Visual Attention Prompted Prediction and Learning

Yifei Zhang[1*], **Bo Pan**[1], **Siyi Gu**[2], **Guangji Bai**[1], **Meikang Qiu**[3], **Xiaofeng Yang**[1], **Liang Zhao**[1]
[1]Emory University
[2]Stanford University
[3]Augusta University
{yifei.zhang2, bo.pan, guangji.bai, xyang43, liang.zhao}@emory.edu, sgu33@stanford.edu, qiumeikang@yahoo.com

Abstract

Visual explanation (attention)-guided learning uses not only labels but also explanations to guide the model reasoning process. While visual attention-guided learning has shown promising results, it requires a large number of explanation annotations that are time-consuming to prepare. However, in many real-world situations, it is usually desired to prompt the model with visual attention without model retraining. For example, when doing AI-assisted cancer classification on a medical image, users (e.g., clinicians) can provide the AI model with visual attention prompts on which areas are indispensable and which are precluded. Despite its promising objectives, achieving visual attention-prompted prediction presents several major challenges: 1) How can the visual prompt be effectively integrated into the model's reasoning process? 2) How should the model handle samples that lack visual prompts? 3) What is the impact on the model's performance when a visual prompt is imperfect? This paper introduces a novel framework for visual attention prompted prediction and learning, utilizing visual prompts to steer the model's reasoning process. To improve performance in non-prompted situations and align it with prompted scenarios, we propose a co-training approach for both non-prompted and prompted models, ensuring they share similar parameters and activation. Additionally, for instances where the visual prompt does not encompass the entire input image, we have developed innovative attention prompt refinement methods. These methods interpolate the incomplete prompts while maintaining alignment with the model's explanations. Extensive experiments on four datasets demonstrate the effectiveness of our proposed framework in enhancing predictions for samples both with and without prompt.

1 Introduction

The "black box" nature of deep learning models often obscures the decision-making process in AI, leading to the emergence of Explainable AI (XAI) aimed at demystifying the rationale behind models [Adadi and Berrada, 2018]. Techniques like CAM, Grad-CAM, and integrated gradients, pivotal in XAI, produce saliency maps highlighting the model's focus areas in the input data [Zhou *et al.*, 2015; Selvaraju *et al.*, 2017; Qi *et al.*, 2019; Bai *et al.*, 2023]. While XAI has advanced in explaining model reasoning, the ultimate aim extends beyond this. It is crucial to leverage XAI to enhance the reasoning and predictive capabilities of Deep Neural Networks (DNNs). Key challenges include incorporating human insight into the model's reasoning and ensuring that such insights positively impact future predictions.

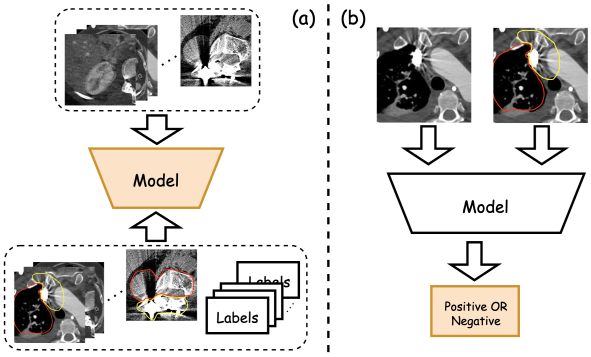

Figure 1: The comparison between attention-guided learning and attention-prompted prediction. (a) explanation-guided learning requires many user-annotated explanations to train the models; (b) attention-prompted prediction enables users to directly guide the model's prediction process by telling the model which areas are "indispensable" (areas in red that look suspicious), "precluded" (areas in yellow that contain artifacts), and "undecided" (other areas).

Explanation-guided learning in natural language processing (NLP) task has been extensively researched, with studies such as [Hsieh *et al.*, 2023; Raffel *et al.*, 2020; Narang *et al.*, 2020] extracting rationales as additional supervision, alongside label supervision, for training language models. In contrast, the visual domain remains not well explored. In recent years, there has been emerging research in visual

*Code and tools are available at https://github.com/yifeizhangcs/visual-attention-prompt

explanation-guided learning. Studies like [Chen *et al.*, 2020; Zhang *et al.*, 2023; Shen *et al.*, 2021; Gu *et al.*, 2023] utilizes human-annotated attention maps to guide the reasoning process of DNNs in Computer Vision (CV) tasks by simultaneously minimizing prediction errors and the disparity between the model's reasoning and the human-crafted true reasoning within the training set, as shown in Figure 1(a). However, visual explanation-guided learning is labor-intensive, time-consuming, and computationally expensive due to the need for large amounts of human-annotated attention maps. In many practical applications, users may have easy and quick (high-level) guidance toward the model for a prediction, e.g., which areas are roughly more important are which are not. For instance, in cancer imaging, clinicians can quickly sketch out areas that are interesting or irrelevant in terms of determining whether the image indicates "cancerous", as shown in Figure 1(b). Hence, there is a need for an efficient way to incorporate these cues to assist the model's decision-making process. Despite many existing works about textual prompts [Oymak *et al.*, 2023; Liu *et al.*, 2023; Ye *et al.*, 2022; Ling *et al.*, 2023], how to prompt in visual space is much less underexplored and the focus of this paper.

Despite the interestingness and significance, achieving visual attention-prompted prediction requires working on visual patterns on images that trigger unique unsolved challenges **1) How to incorporate the visual attention prompt into the model's prediction?** Traditional classifiers only take images for decision-making, but how to embed the prompt such that it can guide the model reasoning as instructed in the visual attention prompt is crucial and very challenging. **2) What if some samples do not have visual attention prompts?** Many samples may not come with prompts, but can their predictions still get some guidance from the samples with prompts? And how? and **3) How to handle the incomplete visual attention prompts?** In practice, it is much easier and quicker to provide an imperfect prompt which may not cover the whole image but just its regions where the user can quickly have some intuition. How can we address the incompleteness and sufficiently utilize it?

To tackle the aforementioned challenges, we introduce the Visual Attention-Prompted Prediction and Learning framework. To address the first challenge, we proposed an attention-prompted prediction framework for integrating visual attention prompts into the model's decision-making process. To counter the second challenge, we developed an attention-prompted co-training mechanism that distills knowledge from the prompted model to the non-prompted model, thereby enhancing future prediction performance for samples without provided prompts. Finally, to tackle the third challenge, we proposed a novel architecture to achieve attention prompt refinement by automatically learning the saliency of "undecided" areas for each pixel. In summary, the main contributions of this paper are as follows: 1) A new framework designed to integrate visual attention prompts into the model's decision-making process; 2) A new attention-prompted co-training algorithm developed to improve predictions for samples without attention prompts by distilling knowledge from the attention-prompted model to the non-attention-prompted model; 3) A novel architecture to refine the incomplete visual attention prompts by automatically learning the saliency of "undecided" areas at the pixel level; 4) Comprehensive experiments on four datasets to demonstrate the effectiveness of our framework in enhancing model predictability for image classification tasks.

2 Related Work

Privileged information The learning with privileged information paradigm, as proposed by Vapnik [Vapnik *et al.*, 2015], introduces a "teacher" that provides additional information to a "student" model during the learning process. The underlying idea is that the teacher's extra explanations help the student develop a more effective model. Lopez et al. [Lopez-Paz *et al.*, 2015] integrated the concepts of distillation and privileged information into 'generalized distillation,' a framework for learning from multiple machines and data representations. Garcia et al. [Garcia *et al.*, 2018] introduced a method for multimodal video action recognition. Pan et al. [Pan *et al.*, 2024] applied privileged information from Large Language Models (LLMs) to Graph Learning.

Attention-guided Learning The integration of human knowledge into interpretable models has been extensively studied in CV and NLP tasks. Attention maps, such as those generated by Grad-CAM [Montavon *et al.*, 2019; Selvaraju *et al.*, 2017] or intrinsic attention mechanism [Vaswani *et al.*, 2017], have been utilized as supervision signals. These signals are aligned with prediction loss to further enhance model performance [Shen *et al.*, 2021; Camburu *et al.*, 2018; Hajialigol *et al.*, 2023; Bai and Zhao, 2022]. HAICS [Shen *et al.*, 2021] proposed a conceptual framework for image classification that incorporates human annotations in the form of scribble annotations as the attention signal. RES [Gao *et al.*, 2022] developed a novel objective to handle inaccurate, incomplete, or inconsistently distributed issues of explanation.

Attention Prompt Prompting, originating from NLP tasks as shown by Devlin et al. [Devlin *et al.*, 2018], has been adapted for CV applications [Dosovitskiy *et al.*, 2020]. Oymak et al. [Oymak *et al.*, 2023] examine prompt-tuning for one-layer attention architectures in contextual mixture models. Jia et al. [Jia *et al.*, 2022] introduced Visual Prompt Tuning (VPT) to add adaptable prompt tokens to the Vision Transformer (ViT) model's patch tokens. Paiss et al. [Paiss *et al.*, 2022] proposed an explainability-based method for improving one-shot classification rates. Finally, Li et al. [Li *et al.*, 2023] developed "saliency prompts" from saliency masks to highlight potential objects in scenes.

3 Problem Formulation

In the context of visual attention-prompted learning and prediction, we consider the samples from a dataset $\mathcal{U}$ to be provided in pair as $(I, D, y) \in \mathcal{U}$, where $I \in \mathbb{R}^{C\times H\times W}$ represents the original image, with C, H, and W denotes the number of channels, height, and width, respectively, y is the class label, and visual attention prompt for the image corresponding to its class label is denoted by $D \in \mathbb{R}^{H\times W}$, with dimensions identical to the original image but with one channel. Attention prompt D serves as an instruction or signal indicating which parts of the input image are particularly relevant or

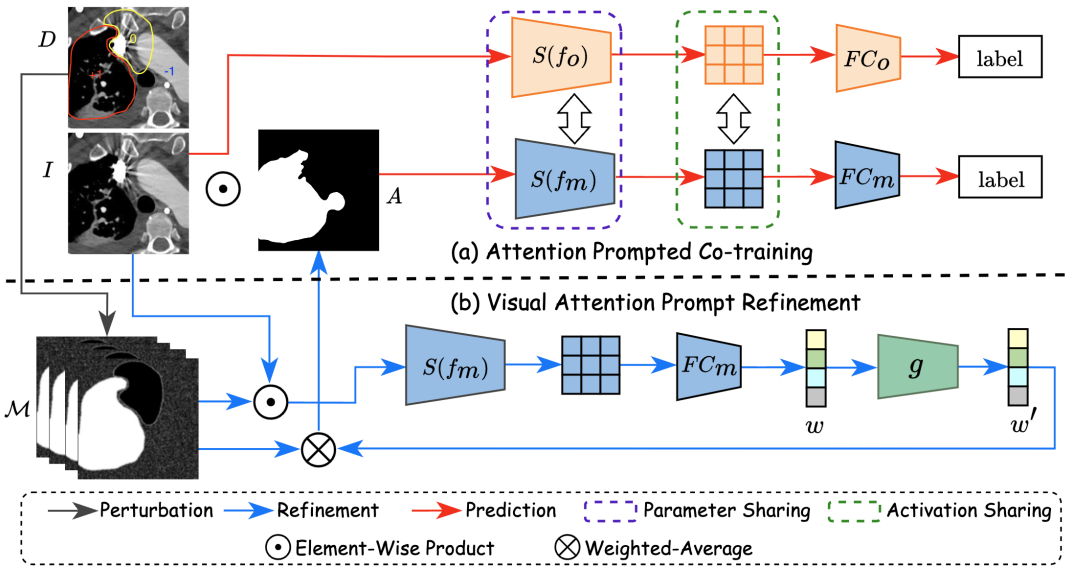

Figure 2: Illustration of the Visual Attention Prompted Prediction and Learning Framework: (a) depicts our proposed Attention-Prompted Co-Training Mechanism, while (b) outlines the proposed Visual Attention Prompt Refinement Architecture.

should be prioritized. The visual attention-prompted prediction process of model f can be denoted as $f : (I, D) \mapsto y$.

Despite the merit of the potential of embedding the visual attention prompt into a prediction to guide it better, achieving it requires tackling nontrivial technical challenges in three key aspects. Firstly, how to use the prompt to guide the model's reasoning so that it can attend to the right places? Secondly, image samples I may not always come with prompts D, but can they benefit from the prompts of the others? Moreover, it is usually much more efficient to provide incomplete prompts covering only part of the image, but how to handle such incompleteness?

4 Methodology

To address the challenges outlined above, we propose a new framework for visual attention-prompted prediction and learning. Section 4.1 introduces the overall framework we propose. Section 4.2 details the attention prompts refinement architecture. Finally, in Section 4.3, we propose the attention-prompted co-training mechanism.

4.1 Overall Framework

In our Visual Attention Prompted Prediction Framework, attention prompts are employed to effectively guide the reasoning process of the AI model. This is achieved by utilizing the attention prompt to mask the original image. Consequently, the model's decision-making process is influenced by areas categorized as "indispensable," "precluded," and "undecided," ensuring that these guided insights are integral to the AI's analytical procedures. In the left of Figure 2(a), "indispensable" and "precluded" areas are shown in red and yellow, respectively, with the remaining areas being "undecided" ones. To address the incompleteness of the prompt, we will respect certain parts that are either "indispensable" or "precluded", but impute the undecided part by eliciting the model's explanation of the reasoning process and aligning it with the given attention prompt. As illustrated in Figure 2(b), the given incomplete prompt is initially inputted into the Visual Attention Prompt Refiner. This step serves to activate and constrain the generation of post-hoc explanations. Subsequently, these elicited post-hoc explanations are refined and utilized to mask the input image, thereby facilitating the attention-prompted prediction process. The details are elaborated on in Section 4.2. As illustrated in Figure 2(a), to handle samples without prompts and allow them to benefit from those with prompts, we propose another non-prompted model, f_o, that takes only the image. This model is co-trained with the prompted model f_m by aligning both the model parameters and activation, as indicated by the dashed rectangles in the figure. $S(f_o)$ and $S(f_m)$ denote the sets of model architectures before the fully-connected layers, while FC_o and FC_m represent the fully-connected layers of two models respectively. The details will be elaborated in Section 4.3.

4.2 Visual Attention Prompt Refinement

In this section, we propose an attention-prone refinement architecture. First, we introduce prompt-guided incomplete prompt refinement architecture. Second, we present an adaptively learnable mask aggregation method.

Prompt-Guided Incomplete Prompt Refinement

To refine incomplete prompts containing areas that are "undecided" yet crucial for prediction, we propose a prompt-guided incomplete prompt refinement architecture designed to enhance the tolerance of our framework to incompleteness and errors within the prompt. This approach aims to learn the saliency of each pixel in the "undecided" areas by aligning the given incomplete prompt with post-hoc explanations. Consider a prompt $D \in \mathbb{R}^{H \times W}$ that includes annotations distinctly highlighting indispensable, precluded,

and undecided areas, we aim to leverage the post-hoc explanations of the prediction for the image prompted with this incomplete prompt, to refine the prompt. However, those gradient-based explainers are usually time-consuming due to the involvement of back-propagation during explanation generation. Therefore, post-hoc explanations obtained without backpropagation such as perturbation-based methods that can directly predict the importance of perturbations to compose explanations are preferred here.

To be specific, each element $\lambda \in D$ corresponds to a pixel in the image and takes a value from the set $\{-1, 0, +1\}$. Here, $\lambda = 0$ indicates that the pixel is precluded, $\lambda = +1$ signifies that the pixel is indispensible, and $\lambda = -1$ denotes that the pixel is undecided. To generate the post-hoc explanation guided by prompt D, we first randomly perturb D as $P(D)$ to generate N binary masks $\mathcal{M} = \{P^{(i)}(D) = M_i\}_{i=1}^{N}$, where $P^{(i)}(\cdot)$ denotes the i-th preturbation process, by setting each pixel with a value equal to -1 to $+1$ with probability p and to 0 otherwise. After obtaining the perturbed masks, we first perform an element-wise multiplication of the original image with each randomly perturbed mask to obtain the masked image. Subsequently, we compute the confidence scores for each mask in $\mathcal{M}$ as $\{(I \odot M_i)\}_{i=1}^{N}$, where I is the image and $\odot$ represents the element-wise multiplication operation. We then calculate the confidence score for each masked image relative to its corresponding class label by $w_i = \mathrm{softmax}_k\, f_m(I \odot M_i), i = 1, \ldots, N$, where k represents the index of the label y and $\mathrm{softmax}_k$ denotes the output of the softmax layer corresponding to the k-th class. Upon calculating the confidence scores for each perturbed mask, we applied weighted averaging and normalized it with the expected pixel value $N \cdot p$ for aggregation. The process resulted in the refined prompt, denoted as A. To summarize, the above computation process can be represented as

$$A = \frac{1}{N \cdot p} \sum_{i=1}^{N} f_m(I \odot P^{(i)}(D)) \cdot P^{(i)}(D) \qquad (1)$$

where the refined prompt is obtained by learning the saliency of "undecided" areas to align with the post-hoc explanation, which is guided by the incomplete prompt.

Adaptively Learnable Masks Aggregation

Aggregating masks by the confidence score can ensure the relationship between confidence and importance, as a higher confidence score indicates greater importance in the aggregation process. However, the confidence score may not necessarily accurately quantify the importance. To overcome this, we require a function that can adaptively learn to quantify the importance score accurately. In the following, we propose to learn such a weight-learning function by satisfying the following inherent constraints of the confidence score:

- *(Monotonicity)*: The function needs to be monotonically non-decreasing, ensuring masks with higher confidence scores receive greater importance in the aggregation.
- *(Endpoint-preserving)*: The function must ensure that inputs of 0 and 1 yield outputs of 0 and 1, respectively. This respects the input range's limits, preserving the key relationship between confidence scores and weights, and mitigates the effects of extremely large or negative values.

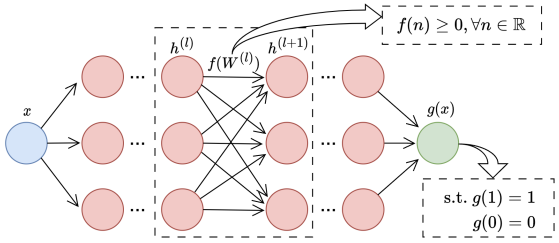

Figure 3: Visualization of proposed weights-learning function based on constrained MLP architecture.

To achieve the formulation and learning of such as weight-learning function, we build a constrained multilayer perceptron (MLP) $g(\cdot)$, to quantify the importance scores. To be more specific, as shown in Figure 3, ensuring the *monotonically non-decreasing* behavior, $g(\cdot)$ is achieved by the principle that the weight matrix must consist of non-negative entries. This crucial constraint is enforced by applying weights derived from a function with a range of positive numbers [Nguyen *et al.*, 2023]. To satisfy the *endpoint-preserving* property, two specific measures have been taken: 1) To guarantee that the function produces an output of 0 when the input is 0, the bias term has been eliminated from all layers of $g(\cdot)$. Additionally, an activation function $\sigma(\cdot)$ has been selected such that $\sigma(0) = 0$, and 2) To ensure that the function yields an output of 1 when the input is 1, a constraint term $g(1) = 1$ is incorporated into the training phase. In details, the mapping between layer k and layer $k+1$ of $g(\cdot)$, which is depicted in Figure 3, is as

$$h^{(k+1)} = \sigma(\phi(W^{(k)})h^{(k)}), k = 1, \ldots, L \qquad (2)$$

where ϕ is a function that has a range within the positive numbers and can be exemplified by functions like the exponential function or a translated hyperbolic tangent, and $L+1$ denotes the total number of layers in the model. With proposed function $g(\cdot)$, the mask aggregation function, as described in Equation 1, can be redefined as

$$A = \frac{1}{N \cdot p} \sum_{i=1}^{N} g(f_m(I \odot P^{(i)}(D))) \cdot P^{(i)}(D) \qquad (3)$$

where the refined prompt is obtained by aggregating masks with the learned weights, as shown in Figure 2(b).

4.3 Attention Prompted Co-training

In this section, we delineate the attention-prompted co-training mechanism. Initially, the framework for parameter-sharing and co-activation is introduced. This is succeeded by a detailed exposition of the learning algorithm, which is predicated on an alternating optimization training strategy.

Parameter-Sharing and Co-Activation Framework

In instances where a prompt is unavailable, the proposed function f_m, which necessitates a prompt, becomes inapplicable. Under such circumstances, an alternative predictor, designated as f_o, is employed. This predictor operates exclusively based on the image data. Rather than learning f_m and f_o independently, we investigate their interrelation. For each

training sample accompanied by a prompt, both f_m and f_o are executed, with the anticipation that f_o will mirror the correct reasoning dictated by f_m and guided by the prompt. This approach not only aims to encourage two models to have similar parameters but also to foster similar activation patterns. For instance, if a prompt directs f_m to disregard the region of an artifact in the image, which results in no activations in that region, f_o should also follow it by disregarding this region. This alignment of activation patterns enables the transfer of knowledge from f_m to f_o in a joint parameter-sharing and co-activation framework. To concretize this concept, we propose the integration of two regularization terms: model parameter-sharing regularization, and co-activation regularization, which collectively embody this approach as

$$\mathcal{L}_{\text{Param}}(\theta_{f_m}, \theta_{f_o}) = \|W_{f_o} - W_{f_m}\|_F^2 \tag{4}$$

$$\mathcal{L}_{\text{Activ}}(\theta_{f_m}, \theta_{f_o}) = \|S(f_o(I)) - S(f_m(I \odot A))\|_F^2 \tag{5}$$

where W_{f_m} and W_{f_o} represent the convolutional layer parameters of two models and $\|\cdot\|_F^2$ represents the squared Frobenius norm [Ma *et al.*, 1994]. Furthermore, the cross-entropy loss associated with the predictions made by the two models, when compared to the target, can be articulated as

$$\mathcal{L}_{\text{Pred}} = -\sum_{i=1}^{K}\sum_{a=1}^{C} \hat{y}_{ia}\left(\log\left(p_m^{ia}\right) + \log\left(p_o^{ia}\right)\right) \tag{6}$$

where $\hat{y}_{ia}$ is the ground truth label for class a of the i^{th} data point in one-hot encoding, $p_m^{ia} = \text{softmax}_a\left(f_m\left(I_i \odot A_i\right)\right)$ predicted probability for class a of the i^{th} sample for the prompted model, and $p_o^{ia} = \text{softmax}_a\left(f_o\left(I_i\right)\right)$ is for the non-prompted model. In conclusion, considering the constraints imposed by the weight-learning function, the learning objective of our study can be formally expressed as

$$\begin{aligned} &\text{minimize} \quad \mathcal{L}_{\text{Pred}} + \lambda_1\mathcal{L}_{\text{Param}} + \lambda_2\mathcal{L}_{\text{Activ}} \\ &\text{subject to} \quad g(1) = 1 \end{aligned} \tag{7}$$

where λ_1 and λ_2 are the weighting hyper-parameters for parameter sharing loss and activation sharing loss, respectively. If we treat the constraint as a Lagrange multiplier [Gordon and Tibshirani, 2012] and solve an equivalent problem by substituting the constraint to a regularization term $\mathcal{L}_{\text{Agg}}(\theta_{\mathbf{g}})$, our overall objective function can be rewritten as

$$\text{minimize} \quad \mathcal{L}_{\text{Pred}} + \lambda_1\mathcal{L}_{\text{Param}} + \lambda_2\mathcal{L}_{\text{Activ}} + \lambda_3\mathcal{L}_{\text{Agg}} \tag{8}$$

where $\mathcal{L}_{\text{Agg}}(\theta_g) = \|g(1) - 1\|$ is commonly chosen to be the ℓ_2-norm, λ_3 is the weighting hyper-parameters for weight-learning function constraint regularization term.

Alternating Training Algorithm

As illustrated in Figure 2, the prompted model f_m exhibits an output-as-input structure during the prompt refinement stage. Specifically, f_m's output is fed back into its input during the prompted prediction stage, creating circular dependencies in training [Han *et al.*, 2017; Xu *et al.*, 2023]. These circular dependencies may introduce challenges in the convergence of the model, potentially leading to instability or oscillatory behavior during training. To mitigate this, we propose an alternating training strategy to break this cyclic dependency and ensure better convergence. This strategy involves three models: the non-prompted model f_o, the prompted model f_m, and the weight-learning function g.

Algorithm 1 Alternating Training

Require: I, D, y
Ensure: f_m, f_o, g
1: **for** $t = 1 : T$ **do**
2: **for** $q = 1 : F$ **do**
3: Compute $\nabla_{\theta_{f_m}}$ based on Equation 9
4: Compute $\nabla_{\theta_{f_o}}$ based on Equation 10
5: $\theta_{f_m} \leftarrow \theta_{f_m} - \eta\nabla_{\theta_{f_m}}$
6: $\theta_{f_o} \leftarrow \theta_{f_o} - \eta\nabla_{\theta_{f_o}}$
7: **end for**
8: **for** $q = 1 : G$ **do**
9: Compute ∇_{θ_g} based on Equation 11
10: $\theta_g \leftarrow \theta_g - \eta\nabla_{\theta_g}$
11: **end for**
12: **end for**

The training algorithm is summarized in Algorithm 1. our algorithm is to fix the parameter for g while updating f_m and f_o with a learning rate η for F iterations from Lines 2-7. The gradients w.r.t. θ_{f_m} and θ_{f_m} are computed as

$$\nabla_{f_m} = \frac{\partial}{\partial\theta_{f_m}}\left(\mathcal{L}_{\text{Pred}} + \lambda_1\mathcal{L}_{\text{Param}} + \lambda_2\mathcal{L}_{\text{Activ}}\right) \tag{9}$$

$$\nabla_{f_o} = \frac{\partial}{\partial\theta_{f_o}}\left(\mathcal{L}_{\text{Pred}} + \lambda_1\mathcal{L}_{\text{Param}} + \lambda_2\mathcal{L}_{\text{Activ}}\right) \tag{10}$$

From Lines 8-11, our algorithm fixes the parameter for f_m and f_o while updating g with a learning rate η for G iterations. The gradients w.r.t. θ_g are computed as

$$\nabla_{\theta_g} = \frac{\partial}{\partial\theta_g}\left(\mathcal{L}_{\text{Pred}} + \lambda_3\mathcal{L}_{\text{Agg}}\right) \tag{11}$$

We repeat the alternating training process outlined in Lines 2-11 for T iterations, continuing until optimal performance is achieved (e.g., no further increase in prediction accuracy on the validation set).

5 Experimental Evaluation

5.1 Datasets

To assess our framework's effectiveness, we employed four datasets: two from real-world scenarios, sourced from MS COCO [Lin *et al.*, 2014], and two from the medical field, namely LIDC-IDRI (LIDC) [Armato III *et al.*, 2011] and the Pancreas dataset [Roth *et al.*, 2015].

Specifically, **LIDC** dataset, featuring lung CT scans with annotated lesions, was preprocessed into 2D images (224×224) and augmented with noise to simulate incomplete prompts. Negative samples were created by slicing surrounding areas of nodules. The final dataset included 2625 nodules and 65505 non-nodules images, split into 100/1200/1200 for training, validation, and testing to reflect limited access to human explanations. **Pancreas** dataset, with normal images from the Cancer Imaging Archive and abnormal images from MSD, was preprocessed similarly to LIDC. It included 281 CT scans with tumors and 80 without. Tumor lesions

Model	Pancreas				LIDC			
	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑
Baseline	85.09	98.82	83.69	90.22	66.40	59.29	69.020	63.47
GRADIA	83.13	<u>99.04</u>	81.12	89.10	67.44	65.65	73.19	68.99
HAICS	86.44	98.99	85.10	91.24	66.86	64.71	74.60	69.14
RES-G	<u>89.89</u>	98.94	89.17	<u>93.79</u>	<u>68.56</u>	**67.93**	71.46	69.27
RES-L	89.79	98.07	<u>89.88</u>	93.78	68.35	65.97	**76.28**	<u>70.55</u>
VAPL	**92.31**	**99.84**	**91.03**	**95.30**	**69.45**	<u>67.43</u>	<u>75.25</u>	**71.13**

Table 1: Comparison of prediction performance on the Pancreas and LIDC datasets between our proposed framework and comparative attention-guided learning methods. The best results for each task are highlighted in boldface, and the second-best results are underlined.

Model	Gender				Scene			
	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑
Baseline	68.35	67.45	69.98	68.69	93.42	94.87	91.68	93.25
GRADIA	70.01	67.83	74.35	70.94	95.03	96.21	92.55	94.34
HAICS	69.29	66.42	73.61	69.83	94.89	95.73	92.94	94.31
RES-G	<u>71.33</u>	<u>69.98</u>	**78.53**	<u>74.01</u>	<u>95.91</u>	96.22	**95.35**	<u>95.78</u>
RES-L	70.39	68.41	73.29	70.77	95.53	<u>96.98</u>	<u>94.56</u>	95.75
VAPL	**73.36**	**71.43**	<u>76.88</u>	**74.05**	**96.39**	**97.43**	94.48	**95.93**

Table 2: Comparison of prediction performance on the Gender and Scene datasets between our proposed framework and comparative attention-guided learning methods. The best results for each task are highlighted in boldface, and the second-best results are underlined.

were treated as inaccurate explanations, while pancreas segmentations were ground truth. Data was split into 30/30/rest for training, validation, and testing, maintaining class balance. For **Gender** classification dataset [Gao *et al.*, 2022], involved extracting 1,600 images from MS COCO based on captions mentioning "men" or "women". Images with both genders, multiple people, or unclear figures were excluded. For **Scene** [Gao *et al.*, 2022] recognition, balanced nature, and urban categories from Places365, again selecting 100 images for training. each image from two datasets comes with explanations that were obtained through a custom user interface (UI). Human annotations were collected for all images, and only 100 were randomly chosen for training to simulate limited access to explanations. For each sample image, the factual region of the human-explanation annotation is treated as the "indispensable" area, the counterfactual region is considered the "precluded" area, and the remaining regions are categorized as "undecided" areas, thereby facilitating the creation of an incomplete visual attention prompt.

5.2 Experimental Setup

Comparison Methods To evaluate the effectiveness of our proposed Visual Attention-Prompted Prediction and Learning framework (hereafter referred to as "VAPL"), comparative studies were conducted with four prominent attention-guided learning methods, namely, GRAIDA [Gao *et al.*, 2022], HAICS [Shen *et al.*, 2021], RES-G, and RES-L [Gao *et al.*, 2022]. These comparison methods were trained following the respective implementation guidelines presented in their papers. Additionally, comparisons were made with a baseline model, comprising a ResNet-18 architecture [He *et al.*, 2016], trained exclusively using prediction loss and employing original images as input.

Implementation Details In this study, the ResNet-18 architecture was uniformly employed as the backbone model across all evaluated methods. The experimental setup was standardized with a batch size of 16, and the number of perturbed masks was set to 5000. Furthermore, a pixel conversion probability of 0.1 was established. The training was conducted over 10 epochs, each comprising 5 iterations for the alternating updating phase, effectively resulting in 50 training epochs for each model. The Adam optimization algorithm [Kingma and Ba, 2014] was utilized with a learning rate of 0.0001. Regarding computational resources, all experiments were executed using an NVIDIA GTX 3090 GPU. To assess the classification performance of the proposed framework, we employed conventional metrics such as accuracy, precision, recall, and the F1 score.

5.3 Results Evaluation

Table 1 shows the quantitative evaluation of two medical image classification tasks: LIDC and Pancreas. Overall, our method outperforms all other explanation-guided learning methods on both LIDC and pancreas datasets. In particular, our model achieves the best accuracy and F1 on the pulmonary nodule classification task and the best accuracy, precision, recall, and F1 on the pancreatic tumor classification task. Our method achieved an accuracy of 92.30% and 69.45% on the two respective datasets, representing an improvement of 8.5% and 4.6% compared to the baseline Resnet-18 approach. In addition, our method attained F-1 scores of 95.30% and 71.13% on the two datasets, marking an enhancement of 5.6% and 12.1% respectively when compared to the baseline Resnet-18 method. Our findings demonstrate that the performance of our prediction methodology is enhanced when incorporating explanations into the prediction phase. While the precision or recall scores reported on the LIDC dataset are marginally lower, our approach achieves the highest F1 scores across both datasets. This outcome signifies that our overall performance surpasses that of all other

Ablation	LIDC-IDRI				Pancreas			
	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑
VAPL	69.45	67.43	75.25	71.13	92.31	99.84	91.03	95.30
VAPL-1	67.18	64.66	75.79	69.78	89.25	98.14	85.76	91.53
VAPL-2	66.68	64.34	72.41	68.91	88.49	97.17	82.03	88.96
VAPL-3	66.43	63.15	72.23	67.38	88.26	96.14	81.43	88.17
VAPL-4	68.14	64.75	73.43	68.81	90.46	98.54	89.59	93.85

Table 3: Ablation study on LIDC-IDRI (left) and Pancreas dataset (right).

attention-guided learning methods.

Table 2 presents a comprehensive comparison of various models across two image classification tasks: Gender and Scene. Notably, the VAPL model demonstrates superior performance, achieving the highest scores in almost all metrics for both tasks, with particularly outstanding results in the Scene task where it achieves an impressive 96.39% in Accuracy and 97.43% in Precision. The RES-G model also shows commendable performance, especially in the Scene task, where it records the highest Recall of 95.35% and a competitive F1 score of 95.78%. In contrast, the Baseline model, as expected, trails behind in most metrics, underscoring the effectiveness of advanced models like VAPL and RES-G. This analysis underscores the significance of specialized models in enhancing classification accuracy in attention-guided learning tasks, with VAPL excelling in precision-oriented metrics and RES-G showing strength in recall-focused areas.

We further provide a sensitivity analysis of the hyper-parameters λ_1, λ_2, and λ_3, which denote the weights of the parameter sharing regularization loss, activation sharing regularization loss, and weight-learning function regularization loss, respectively. Figure 4 illustrates the accuracy of the proposed model for various weights in the context of pancreatic tumor classification, with the red dashed lines representing the baseline model's performance. Overall, the model exhibits relatively high sensitivity to variations in the weights for the parameter sharing and activation sharing regularization losses, while demonstrating lower sensitivity to the weight-learning function regularization loss. Additionally, a concave curvature is observed for all three hyper-parameters. Optimal overall performance is achieved when λ_1, λ_2, and λ_3 are set to 0.1, 1, and 10, respectively.

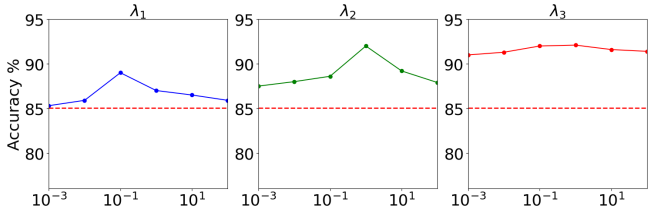

Figure 4: Sensitivity Analysis on the Pancreas dataset.

5.4 Ablation Study

To evaluate the efficacy of the proposed visual attention-prompted learning and prediction (VAPL) framework, an ablation study was conducted with four distinct variants: (**VAPL-1**), which involves the omission of the visual attention prompt refinement architecture, thereby relying exclusively on the incomplete visual attention prompt during the learning and prediction phases. (**VAPL-2**), which pertains to the exclusion of parameter-sharing regularization between the two models, while retaining co-activation regularization during training. (**VAPL-3**), which entails the removal of co-activation regularization, and maintaining only parameter-sharing regularization during the training process. (**VAPL-4**), characterized by the absence of the proposed weight-learning function and instead employing a weighted averaging method based solely on confidence scores computed by classifiers.

The outcomes of the ablation study conducted on the LIDC and Pancreas datasets are presented in Table 3. While each component individually contributes to the overall performance of the model, it is observed from **VAPL-2** and **VAPL-3** that co-training with parameter and activation sharing regularization offers a notably larger enhancement in predictability. Furthermore, the results about **VAPL-1** highlight the significant contribution of the proposed visual attention prompt refinement method to improved performance. This is particularly evident in the Pancreas dataset, which demonstrates a greater prevalence of prompt incompleteness issues.

6 Conclusion and Discussion

This research paper introduces a Visual Attention Prompted Prediction and Learning Framework, a novel approach that integrates visual attention prompts into the decision-making process of models. The framework effectively tackles challenges like incomplete information from visual prompts and predictions for samples lacking such prompts through an attention-prompted co-training mechanism with parameter-sharing and co-activation regularization. This helps align activation patterns and facilitates knowledge transfer between prompted and non-prompted models. Additionally, the framework incorporates a prompt-guided refinement method with an adaptively learnable mask aggregation function to manage prompt incompleteness. Its efficacy is validated across four datasets from various domains, showing improved predictive accuracy for both samples with and without attention prompts. Despite the aforementioned advantages, we also acknowledge that prompts from users can introduce bias, though our framework is tolerant of it. In the future, we plan to explore the application of our proposed framework to graph learning and prediction, incorporating prompts with graph attention mechanisms, such as subgraphs.

Acknowledgments

This work is supported by the National Science Foundation (NSF) Grant No. 1755850, No. 1841520, No. 2007716, No. 2007976, No. 1942594, No. 1907805, a Jeffress Memorial Trust Award, Amazon Research Award, Oracle for Research Grant Award, Cisco Faculty Research Award, NVIDIA GPU Grant, Design Knowledge Company (subcontract number: 10827.002.120.04), CIFellowship (2021CIF-Emory-05), and the Department of Homeland Security under Grant No. 17STCIN00001.

References

[Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[Armato III *et al.*, 2011] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[Bai and Zhao, 2022] Guangji Bai and Liang Zhao. Saliency-regularized deep multi-task learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 15–25, 2022.

[Bai *et al.*, 2023] Guangji Bai, Qilong Zhao, Xiaoyang Jiang, Yifei Zhang, and Liang Zhao. Saliency-guided hidden associative replay for continual learning. *arXiv preprint arXiv:2310.04334*, 2023.

[Camburu *et al.*, 2018] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

[Chen *et al.*, 2020] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 91–107. Springer, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Gao *et al.*, 2022] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Zhao Liang. Res: A robust framework for guiding visual explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 432–442, 2022.

[Garcia *et al.*, 2018] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[Gordon and Tibshirani, 2012] Geoff Gordon and Ryan Tibshirani. Karush-kuhn-tucker conditions. *Optimization*, 10(725/36):725, 2012.

[Gu *et al.*, 2023] Siyi Gu, Yifei Zhang, Yuyang Gao, Xiaofeng Yang, and Liang Zhao. Essa: Explanation iterative supervision via saliency-guided data augmentation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 567–576, 2023.

[Hajialigol *et al.*, 2023] Daniel Hajialigol, Hanwen Liu, and Xuan Wang. Xai-class: Explanation-enhanced text classification with extremely weak supervision. *arXiv preprint arXiv:2311.00189*, 2023.

[Han *et al.*, 2017] Tian Han, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. Alternating back-propagation for generator network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[He *et al.*, 2016] 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.

[Hsieh *et al.*, 2023] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2023] Hao Li, Dingwen Zhang, Nian Liu, Lechao Cheng, Yalun Dai, Chao Zhang, Xinggang Wang, and Junwei Han. Boosting low-data instance segmentation by unsupervised pre-training with saliency prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Ling *et al.*, 2023] Chen Ling, Xuchao Zhang, Xujiang Zhao, Yanchi Liu, Wei Cheng, Mika Oishi, Takao Osaki, Katsushi Matsuda, Haifeng Chen, and Liang Zhao. Open-ended commonsense reasoning with unrestricted answer candidates. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[Liu *et al.*, 2023] Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-Based Systems*, 259:110064, 2023.

[Lopez-Paz *et al.*, 2015] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[Ma *et al.*, 1994] Changxue Ma, Y. Kamp, and L.F. Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–265, 1994.

[Montavon *et al.*, 2019] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

[Narang *et al.*, 2020] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*, 2020.

[Nguyen *et al.*, 2023] An-Phi Nguyen, Dana Lea Moreno, Nicolas Le-Bel, and María Rodríguez Martínez. Mononet: Enhancing interpretability in neural networks via monotonic features. *Bioinformatics Advances*, 3(1):vbad016, 2023.

[Oymak *et al.*, 2023] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. *arXiv preprint arXiv:2306.03435*, 2023.

[Paiss *et al.*, 2022] Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. In *European Conference on Computer Vision*, pages 334–350. Springer, 2022.

[Pan *et al.*, 2024] Bo Pan, Zheng Zhang, Yifei Zhang, Yuntong Hu, and Liang Zhao. Distilling large language models for text-attributed graph learning. *arXiv preprint arXiv:2402.12022*, 2024.

[Qi *et al.*, 2019] Zhongang Qi, Saeed Khorram, and Fuxin Li. Visualizing deep networks by optimizing with integrated gradients. In *CVPR Workshops*, volume 2, pages 1–4, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[Roth *et al.*, 2015] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pages 556–564. Springer, 2015.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Shen *et al.*, 2021] Haifeng Shen, Kewen Liao, Zhibin Liao, Job Doornberg, Maoying Qiao, Anton Van Den Hengel, and Johan W Verjans. Human-ai interactive and continuous sensemaking: A case study of image classification using scribble attention maps. In *Extended Abstracts of CHI*, pages 1–8, 2021.

[Vapnik *et al.*, 2015] Vladimir Vapnik, Rauf Izmailov, et al. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Xu *et al.*, 2023] Jintao Xu, Chenglong Bao, and Wenxun Xing. Convergence rates of training deep neural networks via alternating minimization methods. *Optimization Letters*, pages 1–15, 2023.

[Ye *et al.*, 2022] Hongbin Ye, Ningyu Zhang, Shumin Deng, Xiang Chen, Hui Chen, Feiyu Xiong, Xi Chen, and Huajun Chen. Ontology-enhanced prompt-tuning for few-shot learning. In *Proceedings of the ACM Web Conference 2022*, pages 778–787, 2022.

[Zhang *et al.*, 2023] Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1977–1987, 2023.

[Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.