

Noise-Aware Video Saliency Prediction

Ekta Prashnani^{1,2}
ekta@ece.ucsb.edu

Orazio Gallo²
ogallo@nvidia.com

JooHwan Kim²
sckim@nvidia.com

Josef Spjut²
jspjut@nvidia.com

Pradeep Sen¹
psen@ece.ucsb.edu

Iuri Frosio²
ifrosio@nvidia.com

¹ University of California,
Santa Barbara,
California, USA

² NVIDIA Research,
Santa Clara,
California, USA

Abstract

We tackle the problem of predicting saliency maps for videos of dynamic scenes. We note that the accuracy of the maps reconstructed from the gaze data of a fixed number of observers varies with the frame, as it depends on the content of the scene. This issue is particularly pressing when a limited number of observers are available. In such cases, directly minimizing the discrepancy between the predicted and measured saliency maps, as traditional deep-learning methods do, results in overfitting to the noisy data. We propose a *noise-aware training* (NAT) paradigm that quantifies and accounts for the uncertainty arising from frame-specific gaze data inaccuracy. We show that NAT is especially advantageous when limited training data is available, with experiments across different models, loss functions, and datasets. We also introduce a video game-based saliency dataset, with rich temporal semantics, and multiple gaze attractors per frame. The dataset and source code are available at <https://github.com/NVlabs/NAT-saliency>.

1 Introduction

Humans can perceive high-frequency details only within a small solid angle, and thus, analyze scenes by directing their gaze to the relevant parts [15, 20]. Predicting a distribution of gaze locations (*i.e.*, a *saliency map*) for a visual stimulus has widespread applications such as image or video compression [8] and foveated rendering [63, 64], among others. This has inspired an active area of research – visual saliency prediction. Early methods focused on low- or mid-level visual features [25, 26, 65], and recent methods leverage high-level priors through deep learning (DL) for saliency prediction and related tasks such as salient object detection [20, 24, 28, 36, 37, 39, 43, 44, 45, 49, 60, 63, 62].

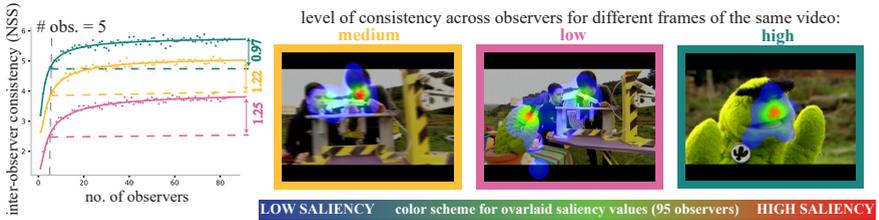


Figure 1: **Motivation for noise-aware training (NAT)**. Frames from a video (DIEM [48]) dataset) are shown with an overlay of the saliency maps reconstructed from the gaze data of 95 observers. The level of gaze consistency across observers varies with frame content, leading to different asymptotic values and convergence rates of the per-frame inter-observer consistency (IOC) curves. Consequently, the accuracy of the saliency maps reconstructed from gaze data varies across frames – especially when a limited number of observers (say, 5 observers) are available. This impedes traditional training that directly minimizes the discrepancy between predicted and measured maps. We introduce NAT to address this.

Given the improved accessibility of eye trackers [57], datasets for saliency prediction are captured by recording gaze locations of observers viewing an image or a video. These gaze locations are then used to estimate a per-frame/image saliency map. Generally speaking, the quality of the reconstructed saliency maps increases with the number of gaze samples. However, two factors make it particularly challenging to reconstruct high-quality maps for videos. First, since a single observer contributes only a few (typically one [48]) gaze locations per video frame, more observers are needed to capture sufficient per-frame gaze data for videos as compared to images (e.g., the CAT2000 dataset has on average ~ 333 fixations per image from 24 observers [10], while LEDOV has ~ 32 fixations per video frame from 32 observers [60]). Therefore, the cost associated with the creation of truly large-scale datasets with tens of thousands of videos can be prohibitively high. Second, for videos of dynamic scenes, it is hard to guarantee high accuracy of the reconstructed saliency maps across all frames from the gaze data of a fixed number of observers. This is because the gaze behavior consistency across observers depends on the scene content [61]: scenes that elicit a high consistency would require fewer observers to reconstruct accurate saliency maps than those for which the inter-observer consistency (IOC) in gaze behavior is low.

Fig. 1 shows 3 frames from a DIEM video [48] with 95-observer saliency map overlays and the per-frame IOC as a function of the number of observers used to reconstruct the saliency map [60, 58, 61]. A converged IOC curve indicates that additional observers do not add new information to the reconstructed saliency map and the captured number of observers (e.g., 95 in Fig. 1) are sufficient for accurate estimation of saliency [60, 62]. As is clear from these plots, when the number of available observers is small, the IOC curve differs from its asymptotic value by a varying amount for each frame. This leads to varying per-frame accuracy of the saliency map reconstructed from few observers. In such cases, traditional training methods, which minimize the discrepancy between the predicted and measured saliency, can lead to overfitting to the inaccurate saliency maps in the training dataset.

We address these issues by proposing a *Noise-Aware Training* (NAT) paradigm: we interpret the discrepancy d between the measured and predicted saliency maps as a random variable, and train the saliency predictor through likelihood maximization. We show that NAT avoids overfitting to incomplete or inaccurate saliency maps, weighs training frames based on their reliability, and yields consistent improvement over traditional training, for

different datasets, deep neural networks (DNN), and training discrepancies, *especially when few observers or frames are available for training*. Therefore, NAT ushers in the possibility of designing larger-scale video-saliency datasets with fewer observers per video, since it learns high-quality models with less training data.

Although existing datasets have been vital to advance video saliency research [30, 48], a significant portion of these datasets consists of almost-static content, as observed recently by Tangemann et al. [59]. Using these datasets for training and evaluation therefore makes it difficult to assess how saliency prediction methods fare on aspects specific to *videos*, such as predicting saliency on temporally-evolving content. Consequently, even an image-based saliency predictor can provide good results for existing video saliency datasets [55]. As a step towards designing datasets with dynamic content, we introduce the Fortnite Gaze Estimation Dataset (ForGED), that contains clips from game-play videos of Fortnite, a third-person-shooter game amassing hundreds of million of players worldwide. With ForGED, we contribute a novel dataset with unique characteristics such as: fast temporal dynamics, semantically-evolving content, multiple attractors of attention, and a new gaming context.

2 Related work

Saliency prediction methods. In recent years, DL-based approaches have remarkably advanced video saliency prediction [8]. Existing works include (i) 3D CNN architectures that observe a short sub-sequence of frames [6, 27, 47]; (ii) architectures that parse one frame at a time but maintain information about past frames in feature maps (*e.g.*, simple temporal accumulation or LSTMs [16, 22, 42, 59, 62]); or (iii) a combination of both [9]. Some methods also decouple spatial and temporal saliency through specific features, such as “object-ness” and motion in a frame [9, 29, 30], adopt vision transformers [45], or predict a compact spatial representations such as a GMM [61]. Overall, existing works largely focus on improving model architectures, output representations [61], and training procedures. In contrast, our NAT paradigm is broadly applicable across all these categories and it only modifies the loss function to account for the level of reliability of the measured saliency maps. We demonstrate the model-agnostic applicability of NAT through experiments on representative DNN architectures – we use ViNET [27] (2021 state-of-the-art that uses 3D CNN), TASED-Net [47] (a 3D CNN-based model), and SaEMA [42].

Metrics and measures of uncertainty for saliency. Popular metrics for training and evaluating saliency models include density-based functions (Kullback-Leibler divergence, KLD, correlation coefficient, CC, similarity, SIM [64]), and fixation-based functions (area under the ROC curve, AUC [11, 12], normalized scanpath saliency, NSS [12, 50, 62]). Fixation-based metrics evaluate saliency at the captured gaze locations, without reconstructing the entire map. We observed that when few locations on a small training set are available, models that directly optimize either type of function show suboptimal performance.

The adoption of correction terms on a incomplete probability distributions has been explored in population statistics [13, 23]. Adapting these concepts to gaze data is possible at low spatial resolutions [61]. However, at full resolution, gaze data tends to be too sparse to collect sufficient statistics in each pixel. IOC curves are also used to estimate the level of completeness of saliency maps [30], and the upper bounds on the performance of a saliency predictor [32, 38, 41, 61]. Such approaches provide an insight on level of accuracy and uncertainty in saliency maps, but depend on the availability of sufficient observers to estimate the full curves. In contrast, NAT is designed specifically for limited-data setting.

Video saliency datasets. Some datasets capture video saliency for specific content (like sports [53], movies [58], faces [46]), while others (like DHF1K [59], LEDOV [29], and DIEM [48]) do for everyday scenes [8]. We perform our experimental analysis using two of the largest datasets, DIEM and LEDOV, which also provide high-quality gaze annotations, and, more importantly, access to per-observer gaze data – a feature that is not available in the most popular DHF1K dataset, among other artifacts [53].

Videos with dynamic content are key to capturing and assessing *video*-specific saliency. However, existing datasets contain mostly-static content, which can be explained by image-based models [53]. Existing datasets with videos of highly-dynamic content are either constrained in visual content variety and real-time gaze capture (*e.g.*, Atari-Head dataset [55]), or capture gaze data from only a single subject (such as a game player [9], or a driver [2]), limiting the accuracy of test-time evaluations. We therefore turn to game-play videos of Fortnite, with its rich temporal dynamics, to further evaluate video-specific saliency. ForGED features videos from Fortnite with gaze data from up to 21 observers per video frame, enabling an effective benchmark for training and evaluating video-specific saliency.

3 Noise-Aware Training (NAT)

The accuracy of the saliency maps in videos varies with frame content, especially when limited gaze data is available. The inaccuracy in the saliency maps can stem from errors in gaze measurements, such as inaccurate localization of Purkinje reflections or calibration issues in gaze tracker [48] – we term these *measurement noise*. Using an insufficient number of observers to estimate the probabilities in different sub-regions of the saliency map is another source of noise, which we term *incomplete sampling*. While the measurement noise can be partially alleviated with techniques such as temporal filtering [14], the best way to overcome *both* sources of noise is to capture sufficient data. Since this can be impractical, we now discuss our proposed strategy to effectively train a DNN for saliency prediction, accounting for the noise level in a measured saliency map (Fig. 2).

Let x_i be the probability distribution of the *ground-truth* saliency map for the i^{th} frame, reconstructed from *sufficient* gaze data (*e.g.*, when the IOC curve is close to its asymptotic value). The traditional approach to train a saliency predictor (abbreviated as TT : traditional training) optimizes:

$$J^{\text{ideal}} = \sum_i d(\hat{x}_i, x_i), \quad (1)$$

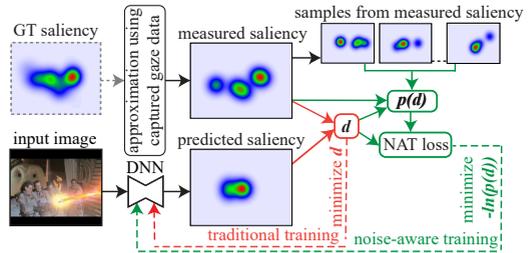


Figure 2: **Overview of NAT.** For an input image, a saliency map is approximated from measured gaze data. This can result in a noisy/incomplete version of the GT saliency – especially when limited gaze data is available. Instead of training a DNN by directly minimizing the discrepancy, d , between the measured and predicted saliency (traditional training), with NAT we first estimate a distribution for d , $p(d)$, that quantifies the uncertainty in d due to the inaccuracies in the measured saliency maps. We then train the DNN to optimize the likelihood of d .

where \hat{x}_i is the predicted saliency map, and $d(\cdot, \cdot)$ is a discrepancy measure such as KLD, CC, NSS, or a mix of these. Since reconstructing an accurate x_i is challenging, the existing methods instead end up optimizing:

$$J^{\text{real}} = \sum_i d(\hat{x}_i, \tilde{x}_i), \quad (2)$$

where \tilde{x}_i is an *approximation* of the unobservable x_i . We adopt the standard practice to estimate \tilde{x}_i from captured gaze data [17, 30, 61]: spatial locations are sampled from x_i during gaze acquisition, followed by blurring with a Gaussian kernel and normalization to obtain the probability density function (pdf) \tilde{x}_i . This can also be seen as a Gaussian Mixture Model with equal-variance components at measured gaze locations. Let us denote this process of sampling spatial locations and reconstructing a pdf (“SR”) as:

$$\tilde{x}_i = SR(x_i; N), \quad (3)$$

where N is the number of spatial locations sampled from x_i via gaze data capture. For videos, N is equivalently the number of observers.

Given that \tilde{x}_i can be prone to inaccuracies/noise, minimizing J^{real} during training can lead to noise overfitting and suboptimal convergence (see Supplementary Sec. 8). Instead of directly minimizing $d(\hat{x}_i, \tilde{x}_i)$, our approach models the uncertainty in $d(x_i, \tilde{x}_i)$ due to the noise in \tilde{x}_i . We first estimate a probability density function for $d(\hat{x}_i, \tilde{x}_i)$, denoted by $p[d(x_i, \tilde{x}_i)]$, and then train the DNN for saliency prediction by maximizing the likelihood of $d(x_i, \tilde{x}_i)$.

We interpret $d(x_i, \tilde{x}_i)$ as Gaussian random variable with statistics $\mathbb{E}[d(x_i, \tilde{x}_i)]$, $\text{Var}[d(x_i, \tilde{x}_i)]$. We first consider an ideal case where x_i is available and therefore we can compute these statistics by sampling and reconstructing several realizations of \tilde{x}_i from x_i (Eq. 3; no gaze data acquisition needed), and then computing sample mean $\mathbb{E}[d(x_i, \tilde{x}_i)]$ and variance $\text{Var}[d(x_i, \tilde{x}_i)]$. The value of these statistics depends on the number of available gaze locations N used to reconstruct \tilde{x}_i and on the complexity of x_i . For example, when x_i consists of a simple, unimodal distribution – e.g., when only one location in a frame catches the attention of all the observers – a small N is sufficient to bring \tilde{x}_i close to x_i , which leads to low $\mathbb{E}[d(x_i, \tilde{x}_i)]$ and $\text{Var}[d(x_i, \tilde{x}_i)]$ values. Alternatively, for a complex multimodal x_i , a larger N is required for \tilde{x}_i to converge to x_i and consequently, $\mathbb{E}[d(x_i, \tilde{x}_i)]$ and $\text{Var}[d(x_i, \tilde{x}_i)]$ are large when N is small (more discussion on this in Supplementary, Sec. 2).

Our NAT cost function is then defined as the following negative log likelihood:

$$J_{\text{NAT}} = -\ln \prod_i p[d(\hat{x}_i, \tilde{x}_i)] = -\sum_i \ln\{p[d(\hat{x}_i, \tilde{x}_i)]\}, \quad (4)$$

that enables us to account for the presence of noise in the training data, for any choice of d . If $\mathbb{E}[d(x_i, \tilde{x}_i)]$ and $\text{Var}[d(x_i, \tilde{x}_i)]$ are known, and assuming that $d(x_i, \tilde{x}_i)$ is a Gaussian random variable, we can simplify Eq. 4 (see Supplementary, Sec. 1) to get:

$$J_{\text{NAT}}^{\text{ideal}} = \sum_i \{d(\hat{x}_i, \tilde{x}_i) - \mathbb{E}[d(x_i, \tilde{x}_i)]\}^2 / \text{Var}[d(x_i, \tilde{x}_i)]. \quad (5)$$

We note that $J_{\text{NAT}}^{\text{ideal}}$ penalizes \hat{x}_i that are far from \tilde{x}_i , as in the traditional case. However, it also ensures that \hat{x}_i is not predicted *too close* to the noisy \tilde{x}_i , which helps prevent noise overfitting (similar to discrepancy principles applied in image denoising [7, 19]). The penalization is inversely proportional to $\text{Var}[d(x_i, \tilde{x}_i)]$, i.e., it is strong for frames where \tilde{x}_i is a good approximation of x_i . In contrast, $\mathbb{E}[d(x_i, \tilde{x}_i)]$ and $\text{Var}[d(x_i, \tilde{x}_i)]$ are large for multimodal, sparse \tilde{x}_i containing gaze data from only a few observers, since in such cases, \tilde{x}_i is not a good

approximation of x_i . This prevents the NAT formulation from overfitting to such uncertain \tilde{x}_i , by weakly penalizing the errors in \hat{x}_i when compared to \tilde{x}_i .

However, Eq. 5 cannot be implemented in practice, as x_i (and consequently $\mathbb{E}[d(x_i, \tilde{x}_i)]$ and $\text{Var}[d(x_i, \tilde{x}_i)]$) is unknown. We only have access to \tilde{x}_i , a noisy realization of x_i . We therefore turn to approximating the statistics of $d(x_i, \tilde{x}_i)$ as:

$$\mathbb{E}[d(x_i, \tilde{x}_i)] \approx \mathbb{E}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)], \text{Var}[d(x_i, \tilde{x}_i)] \approx \text{Var}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)]. \quad (6)$$

Here, $\tilde{\tilde{x}}_i = SR(\tilde{x}_i; N)$ is the pdf obtained by sampling N spatial locations from \tilde{x}_i , followed by blurring (N is also the number of gaze fixations sampled from x_i by real observers). The difference between how \tilde{x}_i is reconstructed from x_i and $\tilde{\tilde{x}}_i$ from \tilde{x}_i is in the manner of obtaining the N spatial locations: the N spatial locations used to reconstruct \tilde{x} are obtained from human gaze when viewing the i^{th} frame; while for reconstructing $\tilde{\tilde{x}}_i$, N spatial locations are sampled from the pdf \tilde{x}_i . Multiple realizations of $\tilde{\tilde{x}}_i$ are then used to estimate $\mathbb{E}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)]$ and $\text{Var}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)]$. Intuitively, the approximation in Eq. 6 holds because the level of consistency across multiple realizations of $\tilde{\tilde{x}}_i$ would be low when \tilde{x}_i is complex (multimodal) with small N and indicates that the underlying GT saliency map x_i must also be complex. Similarly, a high consistency across multiple realizations of $\tilde{\tilde{x}}_i$ points towards a reliable \tilde{x}_i . Therefore, the spatial noise introduced by sampling from \tilde{x}_i serves as a proxy of the various noise introduced by the insufficient gaze-capturing process. We observe empirically that these approximations hold with a mean absolute percentage error of 10 – 21% on real cases (see Supplementary Sec. 4).

Using Eq. 6, the NAT formulation from Eq. 5 is modified to minimize:

$$J_{\text{NAT}}^{\text{real}} = \sum_i \{d(\hat{x}_i, \tilde{x}_i) - \mathbb{E}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)]\}^2 / \text{Var}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)], \quad (7)$$

where all the terms are now well-defined and a DNN can be trained using this cost function. When implementing Eq. 7, for numerical stability, a small offset of $5e^{-5}$ is applied to the denominator, and $\mathbb{E}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)]$ and $\text{Var}[d(\tilde{x}_i, \tilde{\tilde{x}}_i)]$ are computed using 10 realization of $\tilde{\tilde{x}}_i$.

Fig. 3 shows the mean and standard deviation of $\text{KLD}(\tilde{x}_i || \tilde{\tilde{x}}_i)$ for some frames in ForGED, as estimated by Eq. 6. Frames with high consistency across several observers are considered more reliable for training – a feature that is exploited by NAT in Eq. 7.

4 The ForGED dataset

Videogames present an interesting and challenging domain for saliency methods – given their market value, dynamic content, multiple attractors of visual attention, and dependence of human gaze on temporal semantics. We therefore introduce ForGED, a video-saliency dataset with 480, 13-second clips of Fortnite game play annotated with gaze data from up to 21 observers per video. Compared to popular existing datasets such as LEDOV [19] and DIEM [18], ForGED provides higher dynamism and a video-game context, with the highest number of frames at a consistent 1080p resolution. We summarize the characteristics of each of the datasets used in our experiments in Tab. 1 and show typical ForGED frames in Fig. 3.

Dataset	Videos	Frames	Resolution	Max. Observers used for training	Total Obs. for testing	Content type	mean \pm std.dev. of optical flow magnitude
DIEM [18]	84	240,015	720p@30Hz	31	51-219	everyday videos (shows, advs, etc.)	9.41 \pm 33.85
LEDOV [19]	538	179,336	\geq 720p@24Hz	32	32	human/animal activities	4.09 \pm 8.65
ForGED (ours)	480	374,400	1080p@60hz	5 - 15	15-21	videogame (Fortnite)	27.26 \pm 39.08

Table 1: Characteristics of video-saliency datasets, including the proposed ForGED dataset.

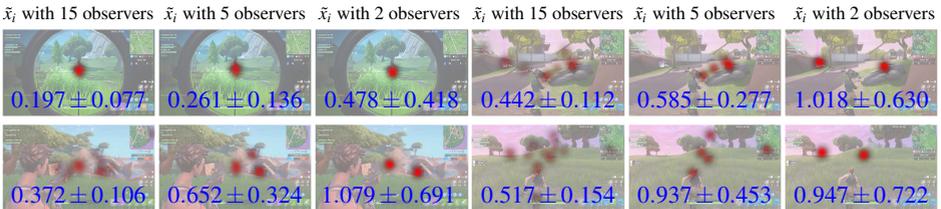


Figure 3: Typical frames from ForGED with gaussian-blurred gaze locations of specified number of observers overlaid in red. For each image, we also show $\mathbb{E}[\text{KLD}(\bar{x}_i||\tilde{x}_i)] \pm \text{Std}[\text{KLD}(\bar{x}_i||\tilde{x}_i)]$. These quantities increase when the saliency map is sparse/multimodal and number of observers is small – a setting that reduces the reliability of a frame for training. *ForGED images have been published with the permission from Epic Games.*

Dynamic content in ForGED. To compare the dynamic content level of ForGED to those of LEDOV and DIEM, we use RAFT [56] to compute the mean and standard deviation of the magnitude of the optical flow on a random subset of 100,000 frames from the three datasets, at a uniform 1080p resolution and 30 fps framerate (Tab. 1). This is in ForGED approximately $3\times$ that of DIEM and more than $6\times$ larger than LEDOV, suggesting that objects move faster (on average) in ForGED. It also has the largest standard deviation suggesting a larger variety of motion magnitudes in ForGED.

Gaze data acquisition and viewing behavior in ForGED. To acquire ForGED, we first recorded 12 hours of Fortnite Battle Royale game-play videos from 8 players of varying expertise using OBS [11]. We then sampled 480 15-second clips to show to a different set of 102 participants with varying degree of familiarity with Fortnite. Each viewer was tasked with viewing a total of 48 clips, randomly sampled from the pool of 480, and interspersed with 3-second “intervals” showing a central red dot on grey screen [60] to ensure consistent gaze starting point for each clip (total 15-minute viewing time per viewer). Each participant viewed the video clips on a 1080p monitor situated approximately 80cm away, while their gaze was recorded with Tobii Tracker 4C at 90Hz. After analyzing the gaze patterns, we discarded the initial 2 seconds of each clip, when observers were mostly spending time to understand the context, to get a total of 374,400 video frames annotated with gaze data. Accumulating the gaze of all frames, we observe that ForGED presents a bias towards the frame center and top-right corner. This is because in Fortnite the main character and the crosshair lie at the screen center – making it an important region – and the mini-map on the top right corner attracts regular viewer attention to understand the terrain. Such a bias is uniquely representative of the observer behavior not only in Fortnite, but also in other third person shooting games with similar scene layouts. Further analysis of gaze data in ForGED, such as IOC curves, visual comparison of gaze data biases in ForGED, LEDOV and DIEM, are presented in the Supplementary, Sec. 3.

5 Results

We compare TT (Eq. 2) and NAT (Eq. 7) on three datasets (ForGED, LEDOV [49], and DIEM [88]) and three DNNs (ViNet [27], the state-of-the-art on DHF1K [59]; TASED-Net, a 3D-CNN-based architecture [47]; and SaleMA, an RNN-based architecture [42]). We further evaluate NAT against TT when density-based (e.g., KLD) or fixation-based (e.g., NSS) discrepancy functions are used as $d(\cdot, \cdot)$ in J^{real} (Eq. 2) and $J_{\text{NAT}}^{\text{real}}$ (Eq. 7). We first evaluated

and improved the author-specified hyperparameters for ViNet, TASED-Net, and SaleMA, by performing TT (Eq. 2) on the entire LEDOV training set (see Tab. 2). We use the improved settings for our experiments (see Supplementary Sec. 6). We also verify that training on existing saliency datasets does not generalize to ForGED (Tab. 3c), given its novel content.

Experimental setup: We want to compare TT and NAT when training with different amounts of data and varying levels of accuracy/gaze-data completeness.

We emulate small-size training datasets from LEDOV, DIEM, and ForGED by controlling the number of fixations, N , used to reconstruct \tilde{x} in the training set and the number of training videos, V , used.

We report the performance evaluation of TT and NAT on test set for each (V, N) value used for the training set. The values for V and N are chosen to gradually increase the training dataset size and accuracy until the maximum V and/or N is reached. To reconstruct \tilde{x} , we choose a kernel of size $\sim 1^\circ$ viewing angle [47, 50, 51] and discuss alternative \tilde{x} reconstruction strategies [57, 59, 40] in Sec. 6 (and Sec. 7 in the Supplementary).

ForGED data are randomly split into 379 videos for training, 26 for validation, and 75 for testing. For LEDOV, we adopt the train / val / test split specified by the authors. DIEM contains gaze data from many observers on a few videos: we use 60 videos with fewest observers for training and evaluate on the remaining videos with 51 – 219 observers. Evaluation is performed on test-set maps reconstructed from the set of *all* the available observers, that is sufficiently large to lead to converged IOC curves even for multimodal maps (see supplementary video for ForGED test set multimodality); consequently, we also assume a negligible noise level in evaluation. We omit experimenting with DHF1K in favor of LEDOV which is similar in scope to DHF1K [55], but contains a larger number of observers (converged IOC curves), while DHF1K lacks accurate per-observer gaze data.

Dataset type and size: We compare NAT and TT on different dataset types, by training ViNet and TASED-Net on ForGED, LEDOV, and DIEM, and changing V and N to assess the performance gain of NAT as a function of the level of accuracy and completeness of the training dataset. Tab. 3a and 3b show the results for ViNet trained on ForGED and LEDOV, whereas Tab. 4a and 5a show the results for TASED-Net. With ViNet, we observe a consistent performance gain of NAT over TT. Although NAT is particularly advantageous when N and V are small, training on the *entire* LEDOV dataset (last row in Tab. 3a) also shows a significant improvement for NAT since, depending on their content, some frames can still have insufficient fixation data. With TASED-Net trained on ForGED, NAT consistently outperforms TT when the number of training videos is ≤ 100 , *i.e.*, when noise overfitting may occur. Notably, NAT on 30 videos / 15 observers and 100 videos / 5 observers is comparable or superior to TT with 379 videos / 5 observers, which corresponds to $\geq 3\times$ saving factor in terms of the data required for training. Similar conclusions can be drawn for LEDOV (Tab. 5a) and DIEM (see Supplementary). We also test the case of practical importance of an *unbalanced* LEDOV dataset, with an uneven number of observers in the training videos. Since NAT, by design, accounts for the varying reliability of the gaze data in training frames, it significantly outperforms TT (last two rows of Tab. 5a).

method, hyperparameter settings	KLD↓	CC↑	SIM↑	NSS↑	AUC-J↑
ViNet, Adam, 0.0001, KLD (default)	0.806	0.697	0.569	3.781	0.881
ViNet, RMSprop, 0.0001, KLD (improved)	0.773	0.710	0.573	3.969	0.889
TASED-Net, SGD, learning rate schedule (default)	1.104	0.554	0.452	2.536	0.828
TASED-Net, RMSprop, 0.001, KLD (improved)	0.754	0.724	0.572	4.227	0.921
SaleMA, Adam, $1e^{-5}$, BCE (default)	1.238	0.511	0.412	2.426	0.894
SaleMA, RMSprop, $1e^{-5}$, KLD (improved)	1.052	0.612	0.463	3.237	0.912

Table 2: LEDOV test-set performance when trained (traditionally) with default and improved settings for ViNet [47], TASED-Net [44], and SaleMA [42].

train videos V		train obs. N	loss	KLD \downarrow	CC \uparrow	SIM \uparrow	NSS \uparrow	AUC-J \uparrow	
30	5	TT	1.358	0.541	0.436	3.261	0.713		
		NAT	1.264	0.593	0.460	3.412	0.773		
		TT	1.779	0.514	0.399	3.130	0.633		
		NAT	1.220	0.620	0.488	3.670	0.764		
		TT	1.218	0.602	0.473	3.542	0.794		
		NAT	1.257	0.605	0.469	3.527	0.773		
	100	5	TT	1.263	0.600	0.473	3.609	0.775	
			NAT	1.149	0.623	0.485	3.620	0.798	
			TT	1.134	0.629	0.494	3.750	0.804	
		200	5	NAT	0.982	0.641	0.489	3.704	0.882
				TT	0.994	0.645	0.495	3.697	0.860
				NAT	1.026	0.625	0.438	3.505	0.918
379	5	TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
		TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
	100	2	TT	1.323	0.565	0.365	3.034	0.890	
			NAT	1.056	0.610	0.447	3.386	0.922	
			TT	1.065	0.623	0.473	3.627	0.917	
		5	NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
			NAT	0.974	0.632	0.470	3.497	0.932	
379	2	TT	0.963	0.631	0.461	3.376	0.936		
		NAT	0.888	0.664	0.508	3.813	0.934		
		TT	0.974	0.632	0.470	3.497	0.932		
	5	NAT	0.888	0.664	0.508	3.813	0.934		
		TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
30	5	TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
		TT	1.323	0.565	0.365	3.034	0.890		
		NAT	1.056	0.610	0.447	3.386	0.922		
	100	2	TT	1.065	0.623	0.473	3.627	0.917	
			NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
		5	NAT	0.974	0.632	0.470	3.497	0.932	
			TT	0.963	0.631	0.461	3.376	0.936	
			NAT	0.888	0.664	0.508	3.813	0.934	
379	5	TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
		TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
	100	2	TT	1.323	0.565	0.365	3.034	0.890	
			NAT	1.056	0.610	0.447	3.386	0.922	
			TT	1.065	0.623	0.473	3.627	0.917	
		5	NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
			NAT	0.974	0.632	0.470	3.497	0.932	
379	2	TT	0.963	0.631	0.461	3.376	0.936		
		NAT	0.888	0.664	0.508	3.813	0.934		
		TT	0.974	0.632	0.470	3.497	0.932		
	5	NAT	0.888	0.664	0.508	3.813	0.934		
		TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
30	5	TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
		TT	1.323	0.565	0.365	3.034	0.890		
		NAT	1.056	0.610	0.447	3.386	0.922		
	100	2	TT	1.065	0.623	0.473	3.627	0.917	
			NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
		5	NAT	0.974	0.632	0.470	3.497	0.932	
			TT	0.963	0.631	0.461	3.376	0.936	
			NAT	0.888	0.664	0.508	3.813	0.934	
379	5	TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
		TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
	100	2	TT	1.323	0.565	0.365	3.034	0.890	
			NAT	1.056	0.610	0.447	3.386	0.922	
			TT	1.065	0.623	0.473	3.627	0.917	
		5	NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
			NAT	0.974	0.632	0.470	3.497	0.932	
379	2	TT	0.963	0.631	0.461	3.376	0.936		
		NAT	0.888	0.664	0.508	3.813	0.934		
		TT	0.974	0.632	0.470	3.497	0.932		
	5	NAT	0.888	0.664	0.508	3.813	0.934		
		TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
30	5	TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
		TT	1.323	0.565	0.365	3.034	0.890		
		NAT	1.056	0.610	0.447	3.386	0.922		
	100	2	TT	1.065	0.623	0.473	3.627	0.917	
			NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
		5	NAT	0.974	0.632	0.470	3.497	0.932	
			TT	0.963	0.631	0.461	3.376	0.936	
			NAT	0.888	0.664	0.508	3.813	0.934	
379	5	TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
		TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
	100	2	TT	1.323	0.565	0.365	3.034	0.890	
			NAT	1.056	0.610	0.447	3.386	0.922	
			TT	1.065	0.623	0.473	3.627	0.917	
		5	NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
			NAT	0.974	0.632	0.470	3.497	0.932	
379	2	TT	0.963	0.631	0.461	3.376	0.936		
		NAT	0.888	0.664	0.508	3.813	0.934		
		TT	0.974	0.632	0.470	3.497	0.932		
	5	NAT	0.888	0.664	0.508	3.813	0.934		
		TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
30	5	TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
		TT	1.323	0.565	0.365	3.034	0.890		
		NAT	1.056	0.610	0.447	3.386	0.922		
	100	2	TT	1.065	0.623	0.473	3.627	0.917	
			NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
		5	NAT	0.974	0.632	0.470	3.497	0.932	
			TT	0.963	0.631	0.461	3.376	0.936	
			NAT	0.888	0.664	0.508	3.813	0.934	
379	5	TT	1.385	0.546	0.370	2.992	0.877		
		NAT	1.298	0.558	0.385	3.161	0.903		
		TT	1.419	0.536	0.370	3.042	0.877		
		NAT	1.172	0.590	0.428	3.372	0.908		
		TT	1.080	0.615	0.481	3.598	0.897		
		NAT	0.995	0.634	0.478	3.590	0.924		
	100	2	TT	1.323	0.565	0.365	3.034	0.890	
			NAT	1.056	0.610	0.447	3.386	0.922	
			TT	1.065	0.623	0.473	3.627	0.917	
		5	NAT	0.969	0.643	0.494	3.749	0.923	
			TT	0.986	0.628	0.475	3.434	0.925	
			NAT	0.974	0.632	0.470	3.497	0.932	
379	2	TT	0.963	0.631	0.461	3.376	0.936		
		NAT							

DNN architectures: Tab. 5a-b compares NAT vs. TT when training two different DNNs (TASED-Net [47] and SaEMA [42]) on LEDOV, with KLD. As also observed earlier, NAT outperforms TT and the performance gap shrinks with increasing training data. The Supplementary (Sec. 5) shows results with SaEMA on ForGED.

6 Discussion and conclusions

NAT for images: Image-based saliency datasets (e.g., CAT2000 [10], SALICON [51]) have many fixations per image resulting in high-quality of the reconstructed saliency maps, as the accuracy rapidly increases with number of fixations (e.g., > 90% accuracy at 20 fixations [62]). It is nonetheless fair to ask if NAT is effective for image-saliency predictors.

We simulate a high-noise, incomplete, dataset by sampling a subset of fixations for each SALICON image¹ and train a state-of-the-art method, EML-Net [28], with TT and NAT. Tab. 6 shows the results

no. of fixations	loss	KLD↓	CC↑	SIM↑	NSS↑	AUC-J↑
5	TT	3.986	0.578	0.537	1.477	0.764
	NAT	1.672	0.660	0.611	1.549	0.817
15	TT	2.877	0.655	0.589	1.669	0.795
	NAT	1.437	0.714	0.640	1.676	0.831

Table 6: Evaluation on EML-Net.

on the official SALICON benchmark test set, and confirms the advantage of NAT.

Alternative methods to reconstruct \tilde{x} : Although reconstructing \tilde{x} by blurring a binary map of fixations is prevalent practice [47, 50, 51], we experiment with another reconstruction strategy for \tilde{x} using Gaussian KDE with a uniform regularization. The optimal KDE bandwidth and regularization weight is estimated by optimizing a gold-standard model [63, 65] (see Supplementary, Sec. 7). Experiments with TASED-Net on ForGED ($N = 5$, $V = 30$) comparing TT with \tilde{x} estimated using a fixed-size blur or KDE-based reconstruction, and NAT, show that while KDE improves TT, NAT still yields the best results (Tab. 7).

Limitations and future work:

Although the test saliency maps of LEDOV, DIEM and ForGED are derived from several observers

\tilde{x}_i for training	loss	KLD↓	CC↑	SIM↑	NSS↑	AUC-J↑
1° blur	TT	1.419	0.536	0.370	3.042	0.877
KDE	TT	1.223	0.573	0.399	3.271	0.897
1° blur	NAT	1.172	0.590	0.428	3.372	0.908

Table 7: Methods for estimating \tilde{x} .

leading to converged IOC on *average*, per-frame inaccuracies of saliency maps can still add uncertainty about the conclusions one can draw. Adopting alternative strategies such as deriving metric-specific saliency from the probabilistic output of a saliency predictor [42, 40], can give a clearer understanding. Nonetheless, in our experiments all the metrics are generally in agreement about the ranking between TT and NAT: a strong evidence in favor of NAT [62]. NAT design principles can also be applied to saliency evaluation (not only training), where variable importance is given to each frame depending on its noise level.

Conclusion: Video gaze data acquisition is time-consuming and can be inaccurate. To reduce the impact of dataset size in the field of visual saliency prediction, we introduce NAT to account for the level of reliability of a saliency map. We also introduce a new dataset which offers a unique video-game context. We show consistent improvements for NAT over TT across a variety of experiments. The adoption of NAT has important practical implications, since it allows acquiring new datasets (or training on old ones) with less data, both in terms of videos and number of observers, without loss of quality.

¹Mouse clicks are used as proxy for gaze in SALICON.

Acknowledgments

We thank Abhishek Badki and Shoaib Ahmed Siddiqui for help with distributed training, Jan Kautz, Ming-Yu Liu, and Arash Vahdat for technical discussions, Michael Stengel for help with gaze-recording software, Paul Weakliem and Fuzzy Rogers for UCSB computational infrastructure, and the Fortnite players and observers for their help with creating ForGED. This work was supported in part through computational facilities purchased using NSF grant OAC-1925717 and administered by the Center for Scientific Computing (CSC) at UCSB. The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 1720256).

References

- [1] Open broadcaster software. <https://obsproject.com/>.
- [2] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. Dr(eye)ve: A dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016.
- [3] Duin Baek, Hangil Kang, and Jihoon Ryoo. Sali360: Design and implementation of saliency based video compression for 360° video streaming. In *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020.
- [4] Cagdas Bak, Aysun Kocak, Erkut Erdem, and Aykut Erdem. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 2017.
- [5] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani. Recurrent mixture density network for spatiotemporal visual attention. *arXiv*, 2016.
- [6] Giovanni Bellitto, Federica Proietto Salanitri, Simone Palazzo, Francesco Rundo, Daniela Giordano, and Concetto Spampinato. Video saliency detection with domain adaption using hierarchical gradient reversal layers. *arXiv*, 2020.
- [7] M Bertero, P Boccacci, G Talenti, R Zanella, and L Zanni. A discrepancy principle for poisson data. *Inverse Problems*, 2010.
- [8] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.
- [9] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2012.
- [10] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015.
- [11] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015.
- [12] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

- [13] Anne Chao and Tsung-Jen Shen. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 2003.
- [14] Sylvain Chartier and Patrice Renaud. An online noise filter for eye-tracker data recorded in a virtual environment. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, 2008.
- [15] Arturo Deza and Miguel P. Eckstein. Can peripheral representations improve clutter metrics on complex scenes? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [16] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified Image and Video Saliency Modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [17] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 2009.
- [18] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Chi conference on human factors in computing systems*, 2017.
- [19] I. Frosio and J. Kautz. Statistical nearest neighbors for image denoising. *IEEE Transactions on Image Processing (TIP)*, 2019.
- [20] Guangshuai Gao, Wenting Zhao, Qingjie Liu, and Yunhong Wang. Co-saliency detection with co-attention fully convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.
- [21] Wilson S Geisler and Jeffrey S Perry. Real-time foveated multiresolution system for low-bandwidth video communication. In *Real-time foveated multiresolution system for low-bandwidth video communication*, 1998.
- [22] Siavash Gorji and James J Clark. Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [23] D Holste, I Grosse, and H Herzel. Bayes' estimators of generalized entropies. *Journal of Physics A: Mathematical and General*, 1998.
- [24] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [25] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 2000.
- [26] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1998.

- [27] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. ViNet: Pushing the limits of visual modality for audio-visual saliency prediction. In *arXiv*, 2021.
- [28] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 2020.
- [29] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. DeepVS: A deep learning based video saliency prediction approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [30] Lai Jiang, Mai Xu, Zulin Wang, and Leonid Sigal. DeepVS2.0: A saliency-structured deep learning method for predicting dynamic visual attention. *International Journal of Computer Vision (IJCV)*, 2021.
- [31] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [32] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012.
- [33] Anton S. Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (SIGGRAPH)*, 2019.
- [34] Jonghyun Kim, Youngmo Jeong, Michael Stengel, Kaan Akşit, Rachel Albert, Ben Boudaoud, Trey Greer, Joohwan Kim, Ward Lopes, Zander Majercik, Peter Shirley, Josef Spjut, Morgan McGuire, and David Luebke. Foveated ar: Dynamically-foveated augmented reality display. *ACM Transactions on Graphics (SIGGRAPH)*, 2019.
- [35] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*. 1987.
- [36] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing (TIP)*, 2017.
- [37] Matthias Kümmeler, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv*, 2014.
- [38] Matthias Kümmeler, Thomas SA Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 2015.
- [39] Matthias Kümmeler, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [40] Matthias Kümmeler, Thomas SA Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

- [41] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 2013.
- [42] Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Noel E O'Connor, Xavier Giro-i Nieto, and Kevin McGuinness. Simple vs complex temporal recurrences for video saliency prediction. *arXiv*, 2019.
- [43] Nian Liu and Junwei Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing (TIP)*, 2018.
- [44] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [45] Nian Liu, Ni Zhang, Kaiyuan Wan, Junwei Han, and Ling Shao. Visual saliency transformer. *arXiv*, 2021.
- [46] Yufan Liu, Songyang Zhang, Mai Xu, and Xuming He. Predicting salient face in multiple-face videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [47] K. Min and J. Corso. TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [48] Parag Mital, Tim Smith, Robin Hill, and John Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 2011.
- [49] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv*, 2017.
- [50] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 2005.
- [51] Navyasri Reddy, Samyak Jain, Pradeep Yarlagadda, and Vineet Gandhi. Tidying deep saliency prediction architectures. In *International Conference on Intelligent Robots and Systems*, 2020.
- [52] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [53] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2008.
- [54] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 1991.

- [55] Matthias Tangemann, Matthias Kümmerer, Thomas S.A. Wallis, and Matthias Bethge. Measuring the importance of temporal features in video saliency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [56] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [57] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications*, 2020.
- [58] Eleonora Vig, Michael Dorr, and David Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.
- [59] W. Wang, J. Shen, F. Guo, M. Cheng, and A. Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [60] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing (TIP)*, 2017.
- [61] Niklas Wilming, Torsten Betz, Tim C Kietzmann, and Peter König. Measures and limits of models of fixation selection. *PloS one*, 2011.
- [62] Xinyi Wu, Zhenyao Wu, Jinglin Zhang, Lili Ju, and Song Wang. SalSAC: A video saliency prediction model with shuffled attentions and correlation-based convlstm. In *AAAI*, 2020.
- [63] Jing Zhang, Jianwen Xie, and Nick Barnes. Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [64] Ni Zhang, Junwei Han, Nian Liu, , and Ling Shao. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [65] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl S. Muller, Jake A. Whritner, Luxin Zhang, Mary M. Hayhoe, and Dana H. Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. *arXiv*, 2019.