ViT-CX: Causal Explanation of Vision Transformers

Supplementary Material

A More Visual Examples to Compare ViT-CX with Previous Methods

Visual examples to compare different explanation methods when explaining **Swin-B** and **DeiT-B**, are given in Figure 1 and 2 respectively. These examples show that our ViT-CX method can highlight the more complete evidences that are important to the model prediction for different ViT models.

A.1 Examples from Swin-B

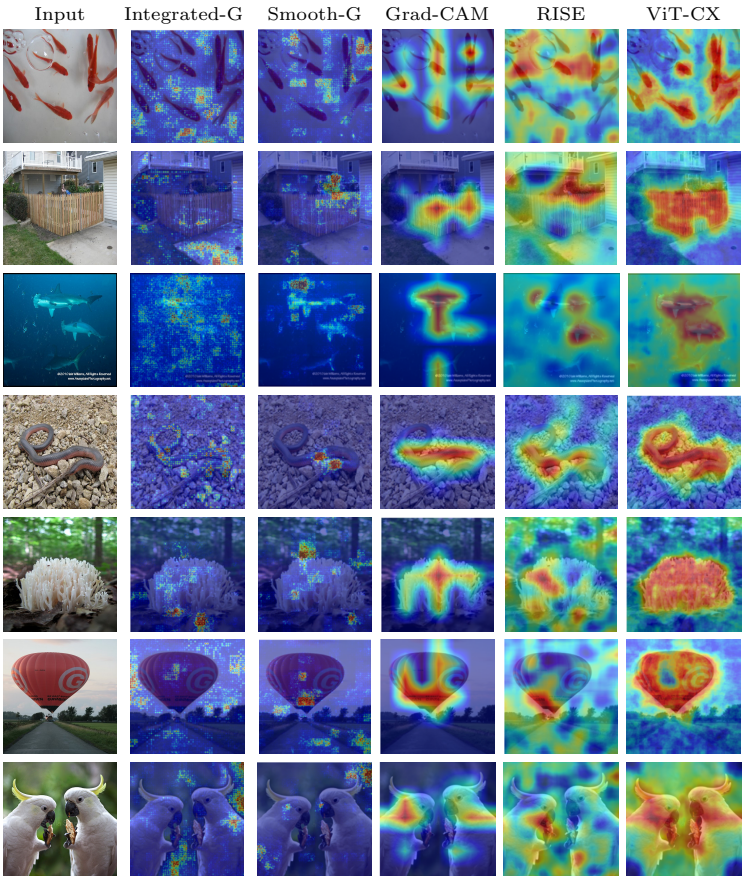


Fig. 1. Sample results of using different XAI methods to explain **Swin-B**. The explained labels are: input(1) - *goldfish* ; (2) - *picket fence* ; (3) - *hammerhead* ; (4) - *thunder snake* ; (5) - *coral fungus* ; (6) - *balloon* ; (7) - *crested cockatoo*.

A.2 Examples from DeiT-B

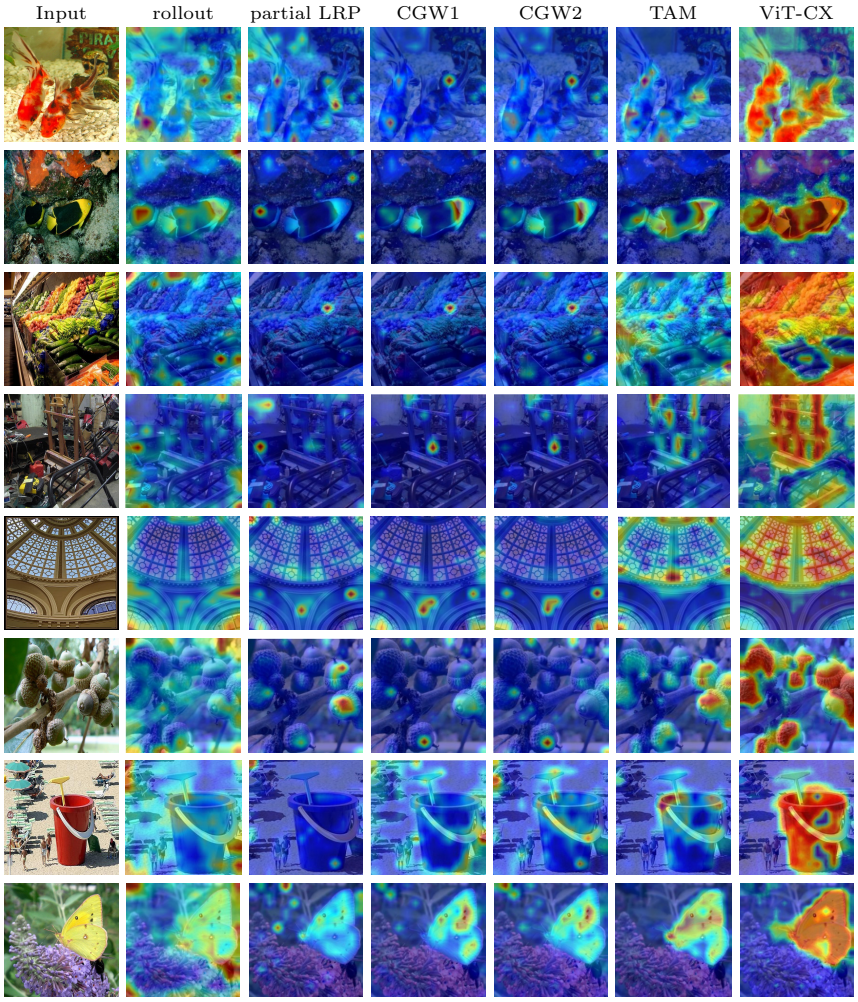


Fig. 2. Sample results of using different XAI methods to explain **DeiT-B**. The explained labels are: input(1) - *goldfish* ; (2) - *rock beauty* ; (3) - *grocery store* ; (4) - *guillotine* ; (5) - *dome* ; (6) - *acorn* ; (7) - *bucket* ; (8) - *sulphur butterfly*.

A.3 Use Cases: ViT-CX on Explaining COVID-ViT

COVID-ViT [8] is based on CT thoracic lung images for the classification of SARS-CoV-2. The structure of COVID-ViT mainly follows that of the vanilla ViT [6], but the attention module is replaced by an efficient attention module [9] in order to speed up the computation. It is worth noting that the efficient attention module does not generate an attention map for each token as the regular attention module does. Thus it is unknown how to apply the attention weights-based explanation methods to explain the classification results of COVID-ViT. In contrast, our ViT-CX method can be easily applied to explain the prediction results given by the COVID-ViT.

We are interested in what features the model relies on to distinguish the COVID patients based on the lung CT images. Here we use the Grad-CAM explanation results as a comparison. Examples are given in Figure 5. As can be seen from those examples, our ViT-CX explanation results localize the suspicious lesion regions that exhibit ground glass opacities while the regions highlighted by Grad-CAM do not show a clear pattern.

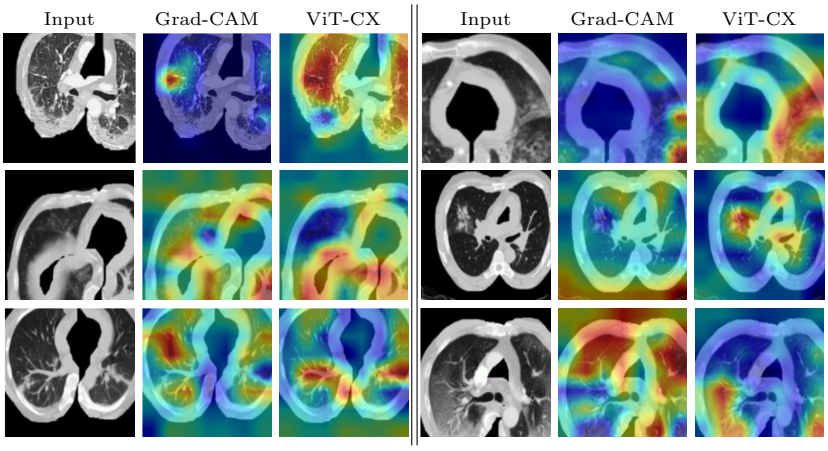
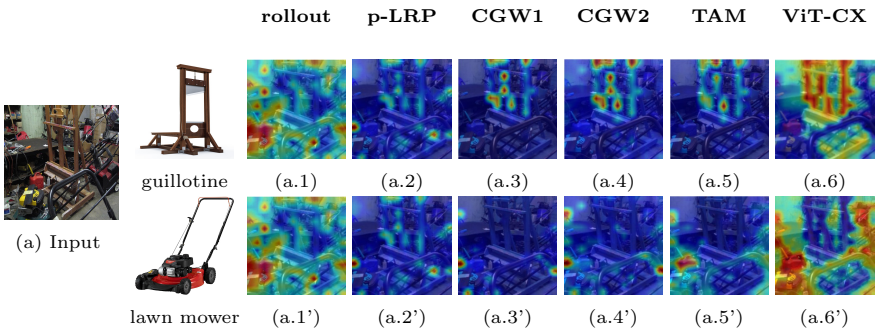


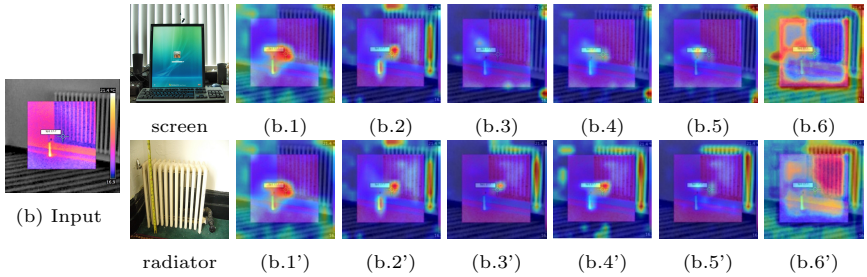
Fig. 3. Explanations for the "COVID" prediction given by the COVID-ViT.

A.4 Use Cases: Explanations to Reveal Why the Model Gives Wrong Predictions

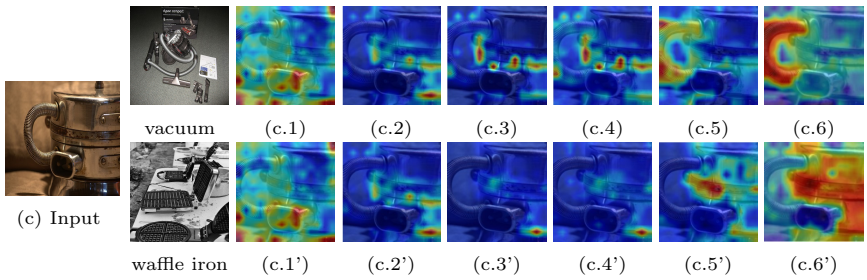
In Figure 4, we show the explanation results for the misclassified images, both for the predicted and ground truth labels. We can see that since rollout and partial LRP are class-agnostic methods, they give the same explanation results regardless of the explained labels, so they do not reveal why the model makes the misclassification. Although CGW1, CGW2 and TAM can make class-specific explanations, the evidence they highlight is incomplete. Using our proposed ViT-CX method to generate explanations with complete evidence highlighted for both the prediction and ground truth labels, users can easily understand why the model fails to classify these images correctly. This can help model users better understand model patterns and provide more hints for AI scientists to improve the ViT models.



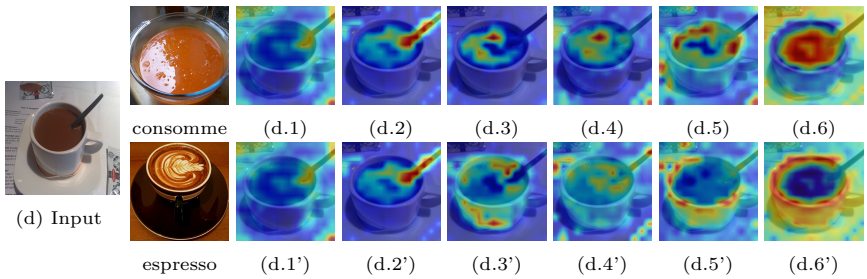
(a.1 - a.6) Explanations for the top-1 prediction label - *guillotine* (23.7%). (a.1' - a.6') Explanations for the ground truth label - *lawn mower* (11.2%)



(b.1 - b.6) Explanations for the top-1 prediction label - *screen* (9.8%). (b.1' - b.6') Explanations for the ground truth label - *radiator* (4.5%)



(c.1 - c.6) Explanations for the top-1 prediction label - *vacuum* (76.7%). (a.1' - a.6') Explanations for the ground truth label - *waffle iron* (17.9%)



(d.1 - d.6) Explanations for the top-1 prediction label - *consomme* (68.3%). (d.1' - d.6') Explanations for the ground truth label - *espresso* (11.4%)

Fig. 4. Examples of using different XAI methods to explain the top-1 prediction label and the ground truth label of the misclassified images (used model: ViT-B; p-LRP is short for partial LRP).

B Sanity Check

We evaluate the sensitivity of ViT-CX to model parameters using the cascading randomization test proposed in [1]. We obtain the explanations obtained by progressively randomizing the parameters of ViT-B/16 from the logit to the penultimate transformer block. We then compute the similarity metrics between the explanations under the original model and those under the perturbed models. Two metrics are considered: Spearman rank correlation and the structural similarity index (SSIM). The results are shown in Table B. We also give examples generated under the perturbed models in Figure 5. As can be seen from the results, the explanation results by our ViT-CX are destroyed when the model parameters are randomized. Therefore, our method is sensitive to the model parameters.

	Rank Correction	SSIM
(a) re-initialized	-0.1631	0.5380
(a)+(b) re-initialized	-0.0557	0.5581
(a)+(b)+(c) re-initialized	-0.0729	0.5561

Table 1. Results of cascading randomization test. The weights of (a). the logits, (b). the last transformer block and (c). the penultimate transformer block are randomly re-initialized, cascade by cascade. We then generate the explanations under the three perturbed model: 1. (a) re-initialized; 2. (a)+(b) re-initialized; 3. (a)+(b)+(c) re-initialized.

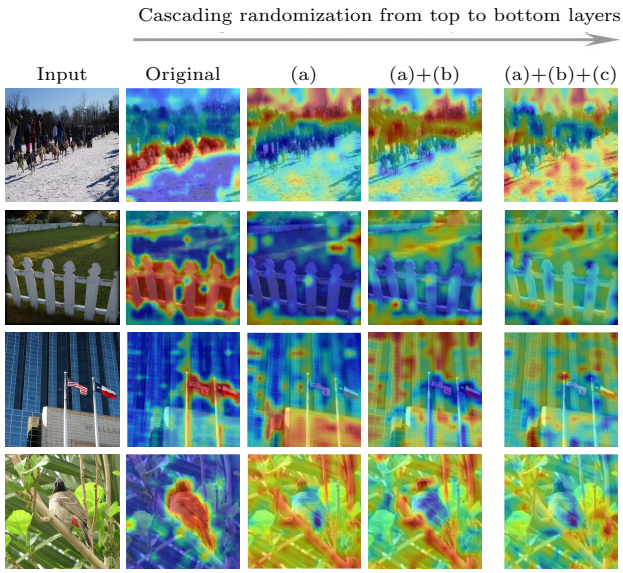


Fig. 5. Sample explanation results under the original model the perturbed models.

C Localization Performance of ViT-CX: Comparison with ViT-based WSOL methods

In addition to using ViT-CX as an XAI tool, we also show that ViT-CX can be used as an object localizer under a weakly supervised object localization (WSOL) setting. WSOL aims to provide object localizations given image-level category labels solely. Several WSOL methods have been proposed [7, 3, 2] and all of them require additional modules based on the ViT image classification models in order to produce semantic-aware localization. The added modules require an extra training process and lead to a decrease in classification accuracy. In contrast, our ViT-CX **does not require any additional training** when a ViT classification model is given. It can achieve comparable localization performance with the previously proposed ViT-based WSOL method.

WSOL Methods and Backbones

We compare the localization performance of ViT-CX with three ViT-based WSOL methods: (a) TS-CAM [7]; (b) LCTR [3]; (c) SCM [2]. In their experiment setting, they use DeiT-S/16 as a backbone, we also run our ViT-CX on the DeiT-S/16 classification model provided by [10]. Since our ViT-CX does not require any model training and can be easily used with larger classification models to obtain prediction bounding boxes; we also run ViT-CX on DeiT-B/16 and DeiT-B/16-Distill.

Dataset and Evaluation Metrics

The evaluation is conducted on the ILSVRC2012 validation set [5], which includes 50,000 images belonging to 1,000 different categories. The commonly used metrics in the WSOL task include top-1 localization accuracy (Top-1 Loc.) and localization accuracy with known ground-truth class (Gt-k.). Specifically, Gt-k. is correct when the intersection over union (IoU) between the ground-truth and the prediction bounding box is larger than 0.5, and does not consider whether the predicted category is correct. Top-1 Loc. is correct when top-1 classification label and Gt-k. are both correct. For each example, the prediction bounding box is obtained by thresholding, and different thresholds are tried. The threshold that gives the best overall performance is adopted.

Experiment Results

The comparison of localization performance between ViT-based WSOL methods and ViT-CX is given in Table 2. We can see that with the same backbone - DeiT-S, the top-1 localization accuracy of ViT-CX (55.5%) is better than TS-CAM (53.4%) and slightly smaller than that of LCTR and SCM (56.1%). With easily applying ViT-CX with the larger and well-trained ViT classification model (DeiT-B, DeiT-B-Distill), higher localization accuracy scores are obtained (56.4% and 59.1%). In terms of localization accuracy with ground-truth known, ViT-CX (65.4% under DeiT-S) is a little worse than the best WSOL method (SCM: 68.8% under DeiT-S). This result fits the nature of ViT-CX as an XAI method to explain what evidences the model uses to predict the top prediction classes.

	Backbone	Top-1 Cls.	Top-1 Loc.	Gt-k.
TS-CAM	DeiT-S	74.3	53.4	67.6
LCTR	DeiT-S	77.1	56.1	<u>68.7</u>
SCM	DeiT-S	76.7	56.1	**68.8**
ViT-CX	DeiT-S	79.8	55.5	65.4
ViT-CX	DeiT-B	<u>81.8</u>	<u>56.4</u>	66.2
ViT-CX	DeiT-B-Distill	**83.4**	**59.1**	67.8

Table 2. Results of comparing ViT-CX with WSOL methods. Top-1 Cls. is the top-1 classification accuracy, Top-1 Loc. is the top-1 localization accuracy and Gt-k. is the localization accuracy with known ground-truth class.

D A Comparison between ViT-CX and ViT Shapley

ViT Shapley [4] (hereafter referred as *ViT-SH*) is another mask-based explanation method for ViTs. Both ViT-CX and ViT-SH generate explanations by determining the importance of image patches. ViT-CX does it via inference, whereas ViT-SH does it by learning a separate model to estimate the Shapley values of individual patches. The estimator is time-consuming to train. Covert et al. trained it only for ImageNette which consists of only **10** ImageNet classes, and it took 0.8 day (as mentioned in Appendix H of their paper). It is unclear how difficult for it to scale up to the **1000** classes of ImageNet. This is why ViT-SH was not originally included in our comparisons.

More importantly, ViT-CX produces better explanations than ViT-SH. Table 3 shows the comparison of them on the 10 classes of ImageNette. We see that ViT-CX clearly outperforms ViT-SH in deletion AUC. Their insertion AUCs are comparable and both close to 1 (maximum possible value).

Figure 6 shows explanations of the two methods on examples from Fig. 12 and 13 of [Covert et al.] without cherry-picking. We can see that ViT-CX performs better than ViT-SH in highlighting objects that correspond to the class being explained, such as the *garbage truck* in the first example. In some examples, ViT-SH assigns high saliency to pixels that are non-relevant to the explained class (e.g., in the first example) or only partially highlights the class-relevant pixels (e.g., only highlighting one of the two *tenches* in the image).

	Del AUC ↓	Ins AUC ↑
ViT-SH	0.691 (0.014)	0.985 (0.002)
ViT-CX	**0.598 (0.016)**	0.981 (0.001)

Table 3. Results of ViT-SH and ViT-CX on ImageNette.

Fig. 6. Visual comparison of ViT-SH and ViT-CX.

E Hyperparameter Sensitivity Analysis of ViT-CX

In Figure 7, we assess the impact of clustering distance threshold δ [1] on the number of masks and explanation quality (measured by deletion/insertion AUC) by testing its values from 0 to 0.3. The results show that 0.1 (used in experiments) is a turning point where further increasing δ does not decrease the number of masks a lot but noticeably decreases the explanation quality. Also, reducing δ does not harm the explanation quality but only increases the explanation time.

In Figure 8, we examine the effect of σ, the standard deviation of the Gaussian noise added to the masked images, on explanation quality by testing its values from 0 to 0.5. The results suggest that the small values of σ (0.05 - 0.2) enhance the explanation quality, while the larger values of σ (above 0.3) can adversely affect explanation quality, which potentially introduces new artifacts in the masked images and results in biased causal impact scores.

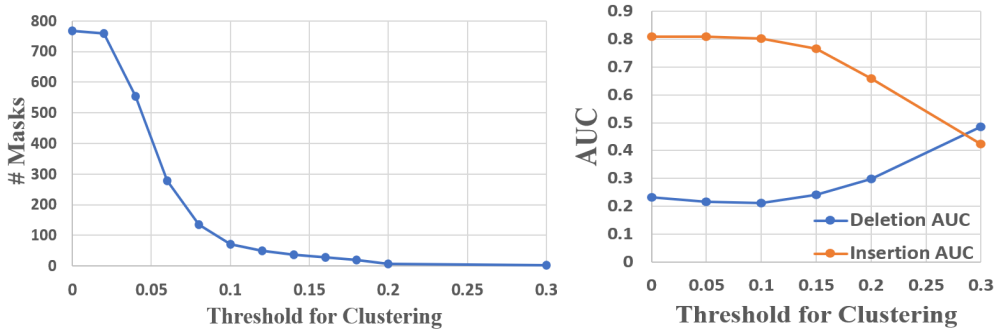


Fig. 7. Relationship between clustering threshold and both the resulting number of masks & explanation quality (based on DeiT-B).

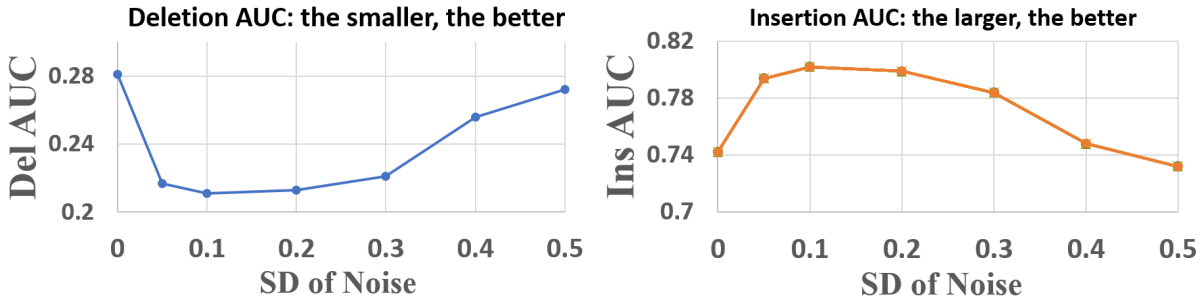


Fig. 8. Impact of the standard deviation (SD) of noise (σ) on the explanation quality (measured by deletion and insertion AUC).

[1] Clusters stop merging if their pairwise distance $> \delta$.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NIPS. pp. 9525–9536 (2018)
2. Bai, H., Zhang, R., Wang, J., Wan, X.: Weakly supervised object localization via transformer with implicit spatial calibration. In: European Conference on Computer Vision. pp. 612–628. Springer (2022)
3. Chen, Z., Wang, C., Wang, Y., Jiang, G., Shen, Y., Tai, Y., Wang, C., Zhang, W., Cao, L.: Lctr: On awakening the local continuity of transformer for weakly supervised object localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 410–418 (2022)
4. Covert, I.C., Kim, C., Lee, S.I.: Learning to estimate shapley values with vision transformers. In: The Eleventh International Conference on Learning Representations (2023)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
7. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2886–2895 (2021)
8. Gao, X., Qian, Y., Gao, A.: Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. arXiv preprint arXiv:2107.01682 (2021)
9. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: WACV. pp. 3531–3539 (2021)
10. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML. pp. 10347–10357. PMLR (2021)