent a scanpath data instance; it is a tuple consisting of an image and a sequence of 2D gaze fixations. Let $S(\mathbf{d})$ denote the identity of the subject from whom the scanpath was collected. PSP considers a scenario where we have a base training set of gaze behavior with size m: $\mathcal{D}_{base} = \{\mathbf{d}_{base}^i\}_{i=1}^m$, which can be used to train a personalized scanpath predictor, but only for the subjects in the base training set. FS-PSP goes beyond PSP, considering the task of personalized scanpath prediction for a subject $\mathbf{s}$ who was not seen during training ($\mathbf{s} \neq S(\mathbf{d}_{base}^i)\ \forall i$) but for whom we have some gaze behavior data on a small subset of images from the base dataset, referred to as the support set. Let $\mathcal{D}_{supp} = \{\mathbf{d}_{supp}^i\}_{i=1}^n$ represent this support set, where $S(\mathbf{d}_{supp}^i) = \mathbf{s}\ \forall i$, and n is small (at most 10). The goal of FS-PSP is, given a set of unseen subjects, to predict their scanpaths on a set of query images $\mathbf{x}^i$ with size q: $\mathcal{D}_{query} = \{\mathbf{x}_{query}^i\}_{i=1}^q$.

3.2. ISP-SENet – Personalized Scanpath Predictor

ISP-SENet is one of the two core components of our framework. In this paper, we propose developing it based on Gazeformer-ISP [10]. This model accounts for the scanpath behaviors of individual subjects, with each subject in the training set associated with a separate embedding vector that is learnable but fixed after training. By assigning a unique embedding vector to each subject, the model achieves significant improvement over a generic model that does not account for subject identity. However, this approach only works for subjects seen during training, as the embedding vector can be retrieved from a lookup table. For an unseen subject, there is no way to compute the embedding vector, thereby preventing prediction for new subjects.

In this work, we propose replacing the fixed embedding vector with the output of a subject-embedding network, SE-Net. SE-Net can compute the subject embedding vector for any subject, as long as a support set of gaze behavior data from that subject is available. SE-Net represents the main technical innovation of our work, which we describe in the following section.

3.3. SE-Net – Subject Embedding Network

SE-Net computes a function f that takes a scanpath (both an image and a trajectory of eye fixations with durations) as input and outputs an embedding vector. This network calculates the subject embedding given a single scanpath behavior data point. For a support set contains more than one gaze behavior data point, $\mathcal{D}_{supp} = \{\mathbf{d}_{supp}^i\}_{i=1}^n$, we follow the approach of prototypical networks [38] and simply take the average as the overall subject's embedding, i.e., $\frac{1}{n}\sum_{i=1}^n f(\mathbf{d}_{supp}^i)$. We will next describe how this network is trained and then provide details about its architecture.

3.3.1. Training SE-Net

SE-Net is trained using scanpath data from the base training set $\mathcal{D}_{base}$, starting with the creation of triplets $(\mathbf{d}, \mathbf{d}_+, \mathbf{d}_-)$, where $\mathbf{d}$ is a scanpath drawn from $\mathcal{D}_{base}$, $\mathbf{d}_+$ is another scanpath randomly drawn from $\mathcal{D}_{base}$ but from the same subject as $\mathbf{d}$, and $\mathbf{d}_-$ is a scanpath randomly drawn from a different subject. The training loss for this triplet is the combination of three classification losses and a contrastive loss term:

$$\mathcal{L}_{cls}(\mathbf{d}) + \mathcal{L}_{cls}(\mathbf{d}_+) + \mathcal{L}_{cls}(\mathbf{d}_-) + \mathcal{L}_{\text{contrast}}. \quad (1)$$

The classification loss $\mathcal{L}_{\text{cls}}$ is computed separately for each of $\mathbf{d}$, $\mathbf{d}+$, and $\mathbf{d}_-$. It is defined as classifying the subject given their scanpath on an image by adding a classification head to the subject embedding layer. The contrastive loss $\mathcal{L}_{\text{contrast}}$ is based on triplet loss [36]:

$$\max\left(\|f(\mathbf{d}) - f(\mathbf{d}_+)\|^2 - \|f(\mathbf{d}) - f(\mathbf{d}_-)\|^2 + m, 0\right). \quad (2)$$

The margin m controls the distance between subject embeddings. A smaller margin allows scanpaths from different subjects to be similar, typically observed in top-down attention scenarios. Conversely, a larger margin is preferable in scenarios like free-viewing, characterized by significant diversity among subjects. Refer to the supplementary for more detailed explanations.

3.3.2. Network architecture

Feature extraction. Our semantic feature extractor F follows the design of HAT [54]. We encode images and scanpaths using hierarchical feature maps from an image encoder (ResNet [18]) and decoder (Deformable attention [58]) and obtain image tokens $F_I \in \mathbb{R}^{\left(\frac{H}{32} \cdot \frac{W}{32}\right) \times C}$ and
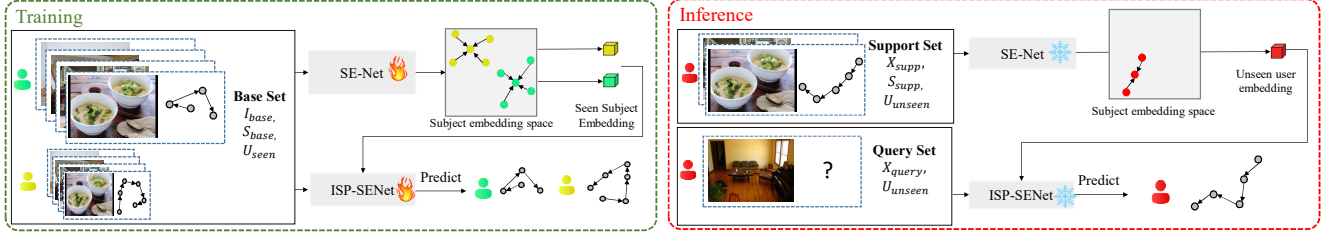
Figure 2. **Overview of ISP-SENet**: Our method for few-shot personalized scanpath prediction has two stages. In the training stage, we train two models on a large amount of image-scanpath pairs $\mathcal{D}_{base}$, corresponding to a set of seen subjects. Initially, we train the Subject Embedding Network (SE-Net) to obtain embeddings for seen subjects, followed by training ISP-SENet to predict scanpaths using these embeddings. In the inference phase, both models are frozen, and we extract embeddings for unseen subjects from the support set, $\mathcal{D}_{supp}$, which consists of n-shot images sampled from the base set. These unseen subject embeddings then guide ISP-SENet in predicting scanpaths for unseen subjects using the query set, $\mathcal{D}_{query}$, which includes a collection of unseen images.
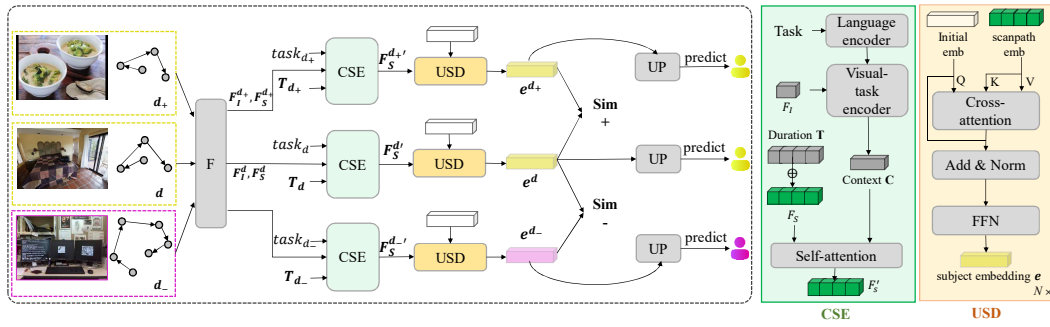

Figure 3. **Structure of SE-Net.** SE-Net employs a feature extractor F to derive image and scanpath semantic features, F_I and F_S, respectively. The CSE module then processes task and duration features, updating the scanpath embedding constrained by all extracted features. An initialized embedding learns human attention information from F'_S to produce the subject embedding e. This triplet network assesses the distances among e^{d+}, e^{d+}, e^{d-}, and the UP module predicts the subject ID. All CSE modules share the same weights, as do the USD and UP modules.

scanpath tokens $F_S \in \mathbb{R}^{L \times c}$ (H, W are the height and width of image, L is the max length of scanpaths, c is embedding dimension). The details are described in the supplementary material.

Context-Scanpath Encoder(CSE). Human attention can be divided into bottom-up and top-down processes, either freely viewing the image or viewing it with a specific task, such as searching for a target. Different tasks influence attention toward specific objects mentioned in the task and related objects. To make SE-Net task-aware, we design an multi-modal visual-task encoder for SE-Net, which takes task and image tokens as input, highlighting the image content relevant to the task. The task is first mapped to a task embedding $t \in \mathbb{R}^c$ by a language model (RoBERTa [28]), then the visual-task encoder is designed as Eq. (3):

$$C = \text{SelfAttn}(\{t \cdot W, F_I\}). \qquad (3)$$

Here, C represents the context embedding, capturing information from both the image and the task. W is a linear layer. $\{,\}$ denotes concatenation along the sequence dimen-

sion, and SelfAttn is multiple self-attention layers.

To obtain the final scanpath embedding, we first ensure the model is aware of each fixation's duration. Position information reflects where the subject tends to focus, while duration indicates the relative interest in specific objects. We use 2D sinusoidal position embeddings [27, 54] to encode the position information and 1D positional embeddings for durations. Given that durations range from 0 to 5000 ms, using 5000 different positional encodings would be redundant, as small differences (*e.g.*, 200 ms vs. 203 ms) are often negligible. Therefore, we collected all durations in the base set and uniformly grouped them into 10 bins, replacing the actual duration with its corresponding bin ID. This replacement is only used in SE-Net, while ISP-SENet still predicts the raw duration. Find details in the supplementary.

Combining the context embedding C, duration embeddings $\mathcal{T}$, and position embeddings Pos, we update the scanpath tokens F'_S representing viewing behavior refined by image content and task as shown in Eq. (4), w/o C denotes that tokens associated with C are discarded, Isolating image con-

tent from subject embeddings to reduce scene-specific bias.

$$F'_S = \text{SelfAttn}(\{C, F_S + \mathcal{T} + \text{Pos}\})_{\text{w/o } C} \qquad (4)$$

User-Scanpath Decoder (USD). The main focus of USD is to capture subject attention traits from the scanpath embedding. We initialize a subject token $e \in \mathbb{R}^c$ and use a cross-attention mechanism to extract the subject embedding from the scanpath embedding. Here, the subject token serves as the query, enabling it to attend to the scanpath embedding (used as key and value) and extract specific patterns that reflect the subject's unique attention characteristics. The detailed implementation is shown in Eq. (5):

$$e = \text{ReLU}(\text{Linear}(e + \text{CrossAttn}(e, F'_S))). \qquad (5)$$

To enable the updated subject embedding to distinguish between different subjects, we add a subject Predictor (UP), a series of linear layers with ReLU activation, to classify the subject ID:

$$\text{Subject ID} = \text{Linear}(\text{ReLU}(\text{Linear}(e))) \qquad (6)$$

4. Experiment

4.1. Dataset Setting

Dataset. We train and evaluate ISP-SENet on three datasets: OSIE [47], COCO-Freeview [12], and COCO-Search18 [51] (target-present). OSIE collects scanpaths under a free-viewing condition from 15 subjects on 700 images. COCO-Freeview collects scanpaths under a free-viewing condition from 10 subjects on 6202 images. COCO-Search18 has same images as COCO-FreeView, but the scanpaths are collected under a search task on 18 categories. We follow the same data split of OSIE as ISP [10], and the same split of COCO-Search18 and COCO-FreeView as HAT[54].

Unseen subject selection. For each dataset, we split the full subject set into around 70% seen and 30% unseen. We repeat experiments twice with different splits, and the result of another split is shown in the supplementary.

n-shot sampling. We selected three values, $n = 1, 5, 10$, as criteria for choosing the support set. For each value of n, we randomly sample n image-scanpath pairs from unseen subjects within the training set and assessed the model's performance across the entire test set of unseen subjects. This process was repeated 10 times with various support sets to ensure robustness, and the average performance metrics were reported. Find margin of error in the supplementary.

4.2. Experiment Setting

Baselines. We use ChenLSTM-ISP and Gazeformer-ISP[10] as baselines designed for personalized scanpath prediction. Both models are pretrained on a base set and

fully fine-tuned on a support set to jointly learn subject embeddings and scanpaths. Since new subject embeddings are not available initially, fine-tuning is required. To reduce overfitting, we also implement ChenLSTM-ISP-S and Gazeformer-ISP-S, which only fine-tune the subject embedding layer and freeze all other parameters.

Metrics. We evaluated ISP-SENet using value-based metrics as outlined by ISP [10], including ScanMatch (SM), MultiMatch (MM), and String-Edit Distance (SED), assess the similarity between predicted and actual scanpaths. ScanMatch [14] evaluates the overall similarity, Multi-Match [1, 15] analyzes five dimensions of scanpath similarity: shape, direction, length, position, and duration, while String-Edit Distance [4] identifies structural differences.

Implementation details. For SE-Net, the number of transformer layers in visual-task encoder and self-attention of CSE, and the number of USD layers are set to 3. The embedding dimension is set to 384. We use AdamW [29] with learning rate 0.0001. We train SE-Net for 25 epochs with a batch size of 16. The max length of scanpaths in OSIE and COCO-FreeView is set to 20, and the max length in COCO-Search18 is set to 10. For ISP-SENet, adhering to the GazeformerISP [10] configuration, we matched the maximum scanpath lengths and embedding dimensions with SE-Net, retaining other hyperparameters. The fine-tuning on support set takes 10 epochs for supervised training and 10 epochs for self-critical sequence training (SCST) [9].

4.3. Main Results

We compared ISP-SENet with four baselines across three datasets as shown in Tab. 1. The consistent performance across $n = 1, 5, 10$ indicates ISP-SENet's effectiveness in minimizing the impact of individual image instances, allowing embeddings to focus on attention patterns rather than being biased by the support set.

Results on OSIE. ISP-SENet outperforms the second-best approach by 5.9% and 2.5% in SM and SED, respectively. It maintains stable performance in the 5-shot and 10-shot settings and achieves a competitive SM score of 0.368 with just one image-scanpath pair.

Results on COCO-FreeView. Both ISP methods exhibit severe overfitting, particularly as COCO-FreeView is a larger dataset with more complex scenes and scanpaths. Nonetheless, ISP-SENet shows substantial enhancements in all settings, outperforming the baseline by 7.9% and 4.5% in SM and SED.

Results on COCO-Search18. We achieved 6.0% and 2.5% on SM and SED compared with the second-best. The consistent performance of COCO-Search18 highlights ISP-SENet's capability in capturing varied attention patterns across different search tasks. This success is at-

n-shot	Method	OSIE			COCO-FreeView			COCO-Search18		
		SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
$n = 1$	ChenLSTM-ISP	0.282	0.763	7.832	0.287	0.805	13.307	0.371	0.760	2.756
	Gazeformer-ISP	0.327	0.792	7.873	0.244	0.787	15.118	0.342	0.770	2.818
	ChenLSTM-ISP-S	0.328	0.793	7.601	0.339	0.814	12.523	0.448	0.803	2.394
	Gazeformer-ISP-S	0.354	0.801	7.503	0.333	0.817	12.538	0.446	0.802	2.463
	ISP-SENet	**0.368**	**0.805**	**7.413**	**0.369**	**0.832**	**12.227**	**0.475**	**0.814**	**2.333**
$n = 5$	ChenLSTM-ISP	0.319	0.773	7.855	0.320	0.815	12.950	0.386	0.773	2.489
	Gazeformer-ISP	0.340	0.791	7.920	0.286	0.800	14.630	0.353	0.774	2.980
	ChenLSTM-ISP-S	0.329	0.791	7.649	0.338	0.814	12.540	0.449	0.803	2.380
	Gazeformer-ISP-S	0.354	0.801	7.499	0.333	0.817	12.539	0.445	0.803	2.457
	ISP-SENet	**0.376**	**0.803**	**7.337**	**0.368**	**0.829**	**12.017**	**0.484**	**0.815**	**2.354**
$n = 10$	ChenLSTM-ISP	0.322	0.777	7.740	0.323	0.819	12.541	0.393	0.781	2.394
	Gazeformer-ISP	0.345	0.794	7.916	0.317	0.805	14.224	0.370	0.785	2.765
	ChenLSTM-ISP-S	0.328	0.791	7.637	0.340	0.814	12.532	0.449	0.803	2.379
	Gazeformer-ISP-S	0.354	0.802	7.505	0.333	0.816	12.545	0.446	0.802	2.464
	ISP-SENet	**0.375**	**0.803**	**7.318**	**0.367**	**0.828**	**11.956**	**0.482**	**0.815**	**2.359**

Table 1. Performance Comparison on Different Datasets under different few-shot settings for FS-PSP.

tributed to the visual-task encoder, detailed further in Sec. 5.

We observed that in the fine-tuning stage of Gazeformer-ISP-S and ChenLSTM-ISP-S, the unseen subject embeddings change minimally from seen subject embeddings. This suggests that simple learnable embeddings (instead of SE-Net) lack adaptability to unseen subjects. Their relatively better performance compared with fine-tuning all parameters likely stems from the similarity between seen and unseen subjects.

Adaptation time. A significant achievement of ISP-SENet is its rapid adaptation time to new subjects. Adaptation time refers to the duration required to update the scanpath prediction model, enabling it to infer the scanpaths of new subjects. For baseline methods, adaptation time is equivalent to the duration of fine-tuning on a support set, which is 161 and 267 seconds for ChenLSTM-ISP and Gazeformer-ISP respectively. In contrast, ISP-SENet's adaptation time is simply the time taken to obtain subject embeddings from SE-Net, which is 3.62 seconds. ISP-SENet achieves nearly real-time adaptation, enhancing efficiency and making the method more practical for real-world applications.

Dataset	ISP-SENet-Seen			ISP-SENet-Unseen		
	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
OSIE	0.390	0.812	7.309	0.375	0.803	7.318
COCO-FreeView	0.401	0.841	11.581	0.367	0.828	11.956
COCO-Search18	0.492	0.815	2.156	0.482	0.815	2.359

Table 2. Performance Comparison. Our method fully trained on all subjects (ISP-SENet-Seen), and trained on seen subjects and adapt to unseen subjects under 10-shot setting (ISP-SENet-unseen).

To establish the upper bound of prediction performance on unseen subjects (ISP-SENet-Unseen), we fully trained ISP-SENet using all available training data and evaluated it on the same subjects designated as unseen in the few-shot setting. The results are presented in Tab. 2. Remarkably, even without training on the unseen subjects, ISP-SENet's performance approaches the upper bound on OSIE and COCO-Search18. While ISP-SENet-Unseen underperforms compared to ISP-SENet-Seen on COCO-FreeView, it is trained on a considerably larger dataset of 43,143 scanpaths featuring more complex scenes and patterns. Even in more complex datasets, ISP-SENet can extract meaningful personalized embeddings from unseen subjects.

We employed scanpath accuracy, same as R@1 implemented in [10], to measure the accuracy of the predicted scanpath for a subject by determining if it is the most similar to the ground truth among all predictions made on the same image across different subjects, reflecting its ability to distinguish subjects. Distinguishing unseen from seen subjects is challenging due to the dataset's limited subject pool, potentially biasing the model towards known subjects. Nevertheless, our method showed considerable improvement in Tab. 3, illustrating that our unseen subject embeddings effectively discern among subjects despite the limited training subjects.

Qualitative results. In Fig. 4, we showcase ISP-SENet's distinct scanpath predictions for the same query image across various unseen subjects, highlighting its ability to capture individual attention patterns. GT 1 and GT 2 represent ground truth scanpaths for different subjects. The first row shows ISP-SENet discerning varying fixation orders across objects. The second row demonstrates how ISP-SENet captures the diverse attention distributions: Subject 1

	COCO-FreeView	COCO-Search18	OSIE
ChenLSTM-ISP	33.66	34.28	21.14
Gazeformer-ISP	31.99	31.73	21.36
ChenLSTM-ISP-S	31.95	33.17	20.17
Gazeformer-ISP-S	31.87	33.53	18.85
ISP-SENet	**35.57**	**35.25**	**22.85**

Table 3. Scanpath accuracy (higher is better) shows the model's ability to distinguish predicted scanpaths from different subjects.

conducts a thorough scene scan, whereas Subject 2 focuses more centrally. The third row reveals Subject 2's distraction by peripheral objects during the search. These observations confirm ISP-SENet's effectiveness in identifying unique attention traits among unseen subjects for personalized scanpath prediction. Find more results in supplementary.

5. Analysis

This section examines the impact of the number of seen subjects on performance, and the interpretability of the model.

num seen	Subject 1			Subject 2		
	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
3 (20%)	0.339	0.806	7.871	0.354	0.781	6.740
7 (50%)	0.365	0.812	7.457	0.364	0.790	6.522
10 (67%)	**0.374**	0.814	**7.420**	0.385	**0.794**	**6.460**
13 (93%)	0.370	**0.815**	7.504	**0.387**	0.793	6.482

Table 4. Performance Comparison of 10-shot setting on OSIE with different numbers of seen subjects in training stage. The table shows the performance for unseen subjects 1 and 2 with varying numbers of seen subjects.

Size of the base training set. Intuitively, training a model with a greater number of subjects enables it to learn a broader range of attention patterns. However, if trained on only a few subjects, the model might struggle with novel subjects whose scanpaths significantly deviate from those of seen subjects, thereby limiting its prediction accuracy. We assessed scanpath prediction performance for unseen subjects with varying numbers of training subjects, as detailed in Tab. 4. We consistently treated two subjects (Subject 1 and 2 in the table) as unseen, sampling the support set ($n = 10$) five times for each and averaging the results. Findings indicate that few-shot performance improves with an increasing number of seen subjects, although the difference between 10 and 13 subjects is minimal, likely due to performance saturation because of the potential similarity between some of OSIE dataset's subjects. See supplementary for COCO-Search18 results.

Interpretability – which fixations reflect attention trait? SE-Net also functions as a tool for quantitatively analyzing individual visual patterns. By examining the attention

weights from the last cross-attention layer of the USD module, we can identify which fixation tokens the subject embeddings consider most critical. We illustrate this analysis in Fig. 5, showcasing the two fixations with the highest attention weights for each subject. For example, in the first row, the highlighted fixations focus on a kid, keyboard, and cables, which are key to distinguishing the three subjects' attention distributions. The second row captures the initial focus of three individuals on a stop sign, warning tag, and one-way tag in a traffic setting, hinting at SE-Net's potential to further study varied driving behaviors across different subjects.

Role of visual-task encoder. COCO-Search18 features 18 search tasks, with participants typically focusing their final fixations on the search target in each image and task. The visual-task encoder informs the model about task-relevant image areas, providing prior knowledge of the search target for the CSE module to adjust its scanpath embedding. Although not tailored for object detection, the task embeddings consistently direct attention to target areas, as shown in the visualizations in Fig. 6. See supplementary for quantitative results. See quantitative results in supplementary.

loss	OSIE			COCO-Search18		
	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
cls	0.361	0.801	7.546	0.448	0.812	2.621
contrast	0.372	0.795	7.389	0.442	0.812	2.400
cls+contrast	0.375	0.803	7.318	0.482	0.815	2.359

Table 5. Ablation on the effect of different losses. cls is solely using classification loss, contrast is solely using contrastive loss, cls+contrast combines both loss, consistent with the settings used in all main experiments.

Ablation study on different losses. To evaluate the impact of different losses, we train SE-Net with classification loss only, contrastive loss only, and their combination (as in the main experiments) on OSIE and COCO-Search18. Results in Tab. 5 show that contrastive loss is crucial for capturing subject-specific attention patterns, while classification loss significantly accelerates convergence. On COCO-Search18, combining classification and triplet loss is more effective due to the similarity of attention patterns across subjects. Classification loss alone overlooks these similarities, while triplet loss struggles to find subtle differences.

Ablation study on different modules. To ablate the effect of different modules in SE-Net, we train SE-Net under two settings on COCO-Search18. 1) w/o task encoder: we remove task encoder, and direct pass task embedding extracted from text encoder to self-attention in CSE module. 2) w/o duration: do not use duration to update scanpath embedding. Results shown in Tab. 6 indicate the importance of each module in inferring unseen subject embeddings.
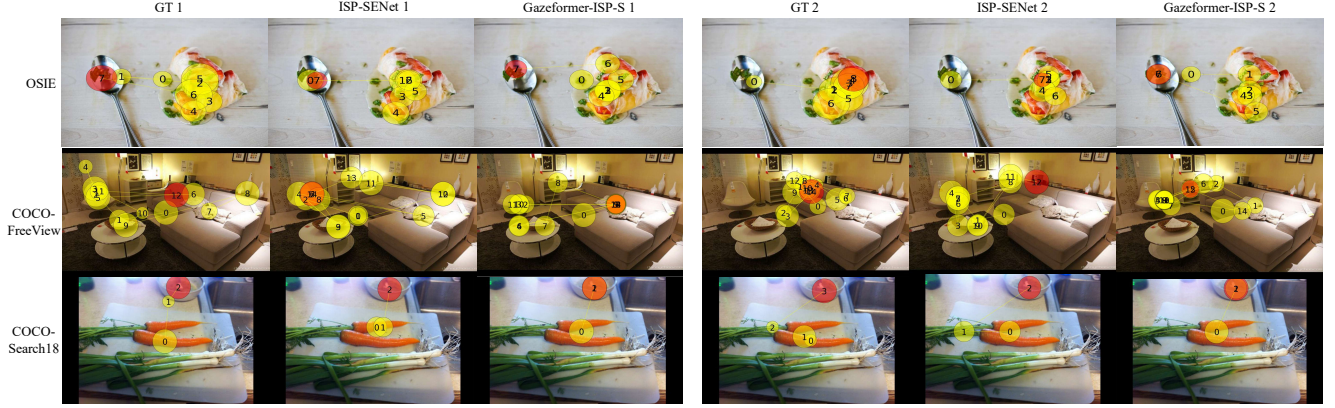
Figure 4. **Qualitative examples of scanpath prediction for different unseen subjects.** GT is the ground truth scanpaths of different unseen subjects. Red circle is the end fixation. In the third row, the subject is searching for "bowl". The results indicates our method is able to capture the temporal order of fixations, fixation distributions, and distractions, while baseline keeps predicting similar scanpaths of different subjects.
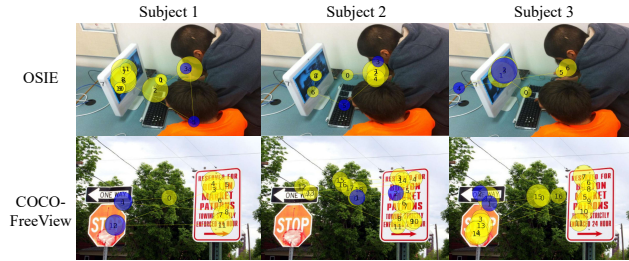

Figure 5. **Model interpretability.** By analyzing a large dataset of seen subject-scanpath pairs, SE-Net determines the most influential fixations in shaping unseen subject embeddings. Fixations highlighted in blue represent the two with the highest weights. This analysis demonstrates that ISP-SENet can effectively identify which fixations are crucial for distinguishing between subjects.
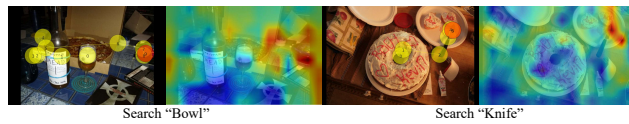

Figure 6. **Correspondence between task and image**. The visualization highlights the areas of the image that the visual-task encoder focuses on. The first and third columns display the attention weights extracted from the visual-task encoder, and the second and fourth columns show the ground truth scanpath visualizations from the support set. Fixations marked in red represent the last fixation, indicating the search target's location. This indicates the visual-task encoder's effectiveness in identifying task-relevant information within the image.

Module	SM ↑	MM ↑	SED ↓
w/o Task Encoder	0.459	0.814	2.459
w/o Duration	0.446	0.814	2.539
ISP-SENet	0.482	0.815	2.359

Table 6. Ablation on the effect of different modules of SE-Net on COCO-Search18.

ized Scanpath Prediction (FS-PSP) task. To tackle this challenge, we created a pipeline that independently learns subject embeddings to capture individual attention patterns and uses these embeddings to predict scanpaths. SE-Net is engineered to extract subject embeddings, separate individual attention traits from the image content, and facilitate generalization to unseen subjects with very few examples, thereby minimizing biases linked to restricted scenes. By utilizing these robust unseen subject embeddings, our scanpath prediction model markedly outperforms methods that require relearning personalization for new subjects with limited data. Furthermore, we have set a benchmark for FS-PSP, encouraging additional research to develop more comprehensive subject embeddings that are effective across diverse eye-tracking tasks and scenarios.

A limitation of SE-Net is, While it distinguishes subjects, it may overemphasize unique traits while overlooking shared patterns beneficial for scanpath prediction. Addressing this could enhance the robustness and generalizability of the method.

6. Conclusions and Discussion

We highlighted the significance of generating personalized scanpath predictions for novel subjects with minimal support samples and introduced the Few-Shot Personal-

References

[1] Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47:1377–1392, 2015.

5

[2] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[3] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2331–2338, 2017. 2

[4] Stephan A Brandt and Lawrence W Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience*, 9(1):27–38, 1997. 5

[5] Romuald Carette, Mahmoud Elbattah, Federica Cilia, Gilles Dequen, Jean-Luc Guerin, and Jérôme Bosche. Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths. In *HEALTHINF*, pages 103–112, 2019. 1

[6] Zhuoqing Chang, J Matias Di Martino, Qiang Qiu, Steven Espinosa, and Guillermo Sapiro. Salgaze: Personalizing gaze estimation using visual saliency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[7] Li Chen, Wanling Cai, Dongning Yan, and Shlomo Berkovsky. Eye-tracking-based personality prediction with recommendation interfaces. *User Modeling and User-Adapted Interaction*, 33(1):121–157, 2023. 1

[8] Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He. Learning from unique perspectives: User-aware saliency modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2023. 2

[9] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10876–10885, 2021. 1, 2, 5

[10] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024. 1, 2, 3, 5, 6

[11] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021. 2

[12] Yupei Chen, Zhibo Yang, Souradeep Chakraborty, Sounak Mondal, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2022. 1, 2, 5

[13] Jiyong Chung, Hyeokmin Lee, Hosang Moon, and Eunghyuk Lee. The static and dynamic analyses of drivers' gaze movement using vr driving simulator. *Applied Sciences*, 12(5):2362, 2022. 1

[14] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42:692–700, 2010. 2, 5

[15] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44:1079–1100, 2012. 5

[16] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 780–789, 2022. 2

[17] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam. On-device few-shot personalization for real-time gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[19] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Contrastive learning for prompt-based few-shot language learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5577–5587, Seattle, United States, 2022. Association for Computational Linguistics. 2

[20] Yue Jiang, Zixin Guo, Hamed Rezazadegan Tavakoli, Luis A Leiva, and Antti Oulasvirta. Eyeformer: predicting personalized scanpaths with transformer-guided reinforcement learning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15, 2024. 1, 2

[21] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012. 1

[22] Sebastian Kapp, Michael Barz, Sergey Mukhametov, Daniel Sonntag, and Jochen Kuhn. Arett: Augmented reality eye tracking toolkit for head mounted displays. *Sensors*, 21(6):2234, 2021. 1

[23] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022. 2

[24] Aoqi Li and Zhenzhong Chen. Personalized visual saliency: Individuality affects image perception. *IEEE Access*, 6:16099–16109, 2018. 2

[25] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8334–8343, 2021. 2

[26] Ruirui Li, Xian Wu, Xian Wu, and Wei Wang. Few-shot learning for new user recommendation in location-based social networks. In *Proceedings of The Web Conference 2020*, pages 2472–2478, 2020. 2

[27] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34:15816–15829, 2021. 4

[28] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019. 4

[29] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[30] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1441–1450, 2023. 1, 2

[31] Sounak Mondal, Seoyoung Ahn, Zhibo Yang, Niranjan Balasubramanian, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Look hear: Gaze prediction for speech-directed human attention. In *European Conference on Computer Vision*, pages 236–255. Springer, 2025. 1, 2

[32] Jordan Navarro, Otto Lappi, Fra**c**cois Osiurak, Emma Hernout, Catherine Gabaude, and Emanuelle Reynaud. Dynamic scan paths investigations under manual and highly automated driving. *Scientific reports*, 11(1):3776, 2021. 1

[33] Takumi Nishiyasu and Yoichi Sato. Gaze scanpath transformer: Predicting visual search target by spatiotemporal semantic modeling of gaze scanpath. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 625–635, 2024. 3

[34] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019. 3

[35] Jean Pfiffelmann, Nathalie Dens, and Sébastien Soulez. Personalized advertisements with integration of names and photographs: An eye-tracking experiment. *Journal of Business Research*, 111:196–207, 2020. 1

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[37] ABM Fahim Shahriar, Mahedee Zaman Moon, Hasan Mahmud, and Kamrul Hasan. Online product recommendation system by using eye gaze data. In *Proceedings of the International Conference on Computing Advancements*, pages 1–7, 2020. 1

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[39] Hyejin Song and Nammee Moon. Eye-tracking and social behavior preference-based recommendation system. *The Journal of Supercomputing*, 75:1990–2006, 2019. 1

[40] Yingjie Song, Zhi Liu, Gongyang Li, Jiawei Xie, Qiang Wu, Dan Zeng, Lihua Xu, Tianhong Zhang, and Jijun Wang. Ems: A large-scale eye movement dataset, benchmark, and new model for schizophrenia recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 1

[41] Florian Strohm, Mihai Bâce, and Andreas Bulling. Learning user embeddings from human gaze for personalised saliency prediction. *Proceedings of the ACM on Human-Computer Interaction*, 8(ETRA):1–16, 2024. 2

[42] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. Scandmm: A deep markov model of scanpath prediction for 360deg images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6989–6999, 2023. 1, 2

[43] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2101–2118, 2019. 2

[44] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 2

[45] Yao Wang, Andreas Bulling, et al. Scanpath prediction on information visualisations. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1, 2

[46] Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9003–9013, 2022. 2

[47] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):1–20, 2014. 2, 5

[48] Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8812–8821, 2021. 2

[49] Jingyi Xu, Hieu Le, and Dimitris Samaras. Generating features with increased crop-related diversity for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19713–19722, 2023. 2

[50] Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu. Personalized saliency and its prediction. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2975–2989, 2018. 2

[51] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2020. 1, 2, 5

[52] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *European Conference on Computer Vision*, pages 52–68. Springer, 2022. 1, 2

[53] Zhanyuan Yang, Jinghua Wang, and Ying J. Zhu. Few-shot classification with contrastive learning. In *European Conference on Computer Vision*, 2022. 2

[54] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using

transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693, 2024. 1, 2, 3, 4, 5

[55] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8808–8817, 2020. 2

[56] Feng Yuan, Matthieu Perreira Da Silva, and Alexandre Bruckert. Personalized visual scanpath prediction using ior-roi weighted attention network. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences Workshops*, pages 66–68, 2023. 1, 2

[57] Wenqi Zhong, Linzhi Yu, Chen Xia, Junwei Han, and Dingwen Zhang. Spformer: Spatio-temporal modeling for scanpaths with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7605–7613, 2024. 3

[58] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3