DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling

Akis Linardos*
University of Barcelona
linardos.akis@gmail.com

Matthias Kümmerer*
University of Tübingen
matthias.kuemmerer@bethgelab.org

Ori Press
University of Tübingen
ori.press@bethgelab.org

Matthias Bethge
University of Tübingen
matthias@bethgelab.org

Abstract

Since 2014 transfer learning has become the key driver for the improvement of spatial saliency prediction—however, with stagnant progress in the last 3-5 years. We conduct a large-scale transfer learning study which tests different ImageNet backbones, always using the same read out architecture and learning protocol adopted from DeepGaze II. By replacing the VGG19 backbone of DeepGaze II with ResNet50 features we improve the performance on saliency prediction from 78% to 85%. However, as we continue to test better ImageNet models as backbones—such as EfficientNetB5—we observe no additional improvement on saliency prediction. By analyzing the backbones further, we find that generalization to other datasets differs substantially, with models being consistently overconfident in their fixation predictions. We show that by combining multiple backbones in a principled manner a good confidence calibration on unseen datasets can be achieved. This new model "DeepGaze IIE" yields a significant leap in benchmark performance in and out-of-domain with a 15 percent point improvement over DeepGaze II to 93% on MIT1003, marking a new state of the art on the MIT/Tuebingen Saliency Benchmark in all available metrics (AUC: 88.3%, sAUC: 79.4%, CC: 82.4%).

1. Introduction

Saliency detection is involved in many sensory modalities. It summarizes the associated mechanisms as the ability of humans and animals to allocate their attention to the most important subsets of the data. In vision, this means attending to the elements of a visual input that stand out from their neighbouring regions, and visual saliency is usually operationalized by measuring fixations locations. Accordingly, in computer vision, saliency prediction currently refers to either predicting fixation locations or detecting salient objects.

Early on, researchers found out that the locations of fixations are statistically influenced by features of the visual stimuli that include both high-level properties such as people [41] and low-level ones such as spatial contrast [33]. Soon after the *Feature Integration Theory* emerged [38], Koch and Ullmann outlined a computational mechanism to model attention [21] which was implemented thirteen years later by Itti et al [15].

The Itti-Koch model was the first to predict a saliency map from any arbitrary image without the need to precompute elementary features allowing for a wide range of applications. This paved the way for many interesting saliency prediction models [2, 20, 42] leading up to the present day where deep learning models are dominating the field [39, 25, 28, 32, 16, 23] driven by large scale saliency datasets [19, 17, 1]. As the saliency domain has substantially less data compared to some of the more prominent computer vision tasks, transfer learning has become the key driver for the improvement of

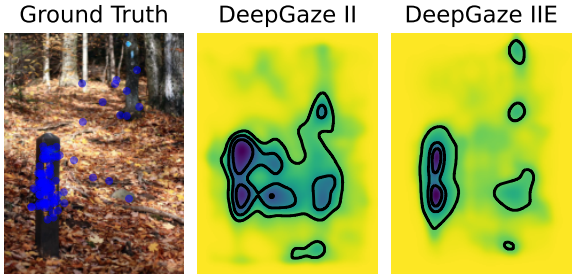

Figure 1. By leveraging the diversity of different backbones, our new saliency model DeepGaze IIE very is able to predict human fixation locations very accurately.

* indicates joint first authorship

spatial saliency prediction. One of the earliest works on transfer learning for deep learning is DeCAF [6], where the authors used features extracted by a deep CNN that was trained on object recognition, leveraging a large dataset to tackle generic tasks that lacked labeled data. Following this transfer learning scheme, they outperformed the state-of-the-art on various vision challenges. Inspired by the huge success of deep convolutional models in the domain of classification and particularly the ImageNet benchmark [5], DeepGaze I [25] was the first to transfer ImageNet learned features to the domain of saliency. Since then all high-performing saliency models use ImageNet as pretext task.

To this day, the problem of spatial saliency is far from solved and the simple case of the MIT300 benchmark [18] illustrates a substantial gap between existing models and the lower bound on the explainable information (e.g., IG of 0.951 vs 1.317 and sAUC of 0.784 vs 0.823). In 2014, the introduction of deep learning and transfer learning in particular, ushered a new era for saliency prediction after several years of stagnating performance. Similarly, there has been only gradual progress in the recent 3-5 years, notwithstanding the significant amount of models proposed during that time (Figure 2). From a machine learning point of view, the task of saliency prediction is conceptually interesting as it requires well-calibrated probabilistic predictions that are less critical in the much more common setting of highly deterministic classification problems.

In this work, we significantly improve spatial saliency modeling by studying how to achieve well-calibrated probabilistic predictions. Beyond proposing a new state-of-the-art model, we make a systematic analysis of the extent to which higher ImageNet performance leads to higher performance in the saliency domain. Specifically, we utilize a broad range of models that have achieved state of the art on ImageNet as fixed feature backbones for the saliency prediction task, using a pointwise nonlinear read out following the DeepGaze II architecture and learning schedule as described in [28]. Additionally, we study the complementarity between these models and leverage it by conducting an ensemble learning approach which ends up yielding a new state of the art, closing the gap between models and inter-observer consistency in all metrics.

To gain additional insights into the differences between the backbones, we study the confidence calibration of the models based on them. Confidence calibration is especially relevant when applying models in out-of-domain contexts where we would expect a good model to realize the domain shift and decrease its confidence accordingly [31]. Many established confidence calibration measures [9] are not applicable in situations of very high stochasticity such as fixation prediction, therefore we propose a new method for testing confidence calibration which can be applied on datasets with high entropy. Instead of being well calibrated or conservatively underconfident, we find that most individual models are highly overconfident on out-of-domain data, while our ensemble models show much better confidence calibration, which makes them more trustworthy on unseen datasets.

2. Related Work

Classic models have relied on hand engineered features to tackle saliency prediction [15, 37, 2, 20, 42]. Saliency prediction has since then moved on to deep learning models, the first of which was eDN [39]. However one major hurdle when tackling saliency prediction with deep learning models is the small size of available data, stemming from the fact that collecting fixation data is both time-consuming and expensive. On top of that, the true saliency of an image is liable to shift when transformations are applied on it, severely limiting potential augmentations [4]. The first work that applied transfer learning to the saliency domain was DeepGaze I [25] which has since then evolved to DeepGaze II that was built on VGG19 [28]. After DeepGaze I virtually every high-performing saliency model used transfer learning, usually based on ImageNet. Among the works that have focused on a principled transfer learning scheme for saliency prediction in the past was [14], which trained a saliency model on deep features from three CNNs (AlexNet, GoogleNet, and VGG16), combining low and high level pre-trained features, with a support vector machine on top and DeepFeat [30] where the authors used a *fixed* architecture on top of three pretrained CNN's features (ResNet, VGG, GoogleNet) to predict saliency. The EML-NET[16] model introduced a scalable method to combine multiple deep convolutional networks of any complexity as encoders for features relevant to visual saliency.

Other models engineered complex deep architectures or build upon existing ones that have shown merit in other tasks, but all of them used transfer learning by pretraining their architectures with larger datasets as a starting point. SalGAN [32] and GazeGAN [4] both used adversarial losses to train their saliency prediction models, which in the first case consist of an encoder-decoder architecture while the second one built on a U-net structure. The MSI-NET[23] tackled the task by integrating global scene information in its encoder-decoder architecture. UNISAL unified the image and video modalities of saliency to harness the entirety of saliency prediction datasets [7].

Arguably, DeepFeat [30] and EML-NET [16] are the two works closest related to our own so contrasting with them makes it easier to highlight the finer shades of contribution in our work. DeepFeat uses a fixed linear readout on top of the pretrained features, whereas we fine tune a readout network that consists of 1×1 convolutions following the DeepGaze II paradigm [28]. Features extracted by multilayer convolutional networks don't have a well defined scale due to many possible transformations, deeming the usage of a

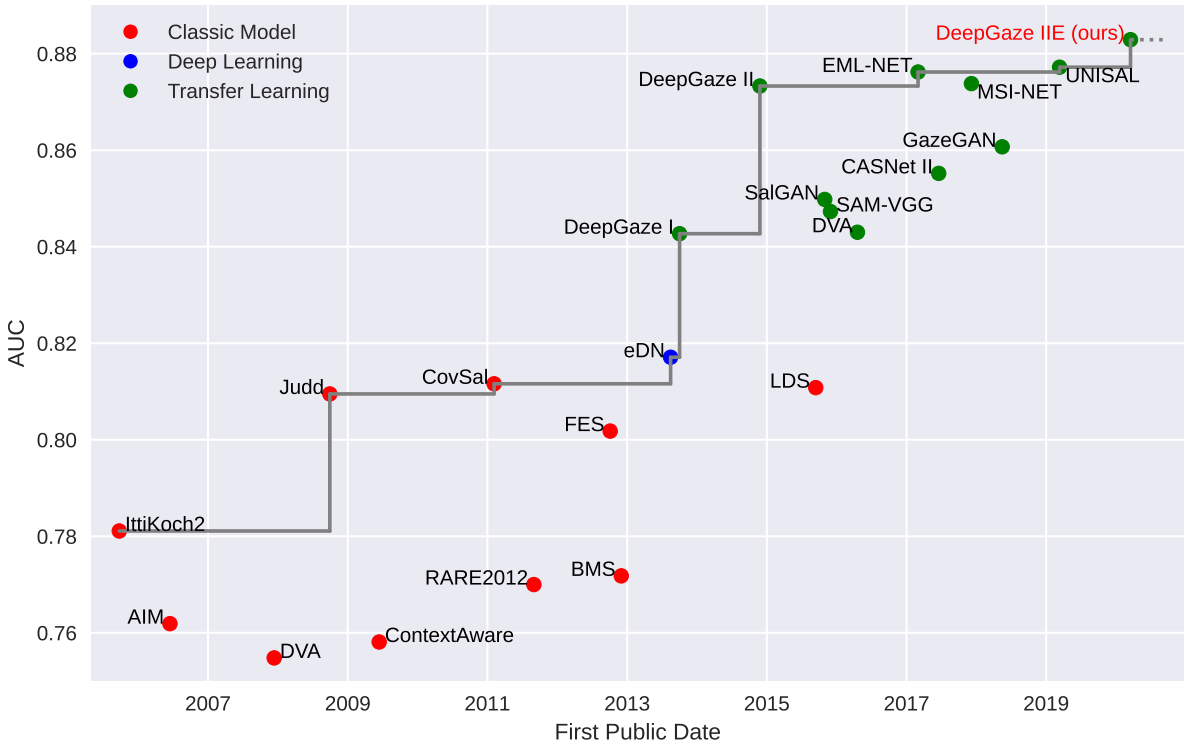

Figure 2. A visualization of progress in saliency prediction over the last 15 years. The displayed dates correspond to the earliest date we found the models available, and usually reflect their first date the model was tested or the date of publication (whichever came first). The AUC corresponds to MIT300 evaluation at the MIT/Tuebingen saliency benchmark [24]. For readability purposes we limited the scale of the plot to models whose AUC score is above 0.75. The gray line indicates state-of-the-art performance with respect to the models listed in the MIT/Tuebingen Saliency Benchmark. We could not include models which are only evaluated on the predecessor benchmark saliency.mit.edu since the evaluation changed slightly, resulting in different model scores.

rigid linear readout too constrained for this type of input. In contrast, a readout network of 1×1 convolutions is able to learn nonlinear transformations adjusting the scale of the input features and leverage interactions between those features. The small kernel size means that the network is unable to learn new spatial features but rather combines the ones given as input, making it an ideal tool for comparing the feature predictivity between different backbones for any given task. Aside from this major difference, we also conduct a series of studies that reveal how different models perform differently and combine their fixation densities to leverage their complementarity.

EML-NET aimed at maximal prediction performance, but we aim at understanding how much relevant information about fixation placement is encoded in deep features. To that end, we compare not only two but a large number of relevant ImageNet trained models. EML-NET is training each CNN model at the encoder stage while we keep ours fixed, which not only is less costly but also a much stronger scientific tool for studying the generalizeability of ImageNet trained features. Added to that, EML-NET combines these models at the encoder stage for a more broad prior knowledge, while in our case we study each model separately, delineating their individual contribution and later combine their predicted densities instead.

Finally, compared to either of these works, we use a much broader array of state of the art ImageNet CNNs as backbones to our architectures and we train an agglomeration of each model configuration, accounting for the uncertainty in our metrics.

3. Methods

3.1. Model and Training Pipeline

The overall pipeline is visualized in Figure 3, where the final model is derived from the combination of multiple backbones after a series of principled analyses steps. An image is first processed with a backbone CNN to extract deep activations, which are subsequently processed in a readout network of 1×1 convolutions. The single output channel of the readout network is blurred, combined with a center-bias and fed through a softmax to yield a two-dimensional fixation distribution (Figure 3a). Essentially, this is an adaptation of the architecture of DeepGaze II [28] with a deeper

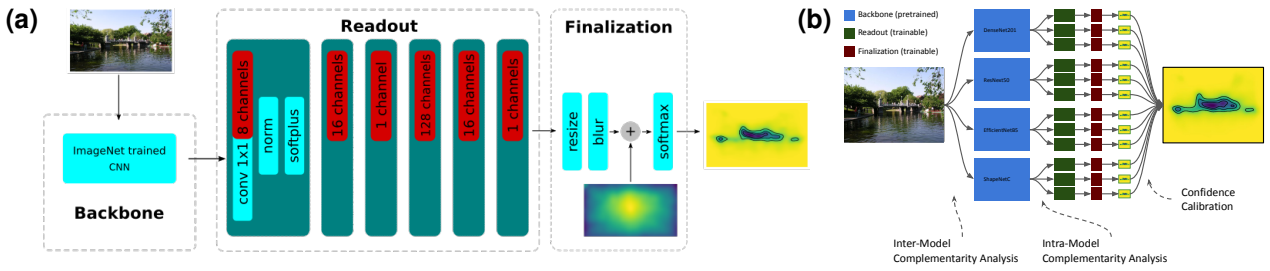

Figure 3. A diagram of our adapted DeepGaze II architecture as was used in all our experiments, as well as our best performing variant, DeepGaze IIE. (a) testing backbones: we collect some layers from CNNs pretrained on ImageNet without any additional training. We apply a readout network on these layers that consists of blocks of 1×1 convolutions, a layernorm and a softplus function. Afterwards, a blur and a center bias prior are applied before a softmax that gives us the final probability density of fixations. (b) The ensemble model DeepGaze IIE: we combine some of the state of the art ImageNet backbones, leveraging inter- and intra- model complementarity which is analyzed in section 4.2. Confidence Calibration is used as an analytical tool to better understand why these models perform best.

readout network, layer norm and softplus instead of ReLUs as activation function, and, most notably, with different backbones instead of the original VGG19 network. The readout network, along with the blur size and the centerbias weight are the only parts of the pipeline that undergo training. The feature extractor's weights are kept fixed during this process. Since our model predicts a fixation density, we have direct access to the likelihood of fixations and therefore we optimize our model for maximum likelihood. We first pretrain our model on the SALICON dataset [17] and then finetune it on the MIT1003 dataset [19]. SALICON includes 10,000 images whose ground truth was collected using mouse traces as indicated by observers rather than a gaze detector. Albeit this seems to sacrifice precision, SALICON makes a good starting point as it has been proven to be very useful for pretraining saliency models. MIT1003 is composed of 1003 natural images tested on 15 subjects (with a presentation time of 3 seconds). The dataset contains images of various dimensions which we resized to either 1024×786 or 768×1024. Images were downsampled by a factor of 1.5 for SALICON and 2.0 for MIT1003/MIT300. We use a learning rate scheduler starting with an initial learning rate of 0.001 which then decays by a factor of 10 every set number of epochs.

We evaluate each configuration of our model following a 10-fold cross validation scheme on the MIT1003 dataset. In simple terms, given an MIT1003 image there is exactly one out of the ten models from this procedure that did neither see this image during training nor for validation in hyperparameter tuning, so that its predicted density is suitable for evaluation. Thus all reported metrics reflect test performance.

3.2. Metrics

As our main guide during our experiments, we used the Information Gain metric [26], which is effectively the difference in average log-likelihood of the model and a baseline model. Therefore, the metric measures the extent by which a model's knowledge has surpassed that of the baseline model. Since it's known that human fixations tend to accumulate towards the center of an image, we use an image-invariant center bias as baseline model.

For a model which predicts a fixation density $p(x \mid I)$ over possible fixation locations x given an image I, the information gain is computed as

$$IG(\text{model}) = \frac{1}{N}\sum_{i}^{N} \log_2 p_{\text{model}}(x_i \mid I_i) - \log_2 p_{\text{baseline}}(x_i),$$

where x_i is the ith fixation of the dataset, taking place in image I_i.

We consider Information Gain the most principled metric [26] and thus mostly rely on it, but we evaluate on other commonly used saliency metrics later on. These include AUC, shuffled AUC, KL divergence, Correlation Coefficient, and Normalized Scanpath Saliency [3]. Saliency metrics are known to be quite inconsistent when evaluated on the same saliency map [3]. However, it has recently been shown that this problem can be mitigated for probabilistic models by evaluating each metric on the saliency maps which has highest expected performance under the fixation density predicted by the model [27].

3.3. Testing Confidence Calibration

One key feature of saliency models is that they predict probabilistic fixation distributions rather than deterministic classes. This means that our models predict not only qualitatively which regions they expect to be fixated, but also quantitatively how much more often they expect a certain salient region to be fixated than any other given region. By comparing this to the actual numbers of fixations in the low-density and high-density regions, we can check how well calibrated the model confidence is—i.e. whether it makes

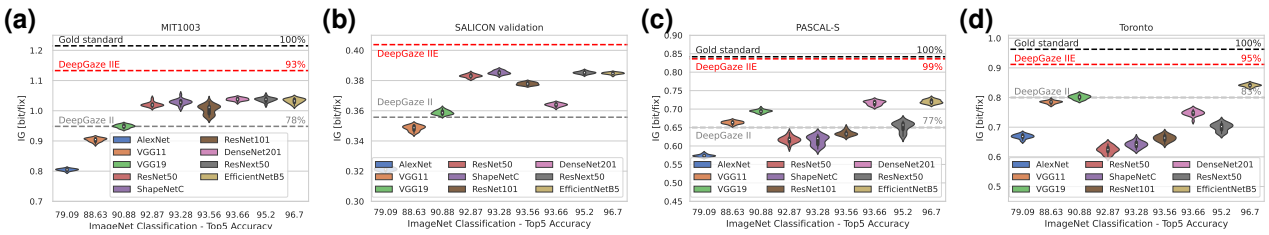

Figure 4. Saliency prediction performance compared to ImageNet accuracy of model backbones. (a) Every violin plot is a representation of the performance distribution of 20 instances that share the same configuration and differ only in the initialization seed of their training. The red dashed line (DeepGaze IIE) indicates the performance of the best model we present in this paper (93%), which averages multiple instances of models with different backbones. The black dashed line (Gold standard = 100%) shows an estimate of the achievable performance by the means of a nonparametric Gaussian KDE model of the fixations. The gray dashed line indicates the performance of the existing DeepGaze II model (78%). (b): Same on the SALICON validation dataset (using the models after the pretraining phase on SALICON). (c) and (d) We evaluate all models of (a) without retraining on the PASCAL-S dataset and the Toronto dataset.

overconfident or underconfident predictions. Confidence calibration has been tested before for deep neural networks on classification tasks like ImageNet [9], and it is known that deep neural networks have a tendency to make overconfident predictions on IID data and even more so on OOD data [12, 31, 11]. Ensembles trained on different augmentation techniques can mitigate this overconfidence to a certain degree [35, 31].

Confidence calibration of classification models is usually tested using the Expected Calibration error, which compares a model's accuracy to its average confidence. If a model is perfectly calibrated, its average confidence matches its accuracy. Fixation prediction can be seen as a high-dimensional classification task where each image pixel constitutes a different class. However, on ImageNet or similar classification tasks, usually only one or very few classes contain most of the probability mass, whereas, in fixation prediction, stochasticity is much higher such that even the most salient pixels have relatively low probability and differences between all pixels are relatively subtle. While this stochasticity makes confidence calibration even more important, accuracy will always be very low and is therefore impractical to be used for an empirical test of calibration. Instead, here we propose an approach which is more suited towards settings with high entropy. First, we sort the pixels of a predicted fixation density by probability and then split them into multiple bins, each of identical probability mass. For example, in Figure 1, the model prediction is split by contour lines into four areas of decreasing size (yellow via green to blue), each of which accumulates 25% of predicted fixation probability. After segmenting the predicted fixation density, we count the empirically measured fixations in each area. If the model is well calibrated, each area should receive the same number of fixations. If the model is overconfident, it would assign a high probability to a region that would receive less than expected fixations, while other regions would receive more than expected fixations. By averaging the number of fixations for each probability quantile, i.e., area, over the full dataset, we can summarize the confidence calibration in a histogram.

4. Experiments and Results

4.1. Transferring ImageNet Features to Saliency in a Principled Way

Our set-up is outlined as follows: first we obtain an architecture trained on ImageNet classification and train a readout network that takes as input a certain number of deep layers whose total number of channels are approximately 2048 (see below for details on the layer selection strategy). These layers are either convolutional or activation layers (ReLU). We use the following networks as backbones for our readout network: AlexNet [22], VGG11 and VGG19 [34], ResNet50 and ResNet101 [10], ShapeNet [8], EfficientNet-B5 [36] and DenseNet [13]. Note that regarding ShapeNet, there are 3 configurations regarding how the model was trained, and we choose the one trained on ImageNet and Stylized-ImageNet, then fine tuned on ImageNet. We will be referring to it as ShapeNet-C.

4.1.1 Selecting Layers From the Backbone

For each network, we conduct two sets of experiments: first we deduce which extracted layers are leading to the best performance (*layer search* stage), then repeat multiple initializations of the exact same configuration to gain a robust metric of final performance (*instance search* stage). Our preliminary results showed that fluctuations appear even between different instances of the same layer configuration and are of the same magnitude as fluctuations between the

top 5 performing layer configurations, indicating that an extensive layer search has marginal value. Thus, we test 10 possible configurations during layer search followed by a training of 20 instances from the top configuration. Given that there appear to be notable fluctuations even between different instances of the same pipeline, evaluating information gain across 20 instances gives us a more robust picture of a model's true performance and an estimate of its epistemic uncertainty.

In general, we find that using approximately 3-4 layers from the ultimate and penultimate layer spaces is ideal while using a single layer consistently results in highly suboptimal performance.

4.1.2 ImageNet accuracy as an indicator of saliency prediction performance

In Figure 4a we show the prediction performances of each backbone on the MIT1003 dataset. For each backbone we show the performance distribution of the 20 trained instances. Our results show that ImageNet performance transfers linearly to saliency up until it reaches a plateau. Specifically, we see a big leap in saliency performance starting from AlexNet and leading up to ResNet-50 which then slows down until it peaks at DenseNet-201, dropping off afterwards. This trend is also visible in all other commonly used saliency metrics (Supplementary Material, Table 1)

4.2. Investigating Model Complementarity

When two distinct models perform almost as well on a dataset, there are two potential assumptions: One, that they are learning the exact same pieces of information and thus achieve similar performance, likely one of them doing it in a slightly better way. Two, the models are doing equally well on the whole dataset but might be achieving that by encoding different and potentially complementary pieces of information.

4.2.1 Mixtures of Fixation Densities

In additional experiments we found substantial variances in per-image performance both for models with different backbones and for model instances using the same backbone but different random seeds (see Supplementary Material). This suggests that not only the different backbones but also the different instances of the models using the same backbones in our experiments encode different information. This finding motivated us to leverage the apparent complementarity of the information our models encode in terms of *inter-model complementarity* (different backbones) and *intra-model complementarity* (different instances within the same model). We average the predicted fixation densities in a pairwise manner across some of our probabilistic models, varying the weights of each predicted density. After conducting this experiment in several pairwise combinations, we find that we consistently get an improvement in performance that peaks when the two models have equal weights (Supplementary Material, Figure 3).

We sought to leverage inter-model complementarity by combining all of our top performing models in a pairwise manner, then triple-wise and finally a quadruple-wise mixture of ShapeNet-C, EfficientNet-B5, ResNext-50, DenseNet-201 (weights being equal in all cases). Model performances consistently improve when adding models until the quadruple mixture achieves top performance (Table 1). Adding ResNet-50 for a total of 5 model mixture decreases performance and therefore we stop at four backbones.

As even within the same backbone, there is a significant variance in per sample performance, we exploit not only inter-model complementarity but also intra-model complementarity. To do that we keep the 4 backbones we found to be best and for each of them average several instances, effectively averaging for 4 models $\times$ 2 instances, then 4 models $\times$ 3 instances etc leading up to 5 instances per model for a total mixture of 20 instances. The split does not change each model's impact on the total average, but rather makes it so each model has a more educated decision by averaging over a greater number of its instances. Leveraging intra-model complementarity, we achieve further boost in performance that saturated at 3 instances per model with a final information gain score of 1.1329 bit/fixation compared to 1.1285 bit/fixation for only one instance per model (Table 2). This best performing model DSREx3 will be called "DeepGaze IIE" in the following ("E" for "ensemble"). In the Supplement, Figure 4, we visualize example predictions of the different models.

4.2.2 Generalization Performance

In Figure 4c and d, we show how well the models with different backbones generalize to the PASCAL-S dataset [29] and the Toronto dataset [2]. It can be seen that not all backbones generalize similarly well. While VGG, DenseNet and EfficientNet show good generalization performance on both datasets, ResNet, ShapeNet and ResNext show substantially worse performance. The DeepGaze IIE ensemble model again shows a substantial performance boost compared to all individual models, with performance close to the gold standard performance (99% on PASCAL-S and 95% on Toronto). Especially on PASCAL-S, the performance gain relative to the best backbone (EfficientNet) is nearly as good as the performance difference between the best and the worst backbone. In Figure 4b, we also show how well the models with different backbones perform on the SALICON validation set (using model weights from pretraining on SALICON). Here, again a very similar pattern can be observed. Since SALICON is a much larger dataset than MIT1003, this provides

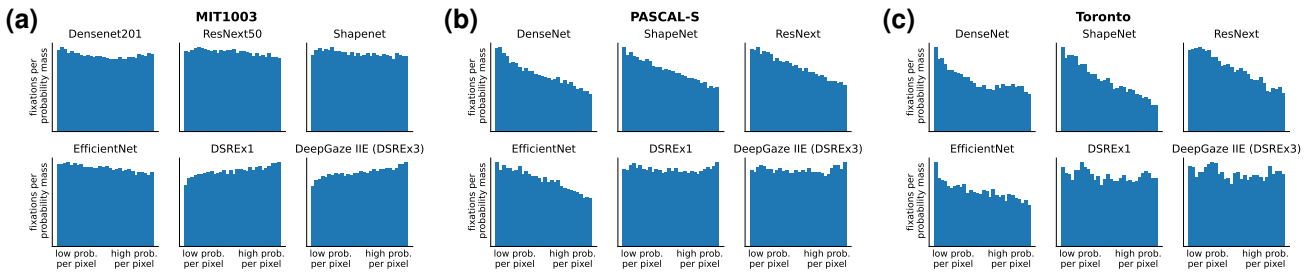

Figure 5. Confidence calibration on different datasets (a: MIT1003, b: PASCAL-S, c: Toronto) for different models (individual histograms). We split predicted fixation densities into multiple quantiles of identical probability mass but sorted by increasing probablity per pixel and quantify the number of actual fixations per predicted probability to assess whether models are overconfident (bar heights decreasing from left to right), underconfident (bar heights increasing from left to right) or well calibrated (uniform histogram). On MIT1003, the dataset used for training, models with individual backbones are quite well calibrated and ensemble models (DSREx1 and DeepGaze IIE=DSREx3) are slightly underconfident. In the generalization setting on PASCAL-S and Toronto, individual models are strongly overconfident while the ensemble models are close to perfectly calibrated.

Table 1. Leveraging *inter*-model complementarity: We mixed our top performing models, starting with pairwise mixtures and leading up to a mixture of four. Note that for illustration purposes, darker shades of red represent more components for the corresponding mixture model.

Backbones	None	DenseNet-201	EfficientNet-B5	DenseNet-201, EfficientNet-B5
None		1.0377	1.0326	1.1077
ResNext-50	1.0368	1.1075	1.1052	1.1256
ShapeNet-C	1.0278	1.1025	1.0986	1.1213
ResNext-50, ShapeNet-C	1.0904	1.1165	1.1143	**1.1285**

additional evidence that DeepGaze IIE is not simply solving an overfitting problem but leverages different information from different backbones.

Finally, we also test our ensemble models on the held-out MIT300 dataset of the MIT/Tuebingen Saliency Benchmark [18, 24]. Our pairwise combination of models is already enough to beat the state of the art, while our final combination of four models with three instances each leads to an even higher leap on the state of the art. The power of ensembling different models with different backbones is further demonstrated by a mixture of the three current top-performing models on MIT300 (UNISAL, EML-Net and MSI-Net), which also outperforms current state-of-the-art, but is still slightly outperformed by DeepGaze IIE. The benchmark results are displayed at table 3. In the Supplement, Table 2, we also report scores on the SALICON test set.

4.2.3 Confidence Calibration

In Figure 5, we visualize confidence calibration for our models (see Section 3.3 for details). Uniform histograms indicate perfect confidence calibration while histograms skewed to the left indicate overconfident models: there are not as many fixations in high-saliency regions as expected by the model. Histograms skewed to the right indicate underconfident models. In Figure 5a, we evaluate confidence calibration for models with four different backbones as well DSREx1 and DeepGaze IIE on the MIT1003 dataset. Evidently, all individual backbones are fairly well calibrated (the histograms are close to uniform), with a slight bias towards overconfidence. The ensemble models DSREx1 and DeepGaze IIE on the other hand are a bit underconfident. When generalizing to the PASCAL-S and Toronto datasets [2, 29] (Figures 5b and c), this effect changes: all individual models are now strongly overconfident, while the ensemble models are close to perfectly calibrated on both datasets. This suggests that individual models make different errors on new images, which are compensated by using an ensemble of models with different backbones. Interestingly, this doesn't hold when we exclusively average models with the same backbone. Apparently, the problem is not noise in the readout network, but overfitting to certain features of the backbone, which happen to be overly correlated with fixations on the MIT1003 dataset. Since ensembling helps, features used by individual models likely differ substantially across backbones.

5. Discussion

Although the models trained for ImageNet classification contain features of high value to saliency prediction, features extracted from ImageNet classification have reached a point of diminishing returns where additional classification

Table 2. Leveraging *intra*-model complementarity: We split the four model mixture (DSRE) into more instances per model and evaluate for each number of instances.

Number of instances	1	2	3	4	5
DSRE	1.1285	1.13193	**1.13294**	1.13285	1.13287

Table 3. Models scores on the MIT300 benchmark. Notably, some models are missing IG as they are not probabilistic and thus impossible to evaluate under this metric. DINet is not included in the public MIT300 leaderboard, therefore we show the scores reported in their paper.

Model	IG ↑	AUC ↑	sAUC ↑	NSS ↑	CC ↑	KLDiv ↓	SIM ↑
DeepGaze IIE (DSREx3)	**1.0715**	**0.8829**	0.7942	**2.5265**	**0.8242**	**0.3474**	0.6993
DSREx1	1.0679	0.8825	0.7938	2.5219	0.8234	0.3489	0.6987
UNISAL+EML-Net+MSI-Net	1.0607	0.8824	**0.7948**	2.5131	0.8239	0.3537	**0.7030**
UNISAL [7]	0.9505	0.8772	0.7840	2.3689	0.7851	0.4149	0.6746
EML-NET [16]		0.8762	0.7469	2.4876	0.7893	0.8439	0.6756
MSI-NET [23]	0.9185	0.8738	0.7787	2.3053	0.7790	0.4232	0.6704
DeepGaze II [28]	0.9247	0.8733	0.7759	2.3371	0.7703	0.4239	0.6636
TranSalNet		0.8730	0.7471	2.3758	0.7991	0.9019	0.6852
GazeGAN [4]		0.8607	0.7316	2.2118	0.7579	1.3390	0.6491
DINet [40]		0.86	0.71	2.33	0.79		

accuracy no longer clearly transfers to higher prediction performance in the saliency domain. However, features from different backbones don't seem to be correlated with saliency in the same way. This is suggested by the fact that models using different backbones generalize in very different ways to new datasets, and even more by the fact that ensemble models substantially outperform even the best individual models both within dataset and on new datasets.

In order to test how useful our models are in practical applications on unseen datasets, we test out-of-domain performance not only with respect to prediction performance, but also with respect to confidence calibration. We find that our individual models tend to be substantially overconfident on out-of-domain data, while our ensemble models are slighlty underconfident on within-domain-data but close to perfectly calibrated on out-of-domain data, which makes them more applicable on unseen datasets. The method which we propose for assessing confidence calibration can be easily applied in settings with a high number of classes and high stochasticity in the ground truth distribution.

With regards to saliency prediction, performance has somewhat stagnated in recent years thus making the observed leap even more significant, especially if we consider that our architectures are not overengineered to the task but rather are part of a principled pipeline that could potentially be applied in other domains. We attribute the success to four factors: First, our choice of readout network, which is less constrained than a linear readout allowing it to make nonlinear transformations of input features but more constrained than a typical CNN as it uses only 1×1 kernels. This allows it to combine the spatial features without creating new ones making it an efficient tool for transfer learning and allowing interpretability in its results, since we don't finetune the backbones. While finetuning the backbone in theory could result in even better performance, we found that by fine tuning the large parameter space we inevitably overfit MIT1003 and consistently produces worse results. The second factor is our utilization of multiple instances of each model. We argue that this is good practice as it models the uncertainty in these models which in some cases such as ResNet101 is much higher than one would expect. Third, we leverage multiple models and combine them in a principled way utilizing both the complementarity between architectures and between instances of the same architectures which we labeled inter- and intra-complementarity respectively. For saliency this sort of combination is really simple, not requiring an oracle network but rather a simple averaging process of the fixation densities. Fourth, we used information gain to guide our experiments and have highlighted how relative performance transfers reliably to other metrics and other datasets. It has been argued that information gain is ideal for principled studies due to its foundation in information theory and its independence of hyperparameters [26]. In the future, maximally diverse backbones should be further explored to yield even better models. This could be done through correlation analysis or by combining ImageNet backbones such as the ones presented here with self-supervised ones, as well as backbones pre-trained on other tasks such as object detection.

Taken together we have shown that our principled ensemble learning approach yields a 15 percent point improvement over DeepGaze II, setting the new state of the art in saliency prediction on the MIT/Tuebingen Saliency Benchmark in all available metrics, a significant leap after 4 years of only gradual progress, highlighting the promise of our approach.

References

[1] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

[2] Neil Bruce and John Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007.

[3] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[4] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, 29:2287–2300, 2019.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[7] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. *arXiv preprint arXiv:2003.05477*, 2020.

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. ISSN: 2640-3498.

[10] 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.

[11] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *arXiv:1812.05720 [cs, stat]*, 2019.

[12] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[14] 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*, pages 262–270, 2015.

[15] 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*, 20(11):1254–1259, 1998.

[16] Sen Jia and Neil D.B. Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020.

[17] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, June 2015.

[18] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[19] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[20] Wolf Kienzle, Felix A Wichmann, Matthias O Franz, and Bernhard Schölkopf. A nonparametric approach to bottom-up visual saliency. In *Advances in neural information processing systems*, pages 689–696, 2007.

[21] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[23] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 2020.

[24] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/, 2019.

[25] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[26] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015.

[27] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing, 2018.

[28] Matthias Kümmerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding Low- and High-Level Contributions to Fixation Prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 4789–4798, 2017.

[29] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 280–287, 2014.

[30] Ali Mahdi, Jun Qin, and Garth Crosby. Deepfeat: A bottom-up and top-down saliency model based on deep features of convolutional neural networks. *IEEE Transactions on Cognitive and Developmental Systems*, 12(1):54–63, 2019.

[31] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] 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 preprint arXiv:1701.01081*, 2017.

[33] KH Ruddock, DS Wooding, and SK Mannan. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial vision*, 10(3):165–188, 1996.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Asa Cooper Stickland and Iain Murray. Diverse ensembles improve calibration. *arXiv:2007.04206 [cs, stat]*, 2020.

[36] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[37] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[38] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[39] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014.

[40] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A Dilated Inception Network for Visual Saliency Prediction. *arXiv:1904.03571 [cs]*, May 2019.

[41] Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

[42] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008.