Appendix

A. Weakly-supervised localization

In this section, we provide quantitative and additional qualitative results for weakly-supervised localization, discussed in the Sec. 5.3 of the main paper. Our quantitative results in Tab. 5, together with the qualitative results in Fig. 5 and Fig. 6, demonstrate the capability of our framework in learning fine-grained representations that can be used for more accurate pathology localization when just image-level annotations are available.

A.1. Quantitative results

Experimental setup: Following the common protocol [11–13], we quantitatively evaluate the applicability of our DiRA framework in a weakly supervised setting using ChestX-ray14 dataset. First, we use min-max normalization to normalize each heatmap; then, following [11], we binarize the heatmaps by thresholding at $\{60, 180\}$, and generate bounding boxes around the isolated regions. To evaluate localization accuracy, we compute the intersection over union (IoU) between the generated and ground truth bounding boxes. According to [11, 12], a localization is correct when the bounding box prediction overlaps with the ground truth box with IoU $\geq \delta$. Following [11], we investigate the accuracy of localization under various δ values, from 10% to 60%. We run each method ten times and report the average accuracy across all runs.

Result: Tab. 5 shows the pathology localization accuracy of our DiRA and underlying discriminative models. As seen, in each of the six IoU thresholds, DiRA models significantly outperform the corresponding discriminative models. In particular, the average of improvement for MoCo-v2, Barlow Twins, and SimSiam across all IoU thresholds is 2.38%, 5.4%, and 9.4%, respectively.

A.2. Qualitative results

Experimental setup: During training, we initialize models with our DiRA pre-trained models, and fine-tune downstream models using only image-level disease labels. We use heatmaps to approximate the spatial location of a particular thorax disease. We generate heatmaps using Grad-CAM [13], a technique for highlighting the important regions in the image for predicting the pathology class.

Results: Fig. 6 presents the visualizations of heatmaps generated by DiRA and the corresponding discriminative models for 8 thorax pathologies in ChestX-ray14 dataset. As seen, DiRA models provide more accurate pathology localizations compared to the underlying discriminative methods. These results demonstrate the impact of restorative learning in providing fine-grained features that are useful for disease localization.

B. Datasets and tasks

We have examined our framework in a diverse suite of 9 downstream tasks, including classification and segmentation in X-ray, CT, and MRI modalities. In this section, we provide the details of each dataset and the underlying task, as well as the evaluation metric for each task.

ChestX-ray14: ChestX-ray14 is a large open source dataset of de-identifie chest X-ray images. The dataset includes 112K chest images taken from 30K unique patients. The ground truth consists of a label space of 14 thorax diseases. We use the official patient-wise split released with the dataset, including 86K training images and 25K testing images. The models are trained to predict 14 pathologies in a multi-label classification setting. The mean AUC score over 14 diseases is used to evaluate the classification performance. In addition to image-level labels, ChestX-ray14 provides bounding box annotations for approximately 1,000 test images. Of this set of images, bounding box annotations are available for 8 out of 14 thorax diseases. During testing, we use bounding box annotations to assess the accuracy of pathology localization in a weakly-supervised setting. The mean accuracy over 8 diseases is used to evaluate the localization performance.

CheXpert: CheXpert is a hospital-scale publicly available dataset with 224K chest X-ray images taken from 65K unique patients. We use the official data split released with the dataset, including 224K training and 234 test images. The ground truth for the training set includes 14 thoracic pathologies that were retrieved automatically from radiology reports. The testing set is labeled manually by board-certified radiologists for 5 selected thoracic pathologies— Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion. The models are trained to predict five pathologies in a multi-label classification setting. The mean AUC score over 5 diseases is used to evaluate the classification performance.

SIIM-ACR: This open dataset is provided by the Society for Imaging Informatics in Medicine (SIIM) and American College of Radiology, including 10K chest X-ray images and pixel-wise segmentation mask for Pneumothorax disease. We randomly divided the dataset into training (80%) and testing (20%). The models are trained to segment pneumothorax from chest radiographic images (if present). The segmentation performance was measured by the mean Dice coefficient score.

NIH Montgomery: This publicly available dataset is provided by the Montgomery County's Tuberculosis screening program, including 138 chest X-ray images. There are 80 normal cases and 58 cases with Tuberculosis (TB) indications in this dataset. Moreover, ground truth segmentation masks for left and right lungs are provided. We randomly

Method	$\delta = 10\%$	$\delta = 20\%$	$\delta = 30\%$	$\delta = 40\%$	$\delta = 50\%$	$\delta = 60\%$
MoCo-v2 [3]	54.89	39.43	24.81	14.59	7.58	2.68
DiRA$_{\text{MoCo-v2}}$	**58.13** (↑ 3.2)	**42.74** (↑ 3.3)	**27.52** (↑ 2.7)	**16.25** (↑ 1.7)	**9.30** (↑ 1.7)	**4.35** (↑ 1.7)
Barlow Twins [4]	50.54	38.01	26.36	16.93	9.31	4.69
DiRA$_{\text{BarlowTwins}}$	**58.98** (↑ 8.4)	**45.26** (↑ 7.2)	**32.71** (↑ 6.3)	**21.71** (↑ 4.8)	**13.62** (↑ 4.3)	**6.26** (↑ 1.6)
SimSiam [5]	30.24	19.80	11.46	5.62	2.30	0.79
DiRA$_{\text{SimSiam}}$	**51.07** (↑ 20.8)	**34.24** (↑ 14.4)	**20.64** (↑ 9.2)	**11.32** (↑ 5.7)	**6.46** (↑ 4.2)	**2.90** (↑ 2.1)

Table 5. **Weakly-supervised pathology localization accuracy under different IoU thresholds (δ):** DiRA models provide stronger representations for pathology localization with only image-level annotations. For each method, we report the average performance over ten runs. The green arrows show the improvement of DiRA models compared with the underlying discriminative method in each IoU threshold.

divided the dataset into a training set (80%) and a test set (20%). The models are trained to segment left and right lungs in chest scans. The segmentation performance is evaluated by the mean Dice score.

LUNA: This publicly-available dataset consists of 888 lung CT scans with a slice thickness of less than 2.5mm. The dataset were divided into training (445 cases), validation (178 cases), and test (265 cases) sets. The dataset provides a set of 5M candidate locations for lung nodule. Each location is labeled as true positive (1) or false positive (0). The models are trained to classify lung nodule candidates into true positives and false positives in a binary classification setting. We evaluate the classification accuracy by Area Under the Curve (AUC) score.

PE-CAD: This dataset includes 121 computed tomography pulmonary angiography (CTPA) scans with a total of 326 pulmonary embolism (PE). The dataset provides a set of candidate locations for PE and is divided at the patient-level into training and test sets. Training set contains 434 true positive PE candidates and 3,406 false positive PE candidates. Test set contains 253 true positive PE candidates and 2,162 false positive PE candidates. We pre-processed the 3D scans as suggested in [6]. The 3D models are trained to classify PE candidates into true positives and false positives in a binary classification setting. We evaluate the classification accuracy by Area Under the Curve (AUC) score at candidate-level.

LIDC-IDRI: The Lung Image Database Consortium image collection (LIDC-IDRI) dataset is created by seven academic centers and eight medical imaging companies. The dataset includes 1,018 chest CT scans and marked-up annotated lung nodules. The dataset is divided into training (510), validation (100), and test (408) sets. We pre-processed the data by re-sampling the 3D volumes to 1-1-1 spacing and then extracting a 64×64×32 crop around each nodule. The models are trained to segment long nodules in these 3D crops. The segmentation accuracy is measured by the Intersection over Union (IoU) metric.

LiTS: The dataset is provided by MICCAI 2017 LiTS Challenge, including 130 CT scans with expert ground-truth segmentation masks for liver and tumor lesions. We divide dataset into training (100 patients), validation (15 patients), and test (15 patients) sets. The models are trained to segment liver in 3D scans. The segmentation accuracy is measured by the Intersection over Union (IoU) metric.

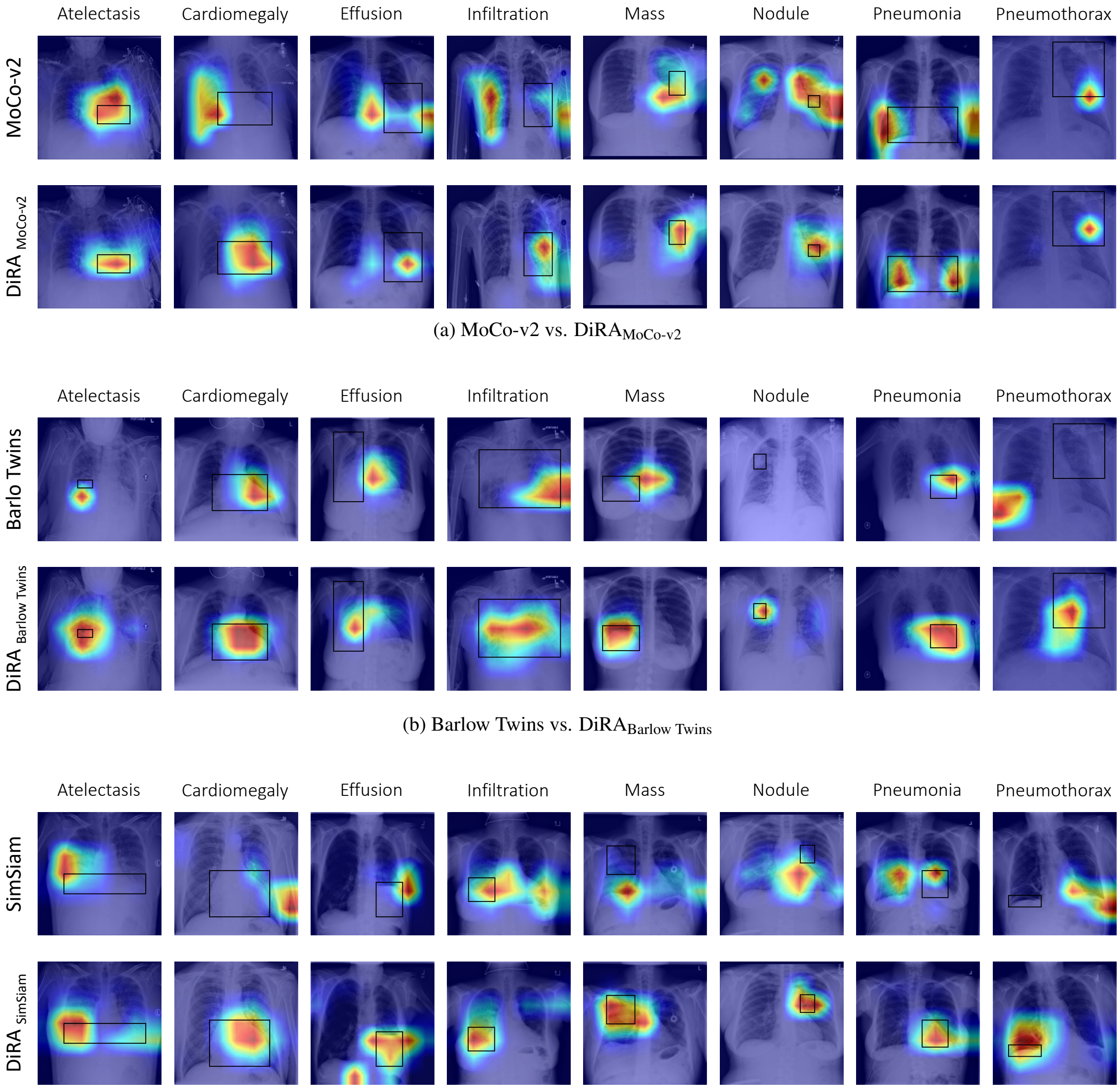
BraTS: The dataset includes brain MRI scans of 285 patients (210 HGG and 75 LGG) and segmentation ground truth for necrotic and non-enhancing tumor core, peritumoral edema, GD-enhancing tumor, and background. For each patient, four different MR volumes are available: native T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 fluid attenuated inversion recovery (FLAIR). We divide dataset at patient-level into training (190 patients) and testing (95 patients) sets. The models are trained to segment brain tumors (background as negatives class and tumor sub-regions as positive class). The segmentation accuracy is measured by the Intersection over Union (IoU) metric.

C. Implementation

C.1. Pre-training settings

We apply DiRA to four existing self-supervised methods [1, 3–5]. To be self-contained, we'll explain each method briefly here. Also, we provide additional pre-training details that supplements Sec. 4.1.

MoCo-v2 [3]: We adopt MoCo-v2— a popular representative of *contrastive learning* methods, into our framework. MoCo leverages a momentum encoder to ensure the consistency of negative samples as they evolve during training. Moreover, a queue $K = \{k_1, k_2, ...k_N\}$ is utilized to store the representations of negative samples. The discrimination task is to contrast representations of positive and negative samples. As MoCo-v2 is adopted in DiRA, the encoder f_θ and projection head h_θ are updated by back-propagation, while f_ξ and h_ξ are updated by using an exponential moving average (EMA) of the parameters in f_θ and h_θ, respectively. The discrimination branch is trained using InfoNCE



(a) MoCo-v2 vs. DiRA$_{\text{MoCo-v2}}$

(b) Barlow Twins vs. DiRA$_{\text{Barlow Twins}}$

(c) SimSiam vs. DiRA$_{\text{SimSiam}}$

Figure 6. **Visualization of Grad-CAM heatmaps:** We provide the heatmap examples for 8 thorax diseases in each column. The first row in each sub-figure represents the results for the original self-supervised method, while the second row represents the original method when adopted in DiRA framework. The black boxes represents the localization ground truths.

loss [7], which for a pair of positive samples x_1 and x_2 defined as follows:

$$\mathcal{L}_{dis} = -log\frac{exp(z_1 \cdot z_2/\tau)}{\sum\limits_{n=0}^{N} exp(z_1 \cdot k_n/\tau)} \qquad (5)$$

where $z_1 = h_\theta(f_\theta(x_1))$ and $z_2 = h_\xi(f_\xi(x_2))$, τ is a temperature hyperparameter, and N is the queue size. Following [3], f_θ is a standard ResNet-50 and h_θ is a two-layer MLP head (hidden layer 2048-d, with ReLU). Moreover, when adopting MoCo-v2 in DiRA, f_θ, h_θ, and g_θ are optimized using SGD with an initial learning rate of 0.03, weight decay 0.0001, and the SGD momentum 0.9.

SimSiam [5]: We adopt SimSiam— a popular representative of *asymmetric instance discrimination* methods, into our framework. SimSiam trains the model without negative pairs and directly maximizes the similarity of two views from an image using a simple siamese network followed by a predictor head. To prevent collapsing solutions, a stop-gradient operation is utilized. As such, the model parameters are only updated using one distorted version of the input, while the representations from another distorted version are used as a fixed target. As SimSiam is adopted in DiRA, the encoder f_θ and projection head h_θ share weights with f_ξ and h_ξ, respectively. The model is trained to maximize the agreement between the representations of positive samples using negative cosine similarity, defined as follows:

$$\mathcal{D}(z_1, y_2) = -\frac{z_1}{\|z_1\|_2} \cdot \frac{y_2}{\|y_2\|_2} \tag{6}$$

where $z_1 = h_\theta(f_\theta(x_1))$ and $y_2 = f_\xi(x_2)$. The discrimination branch is trained using a symmetrized loss as follows:

$$\mathcal{L}_{dis} = \frac{1}{2}\mathcal{D}(z_1, stopgrad(y_2)) + \frac{1}{2}\mathcal{D}(z_2, stopgrad(y_1)) \tag{7}$$

where stopgrad means that y_2 is treated as a constant in this term. Following [5], f_θ is a standard ResNet-50 and h_θ is a three-layer projection MLP head (hidden layer 2048-d), followed by a two-layer predictor MLP head. Moreover, when adopting SimSiam in DiRA, f_θ, h_θ, and g_θ are optimized using SGD with a linear scaling learning rate (lr×BatchSize/256). The initial learning rate is 0.05, weight decay is 0.0001, and the SGD momentum is 0.9.

Barlow Twins [4]: We adopt Barlow Twins— a popular representative of *redundancy reduction instance discrimination learning* methods, into our framework. Barlow Twins makes the cross-correlation matrix computed from two siamese branches close to the identity matrix. By equating the diagonal elements of the cross-correlation matrix to 1, the representation will be invariant to the distortions applied to the samples. By equating the off-diagonal elements of the cross-correlation matrix to 0, the different vector components of the representation will be decorrelated, so that the output units contain non-redundant information about the sample. The discrimination loss is defined as follows:

$$\mathcal{L}_{dis} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{i \neq j} \mathcal{C}_{ij}^2 \tag{8}$$

where $\mathcal{C}$ is the cross-correlation matrix computed between the outputs of the h_θ and h_ξ networks along the batch dimension. λ is a coefficient that determines the importance of the invariance term and redundancy reduction term in the loss. Following [4], f_θ is a standard ResNet-50 and h_θ is a three-layer MLP head. Moreover, when adopting Barlow Twins in DiRA, f_θ, h_θ, and g_θ are optimized using LARS optimizer with the learning rate schedule similar to [4].

TransVW [1]: TransVW defines the similar anatomical patterns within medical images as anatomical visual words, and combines the discrimination and restoration of visual words in a single loss objective. As TransVW is adopted in DiRA, the encoder f_θ and projection head h_θ are identical to f_ξ and h_ξ, respectively. In particular, the discrimination branch is trained to classify instances of visual words according to their pseudo class labels using the standard cross-entropy loss:

$$\mathcal{L}_{dis} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} \mathcal{Y}_{bc} \log \mathcal{P}_{bc} \tag{9}$$

where B denotes the batch size; C denotes the number of visual words classes; $\mathcal{Y}$ and $\mathcal{P}$ represent the ground truth (one-hot pseudo label vector obtained from visual word classes) and the prediction of h_θ, respectively. Following [1], we use 3D U-Net as the f_θ and g_θ. h_θ includes a set of fully-connected layers followed by a classification head. f_θ and g_θ are trained with the same setting as [1].

Joint training process: Following [8, 9], we perform the overall pre-training with the discrimination, restoration, and adversarial losses in a gradual evolutionary manner. First, the encoder f_θ along with projector h_θ are optimized using the discrimination loss $\mathcal{L}_{dis}$ according to the learning schedule of the original discriminative methods [1, 3–5], empowering the model with an initial discrimination ability. Then, the restoration and adversarial losses are further fused into the training process incrementally. To stabilize the adversarial training process and reduce the noise from imperfect restoration at initial epochs [9], we first warm up the f_θ and g_θ using the $\mathcal{L}_{dis} + \mathcal{L}_{res}$, and then add the adversarial loss $\mathcal{L}_{adv}$ to jointly train the whole framework; the optimization of the framework by incorporation of $\mathcal{L}_{res}$ and $\mathcal{L}_{adv}$ takes up to 800 epochs. Following [2], we use the early-stop technique on the validation set, and the checkpoints with the lowest validation loss are used for fine-tuning.

C.2. Fine-tuning settings

Preprocessing and data augmentation: Following [10], for 2D target tasks on X-ray datasets (ChestX-ray14, CheXpert, SIIM-ACR, and Montgomery), we resize the images to 224×224. For thorax diseases classification tasks on ChestX-ray14 and CheXpert, we apply standard data augmentation techniques, including random cropping and resizing, horizontal flipping, and rotating. For segmentation tasks on SIIM-ACR and Montgomery, we apply random brightness contrast, random gamma, optical distortion, elastic transformation, and grid distortion. For 3D target tasks, we use regular data augmentations including random flipping, transposing, rotating, and adding Gaussian noise.

Training parameters: We endeavour to optimize each downstream task with the best performing hyperparameters. In all 2D and 3D downstream tasks, we use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We use early-stop mechanism using the 10% of the training data as the validation set to avoid over-fitting. For 2D classification tasks on ChestX-ray14 and CheXpert datasets, we use a learning rate $2e-4$ and *ReduceLROnPlateau* as the learning rate decay scheduler. For 2D segmentation tasks on SIIM-ACR and Montgomery, we use a learning rate $1e-3$ and *cosine* learning rate decay scheduler. For all 3D downstream tasks, we use *ReduceLROnPlateau* as the learning rate decay scheduler. For downstream tasks on LUNA, PE-CAD, LIDC, and LiTS, we use a learning rate $1e-2$. For BraTS dataset, we use a learning rate of $1e-3$.

References

[1] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 40(10):2857–2868, 2021. 13, 15

[2] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3499–3509, October 2021. 15

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 13, 14, 15

[4] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv:2103.03230*, 2021. 13, 15

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021. 13, 14, 15

[6] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. 13

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 14

[8] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. 15

[9] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2004–2013, June 2021. 15

[10] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B. Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In Shadi Albarqouni, M. Jorge Cardoso, Qi Dou, Konstantinos Kamnitsas, Bishesh Khanal, Islem Rekik, Nicola Rieke, Debdoot Sheet, Sotirios Tsaftaris, Daguang Xu, and Ziyue Xu, editors, *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13, Cham, 2021. Springer International Publishing. 15

[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017. 12

[12] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 12

[13] 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 (ICCV)*, Oct 2017. 12