Supplementary Material

In this supplementary manuscript, we showcase (i) qualitative comparison of final localization maps between different weakly supervised object localization (WSOL) methods for a few query images, (ii) intuitive comparison of GAR and LRP across different transformer backbones, iii) comparison of the patch drop mask visualization across the transformer blocks with respect to the base transformer model DeiT-B, (iv) ablation of hyper-parameters chosen for the p-ADL layer.

1. Qualitative Comparison with SOTA methods

Figure 1 shows extensive comparison of localization maps for competing state-of-the-art (SOTA) methods against our proposed method ViTOL-GAR with 20 randomly sampled images from ImageNet dataset. We consider three models to draw comparisons: a) CAM [3] based on VGG-16 backbone, b) ADL [1] trained on ResNet-50 backbone, and c) TS-CAM [2] trained on transformer DeiT backbone. We observe that ViTOL-GAR clearly outperforms competing approaches in its localization ability. Our method generates localization maps which (a) cover the entire region of the object, (b) are class dependent (e.g. Figure 1 (Row Swing, Stethoscope) Set 2)) and (c) invariant to background noise (e.g. Figure 1 (Row Bam Spider Set 1, Cicada Set 2)).

2. Ablation Study

2.1. GAR and LRP for different backbones

Two attention map generation methods, GAR and LRP, have been used in our proposed approach. From the results in main manuscript, we observe that both these methods are effective in localizing the object. In this study, we aim to further understand differences in performance of these methods with different backbone architectures, namely, ViT-B, DeiT-S and DeiT-B. In Table 1, we observe that GAR performs consistently well across all backbones. Whereas LRP observes a drop of 1.59%, 0.78% with ViT-B and DeiT-S backbones respectively on ImageNet dataset.

GAR uses attention maps (attention matrix) of each layer and gradient maps corresponding to same attention map with respect to the desired class to calculate the final localization map. In addition to using these gradient maps, LRP also generates local relevance maps which is calculated based on *local gradients* based on DTD principle in each layer using chain rule. This accumulation of local relevance across layers in LRP may negatively affect performance where embeddings do not effectively represent the regions belonging to the object. This can occur in cases of misclassification. Thus, features from a relatively weak classifier ViT-B may not be as representative of the correct class as compared to a strong classifier (DeiT-B). Intuitively, we believe this could be the reason for inconsistent performance of LRP with relatively weak classifier backbones such as ViT-B, DeiT-S as compared to DeiT-B. Moreover, we observed that LRP takes on average$\approx 2.26x$ *more time for running inference* over GAR while providing similar localization performance.

	ImageNet		
Method	ViT-B$^+$	DeiT-S$^+$	DeiT-B$^+$
GAR	68.14	69.01	69.17
LRP	66.55	68.23	70.47

Table 1. **GAR vs LRP for different transformer backbones**: MaxBoxAccv2 is shown for ImageNet. Superscript (+) denotes the architectures with p-ADL layer

2.2. Study of Patch Drop Mask

In ViTOL, we use the p-ADL layer to enhance the localization capability. Patch drop mask and patch importance map are two key components of the p-ADL layer. Patch drop mask drops the most highlighted patches based on a drop threshold parameter (λ) and forces the model to look at less highlighted patches of the object of interest. However, only dropping the patches degrades the classification ability. Therefore, we use a patch importance map to retain the most highlighted patches to preserve the model's classification ability. We choose patch importance map or patch drop mask randomly at a chosen embedding drop rate (α).

In this study, we show that the p-ADL layer improves the attention in each encoder block progressively. In Figure 2, we draw a comparison between the base transformer model against ViTOL by visualizing the drop mask of several en-

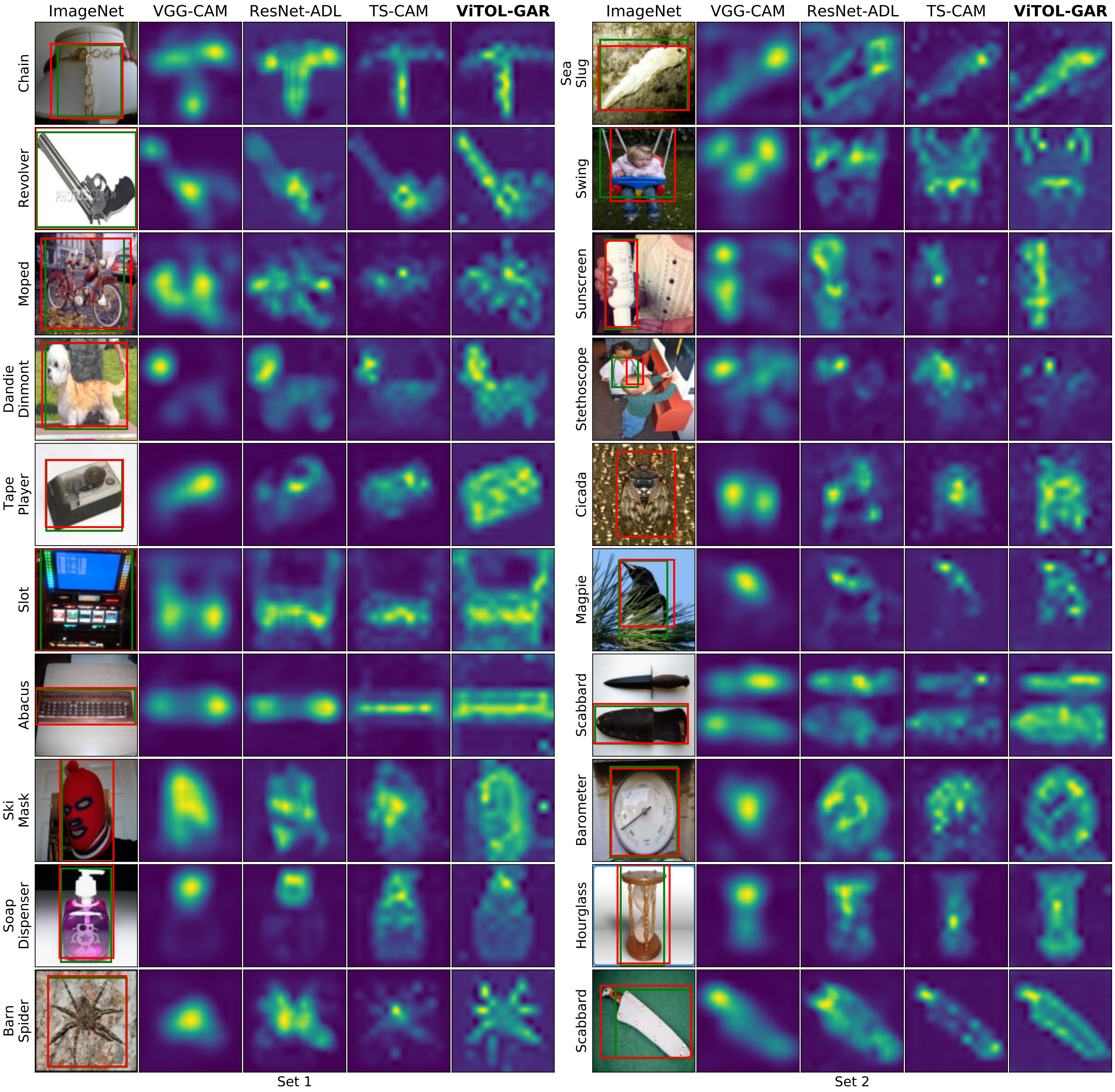

Figure 1. Comparison with state-of-the-art methods on 20 randomly sampled images from ImageNet validation split.

coder blocks. In our experiments, we use $\lambda = 0.9$. Therefore, we drop the patches which have an attention value that is greater than λ times the maximum attention value. Ideally, if the patch drop mask learns to drop the entire object region, it implies that each patch in the object region attains a value in the top 10% of the values in the attention map. Our map generation mechanism uses this information to generate a self-attention map which uniformly attends to each patch covering the object. In Figure 2 (Row 4), we observe that the Blocks 1 to 3 progressively drop the discriminative regions in the attention maps. However, from Blocks 4 and 5, the model potentially starts re-discovering the discriminative patches through the patch importance map. This behavior results in a model which is attentive to both, discriminative patches, as well as other features covering entire object. Thereby resulting in a good quality localization map. Note, in Figure 2, we only show those encoder blocks where the patch drop mask is chosen by the model.

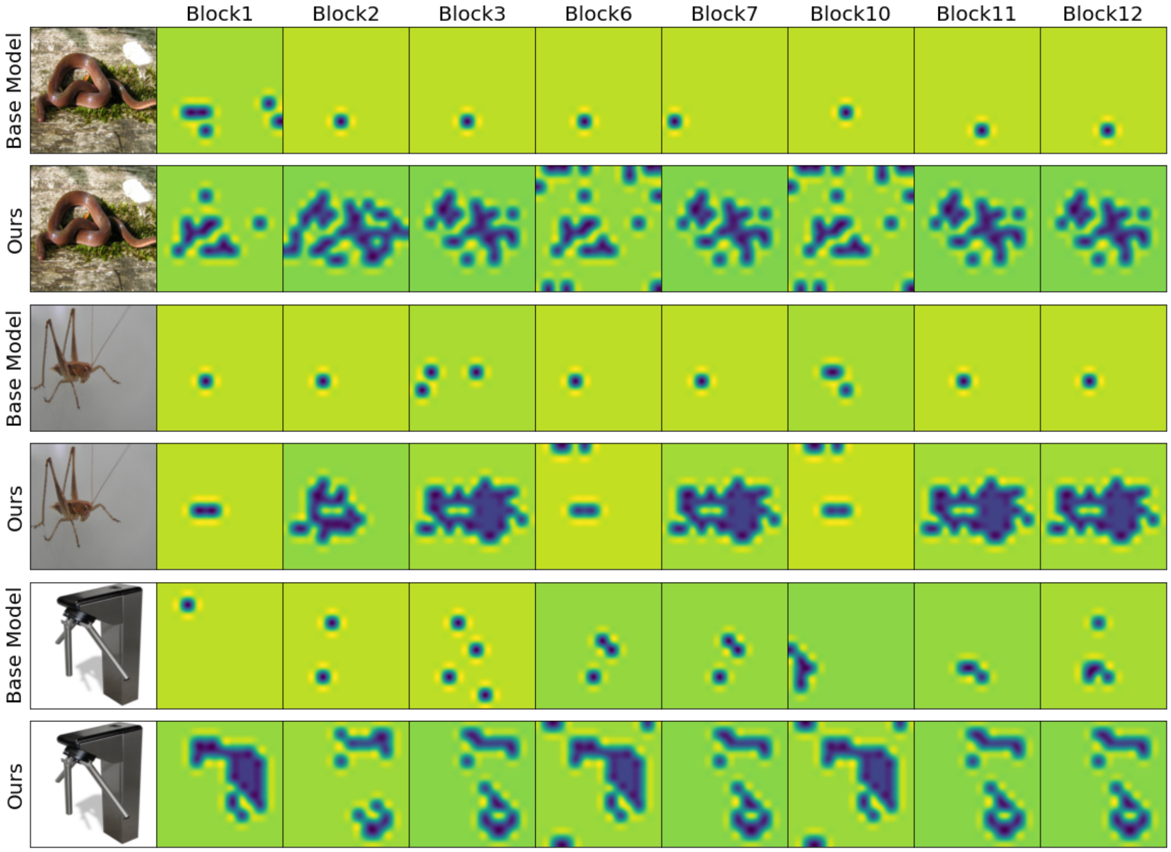

Figure 2. **Patch Drop masks:** Each alternate row shows the comparison of drop masks for some randomly sampled images in p-ADL layers of few encoder blocks. Base model refers to the pre-trained DeiT-B model. Ours is the DeiT-B model trained with p-ADL layers. Image samples are randomly chosen from ImageNet-1k validation split.

In general, in Figure 2 we compare the patch drop mask behavior of the base model against our method. In the base model, the patch drop mask drops a very small portion of the entire object. This indicates that the entire object is not highlighted across different encoder blocks. This results in a poor localization performance. In contrast, the highlighted area in our method increases as the p-ADL layer drops the most discriminative patches and forces the model to focus on other parts of the object as well. In the encoder block 1, the drop region does not cover most of the object of interest. However, as the p-ADL layer drops object specific regions across different encoder blocks, the network focuses on the other parts of the object. This enables the model to focus on the informative pacthes of the object. In some intermediate encoder blocks, patch importance mask is also selected which highlights the most discriminative region. And this discriminative portion is again dropped in subsequent encoder blocks. A similar pattern of highlighting and dropping region is observed across various query images of Figure 2. Block 11 and 12 highlights most of the patches covering the object.

α	λ	MaxBoxAccv2
0.5	0.9	66.37
0.75	**0.9**	**72.42**
0.9	0.9	71.57
0.5	0.8	68.2
0.75	0.8	67.2
0.9	0.8	70.2
0.5	0.7	68.42
0.75	0.7	68.81
0.9	0.7	69.37

Table 2. Ablation of p-ADL parameters for experiments on the CUB dataset.

2.3. p-ADL Hyperparameters Ablation

In Table 2, we study the effect of changing different parameters of p-ADL layer. We vary the embedding drop rate $\alpha \in \{0.5, 0.75, 0.9\}$ and the patch drop threshold $\lambda \in \{0.7, 0.8, 0.9\}$ to show the trend in MaxBoxAccV2 localization metric. In our work, we choose $\alpha = 0.75$ and $\lambda = 0.9$ and use them consistently for all the experiments, as these result in the best localization score.

References

[1] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019. 1

[2] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021. 1

[3] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1