n-Reference Transfer Learning for Saliency Prediction

Yan Luo[1](✉), Yongkang Wong[2], Mohan S. Kankanhalli[2], and Qi Zhao[1]

[1] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, USA
luoxx648@umn.edu, qzhao@cs.umn.edu

[2] School of Computing, National University of Singapore, Singapore, Singapore
{wongyk,mohan}@comp.nus.edu.sg

Abstract. Benefiting from deep learning research and large-scale datasets, saliency prediction has achieved significant success in the past decade. However, it still remains challenging to predict saliency maps on images in new domains that lack sufficient data for data-hungry models. To solve this problem, we propose a few-shot transfer learning paradigm for saliency prediction, which enables efficient transfer of knowledge learned from the existing large-scale saliency datasets to a target domain with limited labeled samples. Specifically, few target domain samples are used as the *reference* to train a model with a source domain dataset such that the training process can converge to a local minimum in favor of the target domain. Then, the learned model is further fine-tuned with the *reference*. The proposed framework is gradient-based and model-agnostic. We conduct comprehensive experiments and ablation study on various source domain and target domain pairs. The results show that the proposed framework achieves a significant performance improvement. The code is publicly available at https://github.com/luoyan407/n-reference.

Keywords: Deep learning · Saliency prediction · n-shot transfer learning

1 Introduction

Saliency prediction is the task that aims to model human attention to predict where people look in the given image. Thanks to the power of deep neural networks [15,24,48] (DNNs), state-of-the-art saliency models [7,50] perform very well in predicting human attention on naturalistic images. Behind the success of this task, a considerable amount of real-world images and corresponding human fixations fuels the process of training the data-hungry DNNs.

However, it is still difficult to predict saliency maps on images in novel domains, which has insufficient or few data to train saliency models with desired

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58598-3_30) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12353, pp. 502–519, 2020.
https://doi.org/10.1007/978-3-030-58598-3_30

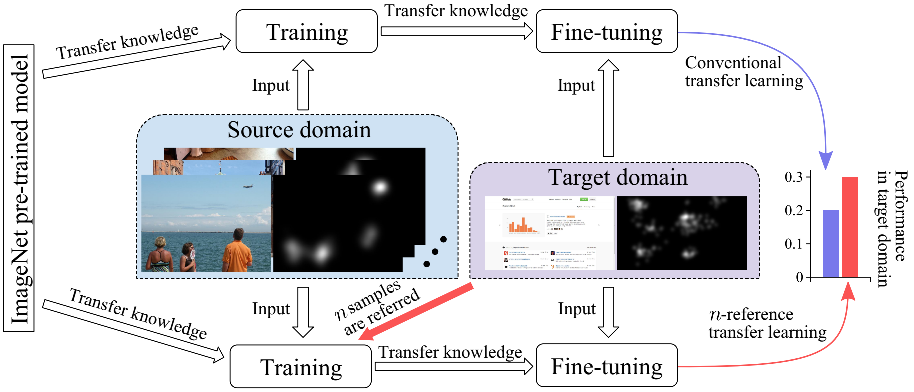


Fig. 1. The proposed n-reference transfer learning framework for saliency prediction. This framework aims to generate a better initialization with n reference samples from the target domain when training on the source domain, followed by fine-tuning to maximize knowledge transfer. It is based on the widely-used two-stage transfer learning framework (i.e., first training and then fine-tuning) and can easily adapt to other fine-tuning strategies

performance. As the time/money cost of collecting human fixations is prohibitive [3,20], a feasible solution is to reuse the existing large-scale saliency datasets along with a few target domain samples to solve this problem. Along this line, we study how to transfer the knowledge learned from the existing large-scale saliency datasets to the target domain in a few-shot transfer learning setting.

The necessity of few-shot transfer learning for saliency prediction lies in the nature of the task. Based on findings drawn from the behavioral experiments, the way that humans attend to regions is significantly affected by the scene context [35,45,47]. The scene context is correlated to the image domain [43]. In other words, each image from a specific domain could be representative of the others from the same domain to some degree, e.g., webpage images generally have a similar layout and design [41]. In visual saliency study, existing datasets [3,42] in non-natural images domain are much smaller than the natural image ones [20, 22]. Moreover, there are numerous images used in the subfields of medicine, biology, etc., which may not have any human fixation data yet. In this work, we assume that it is feasible and viable to collect human fixations on a small number of images to enable few-shot learning.

Compared to n-reference transfer learning for classification task [1], we focus on how to use very few target domain samples as references to learn a better initial model for fine-tuning. Moreover, there exists no such works for saliency prediction task. Models designed for classification may not work for saliency prediction. First, visual samples in existing classification tasks often contain limited visual concepts (i.e., pre-defined object classes), while objects of any class may appear in the images used for saliency prediction. In this sense, saliency prediction often handles images with higher diversity than the ones used for classification. Second, the output of classification models [1,15,24,32,44] is a discrete label, while saliency models [7,50] output a matrix of real numbers.

In this work, we follow the widely-used two-stage transfer learning framework [1,13,41], i.e., first training and then fine-tuning, and propose a n-reference transfer learning framework. Specifically, in the training stage, it aims to use a small number of samples in the target domain as references to guide the knowledge learned from the source domain dataset. In this way, the learned model is adapted to the target domain and can be seen as a better initialization than the one trained without the references. The small number of target domain samples are used as references in both the training stage and as the training data in the fine-tuning stage. The proposed framework is shown in Fig. 1.

Mathematically, we use cosine similarity between two gradients to facilitate the reference aware model training, where the two gradients are respectively computed by samples in the source and target domain. If the angle between the two gradients is greater than 90 degrees, which implies that the directions of the model update are significantly different from each other, we optimize the gradient for the update to have smaller differences with the target-domain referenced gradient in cosine similarity. The intuition behind is to mimic the process of human learning with the reference sample, i.e., we adaptively learn from new information so that the newly absorbed knowledge will not contradict the observation of the reference samples [31,33]. The proposed framework is gradient-based and it is model-agnostic.

To comprehensively evaluate the proposed framework, we employ SALICON [20] and MIT1003 [22] as the source domain datasets (i.e., the knowledge sets), and WebSal [42] and the art subset in CAT2000 [3] as the target domain data. We randomly select 1, 5, or 10 samples from the target domain data as references. The contributions of this work can be summarized as follows:

- To study how humans perceive scenes from a partially explored domain, we propose a model-agnostic few-shot transfer learning paradigm to transfer knowledge from the source domain to the target domain. This is the first work that studies few-shot transfer learning for saliency prediction.
- We propose a n-reference transfer learning framework to adaptively guide the training process. It guarantees that the knowledge learned with the source domain data would not contradict the references in the target domain, and produce a good initialization for further fine-tuning. The proposed framework is model-agnostic and can generally work with existing saliency models.
- Comprehensive experiments show the proposed framework works on various combinations of source domain and target domain pairs. The experiment with various baseline models show that the proposed approach can efficiently transfer the knowledge from the source domain to the target domain.

2 Related Works

2.1 Saliency Prediction

Saliency prediction aims to mimic human vision system to perceive interesting regions in a cluttered visual world. Itti et al. [19] develop the first bottom-up stimulus-driven saliency model. Since then, many works emerge to interpret visual saliency from various perspectives [14,16,22,52]. With the advent of DNNs [15,24,48], saliency prediction benefitted from data-driven discriminative features instead of relying on hand-crafted features [6,25,26,37]. Recently, Cornia et al. [7] introduce a network that integrates ResNet-50 [15] and convolutional LSTMs to better attend to salient regions by iteratively refining the predictions. Yang et al. [50] propose a dilated inception network (DINet) that stacks dilated convolutions with different dilation rates upon ResNet-50 to capture wider spatial information. It achieves state-of-the-art performance on various benchmarks. A widely-used practice to transfer the knowledge learned from image classification to saliency prediction is by using the weights pre-trained on ImageNet as model initialization [6,25,26,37]. In contrast, this work studies the few-shot cross-domain transfer learning problem, which takes place between two domains. Without loss of generality, we follow [33] to adopt both ResNet-50 and DINet as the baseline models in this work.

2.2 Few-Shot Learning

Few-shot learning [11,27,32,44] aims to study how to learn classifiers for unseen visual concepts with only a few samples per class. Lake et al. [28] introduce a Bayesian program learning framework that can learn from one example for predicting character strokes. Matching networks [46] use an attention mechanism that is analogous to a kernel density estimator so that it can learn from a few examples rapidly. Sung et al. [44] propose a relation network to learn a transferable deep metric to compare the relation between the small number samples. In [29], Lee et al. study how to learn feature embeddings with a few samples that can minimize generalization error across a distribution of tasks. As the process of collecting human fixations is prohibitive [20], learning with very few samples is promising for saliency prediction to overcome the need for big data.

2.3 Transfer Learning

Transfer learning, a.k.a. domain adaptation or domain transfer, is a paradigm to utilize training data in the source domain to solve the problem in the target domain [8,9,30,38,41]. In general, it can be seen as a two-stage learning framework, i.e., first training a model with source domain data and then fine-tuning the pre-trained model with target domain data. There are many DNN-based works [1,2,12,13,49] that use this learning framework for classification tasks. Specifically, Guo et al. [13] study and design a variant of the standard fine-tuning method for better transferability. However, it requires many training samples to determine whether it should fine-tune or freeze the parameters in a particular layer. Recently, Bäuml and Tulbure [1] introduce a learning framework that transfers the knowledge learned from the source domain to the target domain with a few samples for tactile material classification. As saliency prediction is by nature class-agnostic, learning to predict human fixations with very few samples (e.g., $\leq$10) in the target domain is more challenging than the same paradigm for classification and has not been explored yet. Different from the aforementioned methods, we propose the first model-agnostic few-shot transfer learning framework for saliency prediction and conduct comprehensive study on multiple combinations of source domain datasets and target domain datasets.

3 Methodology

In this section, we first formulate the problem and discuss its theoretical generalization bound. Then, we delve into the details of the proposed framework.

3.1 Problem Statement

In this work, we denote the images as $I^S, I^T \in \mathbb{R}^m$ and the human fixation maps as $y^S, y^T \in \mathcal{Y}$ ($\mathcal{Y} \equiv [0,1]^m \subseteq \mathbb{R}^m$), where m is the dimensions of the image and S (T) indicates the source (target) domain. In general, given an image I, the prediction function $f : \mathbb{R}^m \xrightarrow{\theta} \mathcal{Y}$ with parameters θ will predict z and then the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ will evaluate the discrepancy between z and y. Transfer learning for saliency prediction task can be considered as a two-stage learning problem. First, the model's parameters are learned with the source domain data through the training process, i.e.,

$$\theta_{\mathsf{TR}} = \arg\min_{\theta} \frac{1}{|D^S|} \sum_{(I_i, y_i) \in D^S} \ell(f(I_i; \theta), y_i)|_{\theta_0} \tag{1}$$

where D^S is the source domain dataset, $|D^S|$ is the number of the samples, and TR stands for training. θ_0 are the initialized parameters and the model is usually pre-trained on ImageNet [10]. Then, θ_{TR} is taken as the initialization for further fine-tuning on the target domain data, i.e.,

$$\theta^*_{\mathsf{FT}} = \arg\min_{\theta} \frac{1}{|D^T|} \sum_{(I_i, y_i) \in D^T} \ell(f(I_i; \theta), y_i)|_{\theta_0 = \theta_{\mathsf{TR}}} \tag{2}$$

In this work, we aim to learn a better initialization by the first stage objective (1), which is in favor of the target domain data. Such initialized parameters (i.e., θ_{TR}) are expected to further achieve better performance by fine-tuning on D^T. To this end, we introduce a referencing mechanism that allows the training process fed with D^S to reference the model update w.r.t. the referenced samples $(I^R, y^R) \in D^T (|D^S| \gg |D^T|)$. Mathematically, this can be formulated as

$$\theta_{\mathsf{TR-Ref}} = \arg\min_{\theta} \frac{1}{|D^S|} \sum_{\substack{(I_i, y_i) \in D^S \\ (I_j^R, y_j^R) \in D^T}} \ell(f_{\mathsf{Ref}}(I_i; \theta, (I_j^R, y_j^R)), y_i)|_{\theta_0} \tag{3}$$

where $\mathsf{TR-Ref}$ indicates the training process references target domain samples when updating the model. f_{Ref} is a variant of f which has the same forward propagation as f but has more complicated backward propagation. $\theta_{\mathsf{TR-Ref}}$ is taken as the initialization in the second stage objective (2) for further fine-tuning. We denote the resulting parameters as $\theta_{\mathsf{FT|Ref}}$.

3.2 Generalization Bound of Saliency Prediction

Here, we discuss the theoretical guarantee of saliency prediction. Following the setting used in [34], given training data $(I_1, y_1), (I_2, y_2), \ldots \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} \in [0,1]^m \subseteq \mathbb{R}^m$, we use the L^p loss, i.e., $\ell^p : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+, p \geq 1$. The prediction function $f(\cdot;\theta)$ is denoted as $f(\cdot)$ for simplicity. I is drawn i.i.d. according to the unknown distribution $\mathcal{D}$ and $y = f^*(I)$ where f^* is the target labeling function. Saliency prediction can be considered as a mathematical problem that finds hypothesis $f : \mathbb{R}^m \to [0,1]^m$ in a set H with small generalization error w.r.t. f^*,

$$R_{\mathcal{D}}(f) = E_{I \sim \mathcal{D}}[\ell(f(I), f^*(I))].$$

In practice, as $\mathcal{D}$ is unknown, we use empirical error for approximation, i.e.,

$$\hat{R}_{\mathcal{D}}(f) = \frac{1}{|D|} \sum_{i=1}^{|D|} \ell(f(I_i), y_i),$$

where $|D|$ is the sample number in dataset D for training.

We introduce the generalization bound of saliency prediction as follows. The proof is provided in the supplementary document.

Theorem 1 (Saliency generalization bound). *Denote H as a finite hypothesis set. Given ℓ^p and $y \in [0,1]^m$, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $f \in H$:*

$$|R_{\mathcal{D}}(f) - \hat{R}_{\mathcal{D}}(f)| \leq m^{\frac{1}{p}} \sqrt{\frac{\log|H| + \log\frac{2}{\delta}}{2|D|}}$$

Remark 1. Theorem 1 shows how the training set scale influences the generalization bound. When $|D|$ tends towards infinity, $R_{\mathcal{D}}(f) \equiv \hat{R}_{\mathcal{D}}(f)$. This conforms to the general intuition that it can train a more general model with more data. Contrarily, when $|D| = 1$, it leads to the largest bound for $|R_{\mathcal{D}}(f) - \hat{R}_{\mathcal{D}}(f)|$. Moreover, it demonstrates the task is challenging with small number of samples.

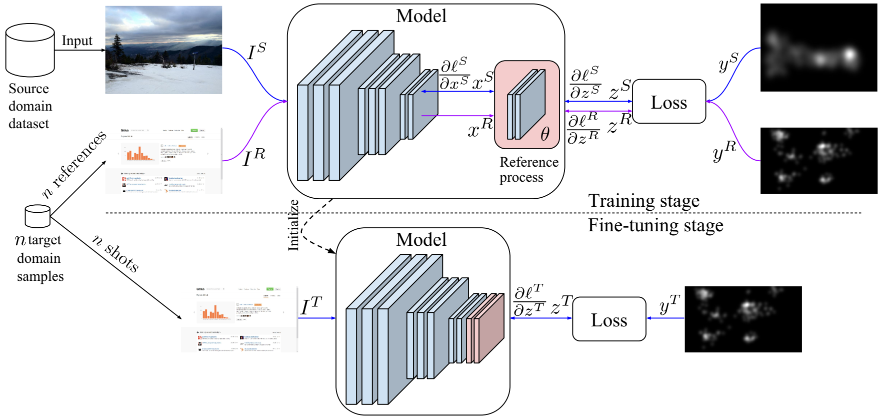

Input
Source domain dataset
I^S
I^R
Model
$\frac{\partial \ell^S}{\partial x^S}$ x^S
x^R
θ
Reference process
$\frac{\partial \ell^S}{\partial z^S}$ z^S
$\frac{\partial \ell^R}{\partial z^R}$ z^R
Loss
y^S
y^R
n references
n target domain samples
n shots
Initialize
Training stage
Fine-tuning stage
Model
I^T
$\frac{\partial \ell^T}{\partial z^T}$ z^T
Loss
y^T

Fig. 2. Proposed n-reference transfer learning framework. Note that we assume that only very few samples from the target domain are available, i.e., $n \leq 10$

3.3 Overall Framework

In this subsection, we introduce the few-shot transfer learning framework that solves the objective function (2) and (3). The overall workflow of the proposed n-reference transfer learning framework is shown in Fig. 2.

Similar to classification model [15,17,48], state-of-the-art saliency models tend to be large. For example, DINet [50] and SAM-ResNet-50 [7] consist of 26M and 70M parameters, respectively. Therefore, instead of inefficiently applying the proposed framework to the whole saliency model, we only apply it to a few downstream layers which are close to the output. The downstream layers produce discriminative features used for prediction with a small number of parameters, and it makes the transfer learning process more cost-effective. Consequently, we split the model into two parts, i.e., the model body $\theta_{\textsf{body}}$ and the model head $\theta_{\textsf{head}}$. This split would be only effective in the training stage and the two parts will be integrated again as they always are in the inference stage. Note that the split is flexible. The effective scope of the proposed framework could cover the whole model and the model body would correspondingly turn to be an empty set. As we only focus on $\theta_{\textsf{head}}$, we simplify it as θ in the following text.

In the forward propagation, as the training image $I^S \in D^S$ and the reference image $I^R \in D_T$ are fed to the model body, the discriminative feature x^S and x^R are generated, respectively. Then, the model head would take x^S and x^R as input to produce prediction z^S and z^R, respectively. Specifically, $z^S = f(x^S; \theta)$. A similar process applies to z^R. The loss function is used to compute the distance between z^S and y^S (and between y^R and y^R as well). In the backward propagation, two gradients are computed by the chain rule

$$\frac{\partial \ell^S}{\partial \theta} = \frac{\partial \ell(f(x^S;\theta), y^S)}{\partial z^S} \frac{\partial z^S}{\partial \theta}, \qquad \frac{\partial \ell^R}{\partial \theta} = \frac{\partial \ell(f(x^R;\theta), y^R)}{\partial z^R} \frac{\partial z^R}{\partial \theta}.$$

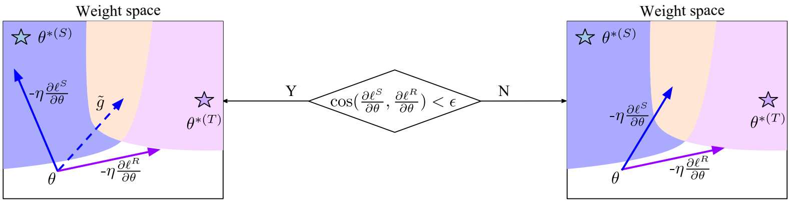

Weight space
$\theta^{*(S)}$
$-\eta\frac{\partial \ell^S}{\partial \theta}$
$\tilde{g}$
$\theta^{*(T)}$
θ
$-\eta\frac{\partial \ell^R}{\partial \theta}$
Y
$\cos(\frac{\partial \ell^S}{\partial \theta}, \frac{\partial \ell^R}{\partial \theta}) < \epsilon$
N
Weight space
$\theta^{*(S)}$
$-\eta\frac{\partial \ell^S}{\partial \theta}$
$\theta^{*(T)}$
θ
$-\eta\frac{\partial \ell^R}{\partial \theta}$

Fig. 3. The reference process computing the gradient that better adapts to the target domain data. $\theta^{*(S)}$ is a local minimum trained by sufficient source domain samples, while $\theta^{*(T)}$ is a local minimum trained by sufficient target domain samples. Given a pre-defined threshold ϵ, if the cosine similarity between the gradient ($\frac{\partial \ell^S}{\partial \theta}$) generated by the source sample and the gradient ($\frac{\partial \ell^R}{\partial \theta}$) generated by the reference sample is smaller than ϵ, it will compute a corrected gradient by optimizing the cosine similarity. It retains $\frac{\partial \ell^S}{\partial \theta}$ otherwise

Specifically, $\frac{\partial \ell^S}{\partial \theta}$ indicates the model update towards a local minimum $\theta^{*(S)}$ which is learned from the samples from D^S, while $\frac{\partial \ell^R}{\partial \theta}$ indicates the model update towards a local minimum $\theta^{*(T)}$ which is learned from the samples from D^T.

As shown in Fig. 2, θ_{head} are updated by the proposed reference process and θ_{body} are updated with the standard gradients in the training stage. During fine-tuning, θ_{head} and θ_{body} are updated with the standard gradients.

3.4 Reference Process

Here, we delve into the formulation of the proposed reference process (Fig. 3). The cosine similarity between $\frac{\partial \ell^S}{\partial \theta}$ and $\frac{\partial \ell^R}{\partial \theta}$ can evaluate the difference of the two gradients. Accordingly, we pre-define a threshold ϵ to determine if the difference is considered as minor and the update with $\frac{\partial \ell^S}{\partial \theta}$ will be close to both $\theta^{*(S)}$ and $\theta^{*(T)}$. If the difference is significant, the proposed reference process will adjust $\frac{\partial \ell^S}{\partial \theta}$ so that it will move more towards $\theta^{*(T)}$. This process is defined as follows

$$\tilde{g} = \begin{cases} \arg\max_g \cos(g, \frac{\partial \ell^R}{\partial \theta}) - \lambda \|g\|_2^2 \big|_{g_0=\frac{\partial \ell^S}{\partial \theta}} & \text{if} \cos(\frac{\partial \ell^S}{\partial \theta}, \frac{\partial \ell^R}{\partial \theta}) < \epsilon, \\ \frac{\partial \ell^S}{\partial \theta} & \text{otherwise}, \end{cases} \tag{4}$$

where λ is the regularization parameter and $\cos(\cdot, \cdot)$ is the cosine similarity, i.e., $\cos(a, b) = a^\top b/|a||b|$ (a and b are the input vectors), and $\tilde{g}$ is the output gradient. The embedded optimization problem in Eq. (4) aims to find a $\tilde{g}$, which is with an initial point $g_0 = \frac{\partial \ell^S}{\partial \theta}$, to be consistent with the reference gradient $\frac{\partial \ell^R}{\partial \theta}$ in terms of cosine similarity. In other words, the reference gradient $\frac{\partial \ell^R}{\partial \theta}$ provides a reference so that $\tilde{g}$ is able to be aware of a rough direction towards the underlying $\theta^{*(T)}$. In this way, the knowledge learned from D^S is transferred to the target domain. We solve the embedded optimization problem with the gradient ascent method because our goal is to maximize the cosine similarity between $\tilde{g}$ and $\frac{\partial \ell^R}{\partial \theta}$.

Subsequently, θ would be updated with $\tilde{g}$, i.e., $\theta \leftarrow \theta - \eta\tilde{g}$, where η is a learning rate. Note that $\frac{\partial \ell^S}{\partial \theta}$ is generated by randomly selected training samples and is the initial point for $\tilde{g}$. As a result, the process of optimizing cosine similarity in the training stage is almost surely stochastic. This can effectively prevent $\tilde{g}$ from overfitting $\frac{\partial \ell^R}{\partial \theta}$.

The proposed reference process yields $\tilde{g}$ to update the model so that the parameters are close to the underlying $\theta^{*(T)}$. As θ is learned with the references from the target domain, by the chain rule, $\frac{\partial \ell^S}{\partial \theta_{body}} = \frac{\partial \ell^S}{\partial x^S}\frac{\partial x^S}{\partial \theta_{body}}$ and $\frac{\partial \ell^S}{\partial x^S}$ can be considered as a function of θ. So θ_{b} will be affected by the references as well.

As the number of references is expected to be far smaller than the training data, we follow a similar idea of the stochastic process to randomly draw a reference from the reference pool at each iteration.

4 Experiments

In this section, we introduce the experimental protocol, present the experimental results, and then have a discussion about the results.

4.1 Experimental Setup

Datasets. We adopt the large-scale saliency prediction dataset SALICON [20] (the 2017 version) and the MIT1003 [22] as the source domain datasets. Accordingly, we adopt WebSal [42] and the art subset in CAT2000 [3] as the target domain datasets. Specifically, SALICON consists of 10000 real-world images, MIT1003 consists of 1003 natural scene images, and WebSal consists of 149 webpage screenshots. CAT2000 includes 20 categories and each category has 100 images. Art is one of the most common categories, whose images are the pictures of human-made works, like the paintings, handcrafts, and etc.

Baseline Models. To study how well the proposed method would generalize to different models, we use two baseline models, i.e., DINet [51] and ResNet-50 [15].

Settings. There are three dimensions to the experiments in this work, i.e., source domain samples, baseline model, and target domain samples. Specifically, the baseline model is trained with the source domain samples. The learned model is further fine-tuned with the target domain samples. This setting is similar in the case of the proposed method. For convenience, we denote the setting as a combination of the initials of the datasets or the models, e.g., $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$ indicates that we use SALICON as the source domain dataset, DINet as the baseline model, and WebSal as the target domain dataset. Similarly, we use initials M, R, and A to represent MIT1003, ResNet-50, and Art, respectively.

To understand how the number of references affects the performance, we evaluate the proposed method with $n = 1, 5, 10$. Moreover, to provide a benchmark of the performance w.r.t. more references, a paradigm that is similar to 3-fold cross validation is applied with more references. For instance, given WebSal as the target domain datasets, we divide it into three subsets, which contain 50, 50, and 49 images, respectively. Then, we alternately use any two subsets as the reference samples and the rest as the validation set. The process is repeated 3 times. We denote the results of this process as an empirical upper bound.

Evaluation Metrics. We adopt the common metrics used in [5] and [20], i.e., normalized scanpath saliency (NSS) [18,40], area under curve (AUC) [4,21], and correlation coefficient (CC) [36]. Higher scores indicate better performance. We use the public implementation[1] provided by [20]. Each experiment is repeated 10 times and the mean metric scores are reported. Due to the space limits, we report the corresponding standard deviation in the supplementary document.

4.2 Training Scheme

We follow the widely-used two-stage transfer learning framework [1,13,41,49], i.e., first train a model with the source domain data and fine-tune with the target domain data. We denote the trained model as TR and the fine-tuned model as FT. In the proposed framework, the n-reference training stage first trains a model with the source domain data and n target domain references (denoted as $\mathsf{TR}-\mathsf{Ref}$), and then further fine-tune with the references (denoted as $\mathsf{FT}|\mathsf{Ref}$).

Regarding the experimental details, we follow DINet [50] to use Adam optimizer [23] with learning rate $\eta = 5\text{e}-5$ and weight decay $1\text{e}-4$. We use batch size 10 for all the experiments. The number of epochs is 10 and we decrease the learning rate for every 3 epochs by multiplying with 0.2. In $\mathsf{TR}-\mathsf{Ref}$, we randomly sample 10 training data without replacement as the training sample at each iteration. Meanwhile, we randomly sample n_r references with replacement as the reference. In this way, the difference between the number of training samples and references will not cause a problem. n_r are 1, 3 and 5 in the experiments with $n = 1, 5, 10$, respectively. This process is the same for the one of FT. We select the model with the best performance over epochs for further fine-tuning. The normalized l_1 loss [50] is used and the threshold ϵ is set to 0 for all the experiments. We implement the proposed framework with PyTorch [39].

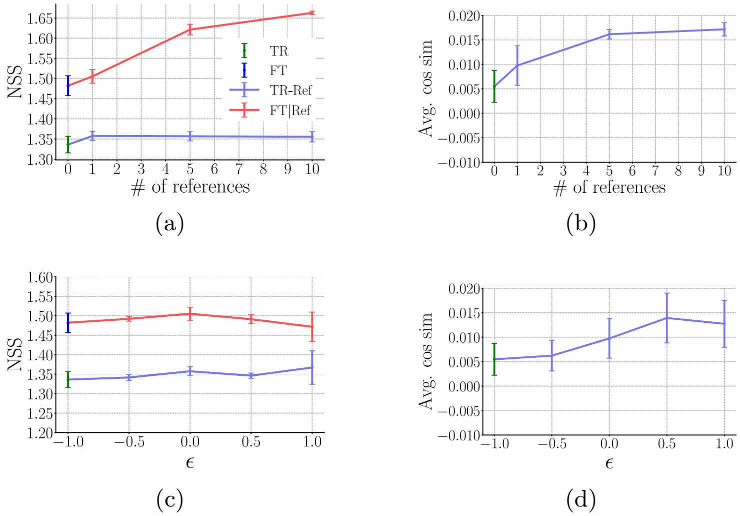
4.3 Performance

The experimental results with the following settings, i.e., $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$, $\langle \mathsf{S}, \mathsf{R}, \mathsf{W} \rangle$, $\langle \mathsf{M}, \mathsf{D}, \mathsf{W} \rangle$, and $\langle \mathsf{S}, \mathsf{D}, \mathsf{A} \rangle$, are shown in Table 1. Within setting $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$, the proposed framework (i.e., $\mathsf{FT}|\mathsf{Ref}$) achieves better performance than FT over all metrics. Particularly, as the number of references increases, the consequently trained models provide better initializations for fine-tuning. In other words, $\mathsf{FT}|\mathsf{Ref}$ yields better performance when the dependent trained model uses more reference samples. Using a different baseline model, we experiment it with setting $\langle \mathsf{S}, \mathsf{R}, \mathsf{W} \rangle$

[1] https://github.com/NUS-VIP/salicon-evaluation.

Table 1. Performance with various settings of $\langle \mathsf{source}, \mathsf{model}, \mathsf{target} \rangle$. Here, S is SALICON, M is MIT1003, W is WebSal, A is Art subset, D is DINet, and R is ResNet. ↑ implies that a higher score is better. The score in bold font indicates the best result under the respective metric. We report the mean score from 10 runs for conventional training (i.e., $n = 0$) and the proposed method. The empirical upper bound (EUB) is generated by 3-fold cross validation on the target domain. The experimental details are provided in Sect. 4.1 and 4.2

		$\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$			$\langle \mathsf{S}, \mathsf{R}, \mathsf{W} \rangle$		
		NSS↑	AUC↑	CC↑	NSS↑	AUC↑	CC↑
FT w/o TR	$n = 10$	0.8252	0.7430	0.3635	0.8846	0.7455	0.3852
TR	$n = 0$	1.3330	0.7796	0.5515	1.2950	0.7749	0.5358
TR−Ref	$n = 1$	1.3621	0.7848	0.5628	1.3569	0.7864	0.5611
FT	$n = 1$	1.4731	0.8005	0.5976	1.3722	0.7923	0.5627
FT|Ref	$n = 1$	**1.5077**	**0.8051**	**0.6121**	**1.4272**	**0.7983**	**0.5817**
TR−Ref	$n = 5$	1.3683	0.7874	0.5659	1.3535	0.7837	0.5593
FT	$n = 5$	1.5803	0.8161	0.6355	1.5043	0.8131	0.6139
FT|Ref	$n = 5$	**1.6085**	**0.8200**	**0.6468**	**1.5491**	**0.8149**	**0.6281**
TR−Ref	$n = 10$	1.3647	0.7839	0.5633	1.3583	0.7857	0.5612
FT	$n = 10$	1.6290	0.8247	0.6531	1.5164	0.8103	0.6200
FT|Ref	$n = 10$	**1.6439**	**0.8276**	**0.6605**	**1.5829**	**0.8143**	**0.6414**
TR−Ref	EUB	1.3822	0.7910	0.5708	1.3626	0.7864	0.5645
FT	EUB	1.8695	0.8488	0.7389	1.8325	0.8462	0.7275
FT|Ref	EUB	**1.8831**	**0.8494**	**0.7442**	**1.8500**	**0.8480**	**0.7321**
		$\langle \mathsf{M}, \mathsf{D}, \mathsf{W} \rangle$			$\langle \mathsf{S}, \mathsf{D}, \mathsf{A} \rangle$		
		NSS↑	AUC↑	CC↑	NSS↑	AUC↑	CC↑
FT w/o TR	$n = 10$	0.8252	0.7430	0.3635	1.2183	0.8339	0.5161
TR	$n = 0$	1.3905	0.7991	0.5700	1.5172	0.8225	0.6003
TR−Ref	$n = 1$	1.4405	0.8085	0.5902	1.5651	0.8287	0.6211
FT	$n = 1$	1.4410	0.8023	0.5784	1.6255	0.8324	0.6449
FT|Ref	$n = 1$	**1.4575**	**0.8070**	**0.5838**	**1.6523**	**0.8380**	**0.6564**
TR−Ref	$n = 5$	1.4452	0.8064	0.5908	1.5870	0.8304	0.6274
FT	$n = 5$	1.5795	0.8217	0.6395	1.8049	0.8480	0.7185
FT|Ref	$n = 5$	**1.6136**	**0.8269**	**0.6515**	**1.8314**	**0.8503**	**0.7274**
TR−Ref	$n = 10$	1.4330	0.8060	0.5872	1.5704	0.8288	0.6204
FT	$n = 10$	1.6462	0.8261	0.6660	1.8325	0.8474	0.7288
FT|Ref	$n = 10$	**1.6691**	**0.8283**	**0.6730**	**1.8584**	**0.8503**	**0.7366**
TR−Ref	EUB	1.4402	0.8087	0.5905	1.5980	0.8340	0.6331
FT	EUB	1.8450	0.8466	0.7330	2.1595	0.8636	0.8464
FT|Ref	EUB	**1.8507**	**0.8478**	**0.7344**	**2.1874**	**0.8649**	**0.8519**



(a)

(b)

(c)

(d)

Fig. 4. The effect of the number of references (a, b) and threshold ϵ (c, d) on NSS metric and average cosine similarity within setting $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$. $n = 0$ indicates that no reference sample is used. Hence, TR−Ref and FT|Ref turn to be TR and FT. The results of c and d are generated with $n = 1$. ϵ determines whether the gradient needs to be corrected or not (see Fig. 3). Comparing to TR, FT, and FT|Ref, only TR−Ref is able to evaluate the cosine similarity between the samples from the source domain and target domain (see Fig. 2)

which FT|Ref achieves consistent improvement. Moreover, using DINet as the baseline model leads to better performance than using ResNet-50.

We study how well the proposed framework generalizes to different target domain data using setting $\langle \mathsf{S}, \mathsf{D}, \mathsf{A} \rangle$. As seen in Table 1, similar performance improvement can be found, which implies the proposed framework can generalize to a different target domain. Furthermore, the study with MIT1003 as the source domain dataset, i.e., setting $\langle \mathsf{M}, \mathsf{D}, \mathsf{W} \rangle$, shows consistent improvement. The overall performance within setting $\langle \mathsf{M}, \mathsf{D}, \mathsf{W} \rangle$ is slightly lower than the one within setting $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$. This implies that SALICON is more efficient than MIT1003 to transfer the knowledge to WebSal. On the other hand, models trained with one sample in target domain have noticeable gaps w.r.t. EUB, and is improved with more training samples. This is consistent with the implication of Theorem 1.

We perform paired t-test and permutation test over images within setting $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$ to evaluate the difference between TR−Ref and FT|Ref. Both corresponding p are less than 0.001. This implies that TR−Ref significantly provides a good initialization to FT|Ref to yield high performance. To validate the effect of knowledge transfer in saliency prediction, we conduct the experiment where models are learned using only the target domain samples, i.e., FT w/o TR in Table 1. We set $n = 10$ as $n = 1, 5$ will yield much worse performance. In all

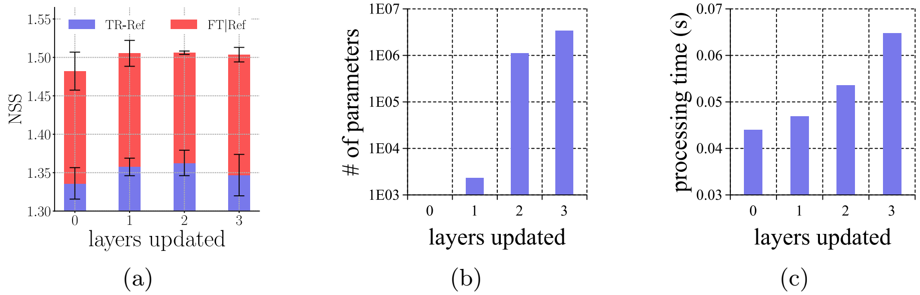


Fig. 5. Ablation study of downstream layers updated within setting $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$ with $n = 1$. Note that when 0 layer is updated, it turns to be TR and FT

settings, the performance of FT w/o TR significantly drops when compare to FT|Ref. These results are even lower than TR and FT, which indicate the importance of efficient initialization with a source domain dataset.

5 Analysis

We study the influences of the number of references, the threshold ϵ, and the layers updated by the proposed framework. All analysis are within setting $\langle \mathsf{S}, \mathsf{D}, \mathsf{W} \rangle$, where the mean score and standard deviation from 3 runs are reported.

5.1 Ablation Study

Effect of Number of References. As shown in Fig. 4, as the number of references increases, the performance of TR−Ref keeps flat or even slightly drops, but the performance of FT|Ref is significantly improved. This implies that the proposed reference process with more reference samples can yield better initialization for fine-tuning. Moreover, the average cosine similarity is increased with more references. This implies that the number of references is helpful to adapt the training process with source domain data to the target domain data.

Effect of Threshold ϵ. We experiment with the proposed framework with $n = 1$, which is more representative and challenging than cases with more references, with various thresholds. An interesting observation in Fig. 4c is that although $\epsilon = 1$ achieves best performance on TR−Ref, it deteriorates the performance of FT|Ref. This shows that when $\epsilon = 1$, all the gradients at each iteration need to be corrected because the cosine similarity between any two gradients is equal or less than 1. As a result, the reference process enforces the training process to overfit the reference samples. This can be verified in Fig. 4d where the average cosine similarity is roughly increased as ϵ is increasing.

Effect of Updated Layers. To understand the effect of layers updated by the proposed 1-reference transfer learning, we experiment with various downstream layers. Consequently, the performance is shown in Fig. 5a, while the number of parameters and the computational cost are reported in Fig. 5b and Fig. 5c, respectively. The layers are downstream layers, which are close to the output. When 0 layer is updated, TR−Ref and FT|Ref are equivalent to TR and FT, respectively. The baseline model in this experiment is DINet.

Figure 5a shows that using the last 2 layers achieves slightly better performance in NSS than using the other numbers of the last layers. However, it takes 69 ms longer in the training process than using the last layers. In light of the trade-off, we use the last layer of the baseline model in Sect. 4.

5.2 Qualitative Comparison

Figure 6 shows the comparison between the predicted saliency maps generated by TR, TR−Ref, FT, and FT|Ref. It can be observed that with the reference process, the proposed framework efficiently leverages the knowledge learned from the source domain, which are based on natural scene images, to subtly identify salience in the new domain. Taking the example in the first row, FT|Ref predicts that the people is salient, which takes the learned knowledge into account, whereas FT predicts that the people is less salient than the text. Figure 7 shows more references lead to better prediction.

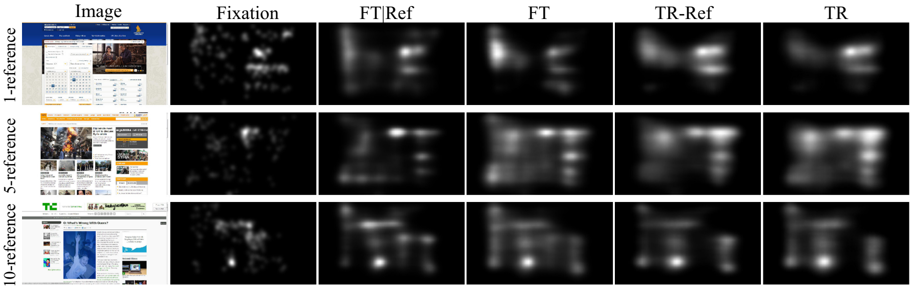


Fig. 6. Qualitative results with human fixations and maps generated by the models trained by the four procedures

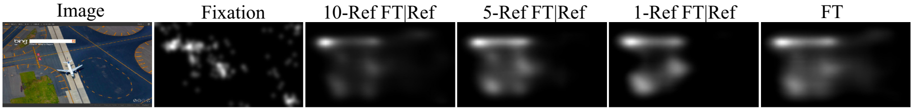


Fig. 7. Qualitative results w.r.t. different n

6 Conclusion

This work studies how to leverage the knowledge learned from a source domain that has adequate images and corresponding human fixations and very few samples (i.e., references) from a new domain (i.e., target domain) to predict saliency maps in the target domain. We propose an n-reference transfer learning framework to guide the training process to converge to a local minimum in favor of the target domain. The proposed framework is gradient-based and model-agnostic. Comprehensive experiments and ablation studies to evaluate the proposed framework are reported. Results show the effectiveness of the framework with a significant performance improvement.

Acknowledgments. This research was funded in part by the NSF under Grants 1908711, 1849107, in part by the University of Minnesota Department of Computer Science and Engineering Start-up Fund (QZ), and in part supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

1. Bäuml, B., Tulbure, A.: Deep n-shot transfer learning for tactile material classification with a flexible pressure-sensitive skin. In: International Conference on Robotics and Automation, pp. 4262–4268 (2019)
2. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 17–36 (2012)
3. Borji, A., Itti, L.: CAT2000: a large scale fixation dataset for boosting saliency research. In: CVPR 2015 Workshop on "Future of Datasets" (2015)
4. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. IEEE Trans. Image Process. **22**(1), 55–69 (2013)
5. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 740–757 (2018)
6. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: International Conference on Pattern Recognition, pp. 3488–3493 (2016)
7. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Trans. Image Process. **27**(10), 5142–5154 (2018)
8. Csurka, G.: A comprehensive survey on domain adaptation for visual applications. In: Csurka, G. (ed.) Domain Adaptation in Computer Vision Applications. ACVPR, pp. 1–35. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58347-1_1
9. Daume III, H., Marcu, D.: Domain adaptation for statistical classifiers. J. Artif. Intell. Res. **26**, 101–126 (2006)

10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
11. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 594–611 (2006)
12. Ge, W., Yu, Y.: Borrowing treasures from the wealthy: deep transfer learning through selective joint fine-tuning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1086–1095 (2017)
13. Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., Feris, R.: SpotTune: transfer learning through adaptive fine-tuning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4805–4814 (2019)
14. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552 (2007)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. Hou, X., Harel, J., Koch, C.: Image signature: highlighting sparse salient regions. IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 194–201 (2011)
17. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
18. Itti, L., Dhavale, N., Pighin, F.: Realistic avatar eye and head animation using a neurobiological model of visual attention. In: Proceedings of SPIE 48th Annual International Symposium on Optical Science and Technology, August 2003
19. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
20. Jiang, M., Huang, S., Duan, J., Zhao, Q.: SALICON: saliency in context. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1072–1080 (2015)
21. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. MIT Technical report (2012)
22. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision, pp. 2106–2113 (2009)
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
25. Kruthiventi, S.S., Ayush, K., Babu, R.V.: DeepFix: a fully convolutional neural network for predicting human eye fixations. IEEE Trans. Image Process. **26**(9), 4446–4456 (2017)
26. Kümmerer, M., Theis, L., Bethge, M.: Deep Gaze I: boosting saliency prediction with feature maps trained on imageNet. In: International Conference on Learning Representations (ICLR 2015), pp. 1–12 (2014)
27. Lake, B., Salakhutdinov, R., Gross, J., Tenenbaum, J.: One shot learning of simple visual concepts. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2011)
28. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
29. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10657–10665 (2019)

30. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Attention transfer from web images for video recognition. In: ACM Multimedia, pp. 1–9 (2017)
31. Li, J., Xu, Z., Wong, Y., Zhao, Q., Kankanhalli, M.S.: GradMix: multi-source transfer across domains and tasks. In: IEEE Winter Conference on Applications of Computer Vision (2020)
32. Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: Revisiting local descriptor based image-to-class measure for few-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7260–7268 (2019)
33. Luo, Y., Wong, Y., Kankanhalli, M., Zhao, Q.: Direction concentration learning: enhancing congruency in machine learning. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
34. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press, Cambridge (2012)
35. Neider, M.B., Zelinsky, G.J.: Scene context guides eye movements during visual search. Vis. Res. **46**(5), 614–621 (2006)
36. Ouerhani, N., Von Wartburg, R., Hugli, H., Müri, R.: Empirical validation of the saliency-based model of visual attention. Electron. Lett. Comput. Vis. Image Anal. **3**(1), 13–24 (2004)
37. Pan, J., et al.: SalGAN: visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081 (2017)
38. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Trans. Neural Netw. **22**(2), 199–210 (2010)
39. Paszke, A., et al.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017)
40. Rothenstein, A.L., Tsotsos, J.K.: Attention links sensing to recognition. Image Vis. Comput. **26**(1), 114–126 (2008)
41. Shan, W., Sun, G., Zhou, X., Liu, Z.: Two-stage transfer learning of end-to-end convolutional neural networks for webpage saliency prediction. In: Sun, Y., Lu, H., Zhang, L., Yang, J., Huang, H. (eds.) IScIDE 2017. LNCS, vol. 10559, pp. 316–324. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67777-4_27
42. Shen, C., Zhao, Q.: Webpage saliency. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 33–46. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_3
43. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. ACM Trans. Graph. **30**(6), 154 (2011)
44. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: relation network for few-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1199–1208 (2018)
45. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. **113**(4), 766 (2006)
46. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems, pp. 3630–3638 (2016)
47. Wolfe, J.M., Horowitz, T.S.: Five factors that guide attention in visual search. Nat. Hum. Behav. **1**(3), 0058 (2017)
48. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5987–5995 (2017)

49. Li, X., Grandvalet, Y., Davoine, F.: Explicit inductive bias for transfer learning with convolutional networks. In: International Conference on Machine Learning, pp. 2830–2839 (2018)
50. Yang, S., Lin, G., Jiang, Q., Lin, W.: A dilated inception network for visual saliency prediction. IEEE Trans. Multimedia **22**(8), 2163–2176 (2020)
51. Yang, Y., Ma, Z., Hauptmann, A.G., Sebe, N.: Feature selection for multimedia analysis by sharing information among multiple tasks. IEEE Trans. Multimedia **15**(3), 661–669 (2012)
52. Zhang, J., Sclaroff, S.: Saliency detection: a Boolean map approach. In: IEEE International Conference on Computer Vision, pp. 153–160 (2013)