← Back to **Author Console** (/group?id=NeurIPS.cc/2023/Conference/Authors#your-submissions)

SimPLR: A Simple and Plain Transformer for Universal Detection and Segmentation

PDF (/pdf?id=Q6R1B1UMQb)

Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1)

01 May 2023 (modified: 15 Jul 2023) NeurIPS 2023 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=Q6R1B1UMQb)

Keywords: Plain Object Detection, Instance Segmentation, Panoptic Segmentation

Abstract:
This paper presents a simple and plain transformer-based architecture for detection and segmentation, we call SimPLR. Our network enables end-to-end prediction using the original Vision Transformer backbone without the need for complex modifications to accommodate multi-scale feature maps. By incorporating scale information into the attention computation, SimPLR effectively captures objects of different sizes. Remarkably, our study demonstrates that a single-scale feature map is adequate for our network, and with the proposed attention mechanism, SimPLR is capable of learning scale equivariant features. We show through our experiments that SimPLR, as a plain detector, achieves performance on par or better than its multi-scale or hierarchical counterparts. The straightforward end-to-end design of SimPLR is also universal as it yields competitive performance in object detection, instance segmentation, and panoptic segmentation while requiring only a small computational budget. The code will be released.

Supplementary Material: pdf (/attachment?id=Q6R1B1UMQb&name=supplementary_material)
Corresponding Author: d.k.nguyen@uva.nl
Reviewer Nomination: d.k.nguyen@uva.nl
Primary Area: Machine vision
Claims: yes
Code Of Ethics: yes
Broader Impacts: n/a
Limitations: yes
Theory: n/a
Experiments: yes
Training Details: yes
Error Bars: yes
Compute: yes
Reproducibility: yes
Safeguards: yes
Licenses: yes
Assets: n/a
Human Subjects: n/a
IRB Approvals: n/a
Submission Number: 1245

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Program Chairs | Senior Area Chairs | Area Chairs | Reviewers Submitted

Authors | Ethics Reviewers... | Reviewer hh4Z | Reviewer ovX2 | Reviewer 4Lq8

Reviewer Fghh | Reviewer qn8d | ✖

18 / 18 replies shown

Add: **Withdrawal**

Official Review of Submission1245 by Reviewer hh4Z

Official Review Reviewer hh4Z 09 Jul 2023, 06:23 (modified: 01 Sept 2023, 04:49)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer hh4Z

Revisions (/revisions?id=KgElyMPqY7)

Summary:
This work describes a model for object detection and segmentation, based on multiscale sampling of visual transformer features. Rather than using multiscale feature maps (e.g. from CNN or Swin-style transformer, possibly combined with FPN), this method uses only the last layer of a plain ViT backbone upsampled to a single fixed size, and uses multiscale box attention layers to sample and pool features from different-sized areas of a single feature map. This is like a variant of BoxeR, where differently sized reference boxes are now all associated with the same feature map. The paper explores a couple different ways of inserting multiscale sampling into the attention mechanism, as well as ablations on effect of feature map size. Evaluations are performed on COCO detection, instance segmentation and panoptic segmentation, with performance similar or slightly better than BoxeR.

Soundness: 3 good
Presentation: 2 fair
Contribution: 2 fair
Strengths:
Extracting features with box attention over multiple reference sizes in a single feature map is an interesting technique to explore (and one that may be particularly well-suited for ViT, which can incorporate longer-range context without a FPN). The system offers promising performance, particularly for a single-feature-map method. To the degree that it may be similar to BoxeR (see below), it's a new and different actualization that opens a different perspective.

Weaknesses:
While this method appears promising, and indeed uses only one ViT feature map before the detection transformer, many of the mechanisms it describes are paralleled by other mechanisms in BoxeR and other works, and I think pinning these down and seeing exactly how this system is different is still under-explored. It's unclear to me exactly what are the exact effects of the differences.

In particular, multi-scale feature maps and multi-scale reference boxes sampling a single map would be roughly the same if the multi-scale maps were all resized copies of a single source map. So, are there any other key differences or advantages, and if so, what are they and what are their exact effects?

Likewise, "scale equivariance" is mentioned several times, but also not measured. How similar are post-attention values between scales or scale perturbations? And how similar are they if using multiscale feature maps instead?

Questions:

- fixed-scale vs adaptive-scale attention: These appear highly related, if using more heads for fixed scale. How does fixed-scale attention using 32 heads and 4 scales relate to adaptive-scale using 8 heads and 4 scales? Each will attend to a total of 8 x 4 = 32 reference windows. After accounting for head concat combination W_o, can one be expressed in terms of the other? Are they the same, different, or one more general than the other?
- eq. 8: why max(.,0) for width and height deltas --- the use in eq 2,3 shows this is applied additively --- so can the boxes get smaller? or is the scale centered instead on w_i + b_wi, with w_i the min size for the scale? and if there is a min size at the current scale, why not a max size stopping at the next scale up?

- FLOPS measurements: could be interesting to break down FLOPs into those in the backbone, and those in the detector part of the system. while many of the comparable systems have 0.5T flops, if ViT-B dominates this, then any differences in computation between detectors won't really be visible.
- eq. 10 doesn't look right -- addition of m_i_k,j presumably should be inside a softmax if used to mask out areas in the attention by setting logits to -inf
- l.150: h_i used for two purposes, height and feature
- l.249 "well on 1/16 and worse with 1/8" --- the table has this the other way around, better for 1/8 and a little worse with 1/16

Limitations:
--

Flag For Ethics Review: No ethics review needed.
Rating: 5: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

− = ≡

Rebuttal by Authors

Rebuttal
✎ Authors (👁 Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))
📅 09 Aug 2023, 11:42 (modified: 23 Aug 2023, 17:38)
👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Ethics Reviewers Submitted, Authors
📄 Revisions (/revisions?id=XKAc0OXJRj)

Rebuttal:
Thanks for the detailed and constructive comments to help our work, acknowledging our approach `promising`; `a new and different actualization` that `opens a different perspective`. Next, we address **all** concerns and will include them in our updated paper.

Our response to each concern:

1. ***Are there any other key differences or advantages ... what are they and what are their exact effects***:

- In ViTDet, multi-scale feature maps are created by not only resizing the last features of ViT but also applying different sets of convolution layers on these feature maps (illustrated in Figure 4 of our supplementary). Consequently, each feature map in the feature pyramid is responsible for objects of a specific scale during the training process. Moreover, Table 1 of the ViTDet paper shows that predicting objects from a single scale input is challenging (no feature pyramid: 51.2 AP-box and 45.4 AP-mask vs. simple feature pyramid: 54.6 AP-box and 48.6 AP-mask). Notably, even with a specialized design of ViT architecture and the use of multi-scale anchors in the prediction layer, UViT still performs worse than ViTDet. In our work, the adaptive-scale attention allows each feature vector in the single-scale input to learn a different context range in order to deal with objects of various scales. Here, we report the GFLOPs breakdown (backbone and detection head), performance, and inference time of several methods:

	Backbone	Head	GFLOPs - Backbone	GFLOPs - Head	AP_{box}	AP_{mask}	PQ	fps
Swin (IN-21K)	Swin-B	Cascade	~320	~620	54.0	46.5	n/a	13
MViT (IN-21K)	MViT-B	Cascade	~220	~620	55.6	48.1	n/a	11
Mask2Former	Swin-B	Mask2Former	~320	~220	n/a	46.7	55.1	-
ViTDet	ViT-B	Cascade	~350	~700	54.0	46.7	n/a	11
BoxeR+SimpleFPN	ViT-B	BoxeR	~350	~200	55.4	47.7	n/1	12
UViT	UViT-B	Cascade	~510	~600	53.9	46.1	n/a	12

	Backbone	Head	GFLOPs - Backbone	GFLOPs - Head	AP box	AP mask	PQ	fps
SimPLR	ViT-B	SimPLR	~350	~170	55.6	48.0	55.6	15 (25 with jit optimization)

(Mask2Former does not report the inference time with Swin-B in their paper.)

- It can be seen that the convolution-based detection head accounts for a high number of GFLOPs compared to the transformer-based detection head. And, our detection head with SimPLR is more efficient than the multi-scale Transformer-based segmentation model like Mask2Former or BoxeR+SimpleFPN. Moreover, SimPLR requires less memory compared to BoxeR+SimpleFPN (23.3 Gb vs 29.4 Gb) during training. While convolution operation is highly optimized for CUDA kernels, our plain Transformer-based detector, SimPLR, still shows faster inference time in terms of frame-per-second, reaching 25 fps with jit optimization.

2. ***"scale equivariance" is mentioned several times, but also not measured***:

- By saying “scale equivariance”, we mean the ability of the detector to predict objects of multiple scales using the single-scale feature input learnt by our adaptive-scale attention mechanism. In Table 2, we show that simply using BoxeR with the single-scale feature of 1/8 scale (referred as “base”) causes a performance drop compared to BoxeR+SimpleFPN (54.5 vs 55.4 for AP-box, and 46.8 vs 47.7 for AP-mask). Our ablation study suggests that the adaptive-scale attention mechanism improves the performance of the plain detector in object detection and segmentation.

3. ***fixed-scale vs adaptive-scale***:

- The difference between fixed-scale and adaptive-scale attention is in the ability of choosing suitable scale information per query vector in the input feature map. While fixed-scale attention assigns a fixed distribution of scale information to all features in the feature map, adaptive-scale attention generates different scale information based on the content of the query vector. Furthermore, the adaptive-scale attention is more general in which it can replicate the behavior of the fixed-scale attention if necessary. When comparing the fixed-scale attention using 32 heads and 4 scales with the adaptive-scale using 8 heads and 4 scales, one important factor is the dimension of each attention head. For example, if the hidden dimension of adaptive-scale using 8 heads and 4 scales is d, the fixed-scale attention using 32 heads and 4 scales requires $4d$ dimension throughout the detection head in order to keep the same information per attention head.

4. ***eq. 8: why max(.,0) for width and height deltas***:

- Thank you for your correction. In our implementation, the $\max(.,0)$ or ReLU is applied after Eq.(2) to prevent the region of interest, r', from having the negative width and height. So the correct one should be $\Delta_{w_i} = \max(qW_{w_i}^T + b_{w_i}, -1) * w_i; \Delta_{h_i} = \max(qW_{h_i}^T + b_{h_i}, -1) * h_i$. Since r'_i with negative w'_i and h'_i is counter-intuitive, we put this constraint to prevent it from happening. We will correct this equation in the paper.

5. ***FLOPS measurements***:

- Our SimPLR introduces an efficient detection head with less number of FLOPs, as reported in the response 1.

6. ***eq. 10 doesn't look right***:

- We did apply the $m_{i_{k,j}}$ before softmax to mask out areas and will update the equation to reflect that better.

7. ***l.150: h_i used for two purposes, height and feature***:

- Sorry for the confusion. We will separate them with different notations.

8. ***l.249: well on 1/16 and worse with 1/8 -- the table has this the other way around...***:

- Sorry for the confusion. We mean that the proposed method, adaptive-scale attention mechanism, works well on the input feature map of 1/16 scale and performs only worse than the baseline method, box-attention, with the input feature map of 1/8 scale by 0.3 AP (54.2 vs 54.5 for AP-box and 46.6 vs 46.8 for AP-mask). But with the 1/8 scale input feature map, the proposed method, adaptive-scale attention, improves over the baseline method, box-attention, by 0.9 AP (55.4 vs 54.5 for AP-box and 47.6 vs 46.8 for AP-mask). We will discuss this clearly in the section.

Replying to Rebuttal by Authors

response

Official Comment Reviewer hh4Z 17 Aug 2023, 19:46 (modified: 28 Aug 2023, 21:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=m56A0kHwmM)

Comment:

Thanks for your responses. The FLOPS breakdowns are good to see.

My questions on better pinning down the different ways of working with scale are addressed some more, but only partially. I still think this could have been explored further, a view that may be shared by the other reviewers. It also looks like a couple reviewers may not have seen the use of a single feature map as very significant --- I think it is, but the fact that they didn't recognize this highlights that a deeper study could have strengthened the paper.

For example, as I mentioned in my initial review, "scale-equivariance" of features is mentioned by not studied or measured. The definition of this term in the rebuttal as simply increased performance for different-scaled objects is much weaker than a definition of feature equivariance, which would say $f(t(x))$ should be close to $g_t(f(x))$, where t is scaling, f is the neural net (or a portion of one), and g_t is a transformation of features dependent on the transform t. Is this the case for the new scale-adaptive box attention mechanisms? How much and why? Do the ViT's long-range context combinations interact with this formulation of multiscale attention?

Even without this, I still think the paper adequately demonstrates effectiveness of without multiscale feature maps, and using its different multiscale attention mechanisms.

I also agree with reviewer qn8d15 that results with a backbone with supervised pretraining would be useful for comparisons with other systems that used that.

Replying to response

Official Comment by Authors

Official Comment

Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))

18 Aug 2023, 16:35 (modified: 28 Aug 2023, 21:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=mLJvU1vsYN)

Comment:

Thanks for your feedbacks about more discussion regarding the scale-equivariant property. Here, we perform the analysis to provide more understanding about the behaviour of our proposed adaptive-scale attention mechanism.

Specifically, we use COCO val annotations to classify the detected objects into three categories: small (area from 0 to 32^2), medium (area from 32^2 to 96^2), large (area from 96^2 to $1e5^2$). Since our adaptive-scale attention mechanism learns to choose suitable scale information per attention head, we calculate the attention values of the corresponding object queries in each attention head of the last encoder layer and show the scale distribution in three categories:

object size \ window scale	32×32	64×64	128×128	256×256
large	20%	24%	20%	36%

object size \ window scale	32×32	64×64	128×128	256×256
medium	31%	33%	33%	3%
small	57%	37%	3%	3%

The fixed-scale attention mechanism has uniform distribution across all window scales in its attention heads (i.e., the distribution of each window scale in the attention head is always 25%). In contrast, our adaptive-scale attention mechanism learns different scale distribution in its attention heads based on the size of objects. While feature vectors corresponding to *small* objects tend to focus on small window scale in it heads, ones with *medium* objects distribute across first three scales. This may come from the fact that the medium size (area from 32^2 to 96^2) lies between the range of these three scales (32×32, 64×64 and 128×128). Interestingly, the queries corresponding to *large* objects prefer not only largest scale (256×256) but also other scales to make the prediction. We conjecture this behaviour is to capture more fine-grained information for segmentation tasks.

To further analyze the behaviour of the network, we also calculate the average attention distance of feature vectors in the last layer of our ViT backbone which corresponds to the above object queries. Similarly, we divide them into three catergories: small, medium, and large.

object size	**avg. attention distance (pixels)**
large	258
medium	210
small	181

It's interesting to see that the average attention distance of feature vectors is proportional to their object sizes. This suggests that both SimPLR detection head and its ViT backbone learn to focus on regions of interest using plain features. Moreover, the average attention distance of features corresponding to *small* objects is around 181 pixels, showing that our ViT backbone captures long-range dependency and provides more context information for our detection head to make the decision.

Official Review of Submission1245 by Reviewer ovX2

Official Review Reviewer ovX2 05 Jul 2023, 16:41 (modified: 01 Sept 2023, 04:49)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer ovX2

Revisions (/revisions?id=wUMu3AfsxS)

Summary:
This paper propose a simple and plain transformer-based object detector and segmenter called SimPLR, which only use single-scale feature rather than multi-scale feature to do such tasks. Using a proposed scale-aware attention mechanism enables SimPLR to extract scale equivariant features, so that it achieves significant performance compared to the multi-scale or hierarchical counterparts. The main idea of this paper deliver clearly.

Soundness: 3 good
Presentation: 2 fair
Contribution: 2 fair
Strengths:

1. Plain architecture formed by transformer, SimPLR can discard many hand-craft design like NMS, which introduces less bias.
2. Using single-scale feature as input of detection head is memory efficient.

Weaknesses:

1. This paper has less novelty. SimPLR paper follows BoxeR, which follows Deformable DETR. Beyond Deformable DETR, Boxer proposes a box-attention mechanism makes it effective to handle both of 2D and 3D detection scenario. After reading this paper, SimPLR, compared with BoxeR, only set multiple anchors within multiple heads of attention on spatial location of each queries, which is not surprise to me.

2. How does scale-aware attention mechanism can learn scale equivariant features? In my view, both of Deformable DETR and BoxeR have scale-aware attention mechanism, because their attention range is located within object extent in an implicit way.
3. The segmentation task and masked instance-attention may not necessary, because both of them may not contribute to you main idea.
4. This is no ablation study on small objects compared with other methods. Single scale feature may has less local details.
5. The baseline of your work is not clear. In line 224, "The baseline of our study is the ViTDet:...". In line 239, "...we study the baseline which is the original box-attention from [30] with a single-scale feature map...". In addition, under the same time schedule, the performance of SimPLR saw little improvement compared to BoxeR in Table 4.
6. In my opinion, ViTDet only use a simple FPN in RPN with the aim to generate anchors/proposals of small objects. The input feature ViTDet detection head is also the single scale feature from the plain vision transformer, which gives the evidence that single scale feature also works.

Questions:

1. Please make it clear difference of SimPLR with BoxeR.
2. Please make it clear why scale-aware attention mechanism can learn scale equivariant features, which is my main concern. For example give some analysis or figures.

Limitations:
Yes

Flag For Ethics Review: No ethics review needed.
Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal

Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))

09 Aug 2023, 14:48 (modified: 23 Aug 2023, 17:38)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Ethics Reviewers Submitted, Authors

Revisions (/revisions?id=8fa09rvAPg)

Rebuttal:
Thanks for comments appreciating that our `plain architecture` with transformers can `discard many hand-craft design` , and the single-scale input is `memory efficient` . We address concerns below and will incorporate discussions in our updated paper.

Overall, we'd like to highlight that:

- Our goal is to not focus heavily on the architectural design but to explore whether it is possible for a Transformer-based detector to learn from only a single-scale feature map.
- Our simple design of adaptive-scale attention mechanism for the single-scale input leads to competitive performance compared to multi-scale counterparts, while being fast and memory efficient.

1. ***Novelty of the paper***: Due to the character limit, please see our response 1 to Reviewer 4Lq8.
2. ***In my view, both Deformable DETR and BoxeR have scale-aware attention mechanism***

- Theoretically, both deformable attention in Deformable DETR and box-attention in BoxeR have the scale-aware property which can locate dynamic objects in an implicit way. However, it is demonstrated through experiments in Table 2 of the Deformable DETR paper that multi-scale feature maps greatly improve its performance (AP-box increases from 39.7 to 41.4). Moreover, in Table 2 of our paper, we show that simply using BoxeR with the single-scale feature of 1/8 scale (referred as “base”) causes a performance drop compared to BoxeR+SimpleFPN (54.5 vs 55.4 for AP-box, and 46.8 vs 47.7 for AP-mask). Our simple design of adaptive-scale attention mechanism allows the detection head to choose suitable scale information per attention head and boosts the performance on the single-scale input.

3. ***Segmentation task ... may not necessary***:

- We believe segmentation tasks like instance and panoptic segmentation are a great way to show the effect of our plain detector. Specifically, in object detection, the detection head is required to predict low-dimension parameters (i.e., x, y, w, h) of bounding boxes which is an easier task compared to segmentation. However, the instance or panoptic segmentation ask the detection head to give the prediction over all pixels of objects, which is more challenging for a plain detector with the single-scale input feature. Thus, our SimPLR still gives strong results on three tasks.

4. ***There is no ablation study on small objects***: Due to the character limit, please see our response 4 to Reviewer 4Lq8.
5. ***The baseline is not clear.***

- It is noted that the original BoxeR uses a hierarchical, convolutional backbone. In Table 1, we adopt BoxeR with ViT backbone using SimpleFPN in order to have a fair comparison. Specifically, we show that the plain detector, SimPLR, with only a single-scale feature map as input outperforms the plain-backbone detector, ViTDet, and is competitive with BoxeR + SimpleFPN.
- In Table 2, our goal is to ablate the effectiveness of SimPLR with several design choices. Therefore, we use BoxeR with the single-scale input of 1/8 scale as our baseline (referred as 'base'). Table 2 shows that simply using BoxeR with the single-scale feature map causes a performance drop and the simple design of SimPLR improves the performance of the detection head on the single-scale input. We will clarify this better in our paper.

6. ***In my opinion, ViTDet only use a simple FPN in RPN to generate anchors of small objects. The input feature ViTDet detection head is also the single scale feature...***

- The official implementation of ViTDet is released under detectron2: https://github.com/facebookresearch/detectron2/tree/main/projects (https://github.com/facebookresearch/detectron2/tree/main/projects). ViTDet uses the feature pyramid generated by SimpleFPN in both Region Proposal Network and its detection head. Specifically, the object proposals are generated by multi-scale feature maps. Given the generated proposals, the detection head will perform the RoIAlign function on the corresponding feature map in the feature pyramid to predict final bounding boxes and masks. So the feature pyramid plays an important role in ViTDet.

7. ***Please make it clear difference of SimPLR with BoxeR.***

- BoxeR proposes the box-attention and an end-to-end architecture which relies on multi-scale feature maps as input. In addition, BoxeR is proposed for hierarchical backbones (i.e., convolutional networks). In this work, we explore the plain and Transformer-based detector where both backbone and detection head operate on the single-scale input. Notably, we found that our proposed detector, SimPLR, is able to compete with multi-scale counterparts. The motivation of our work comes from (1) the transformer-based detection head with multi-head adaptive-scale attention allows us to encode context of different scales into each feature; (2) the global attention in ViT allows feature vector to learn the long range dependency within the feature map. In Table 2, we show that simply using BoxeR with the single-scale feature of 1/8 scale (referred as “base”) causes a performance drop compared to BoxeR+SimpleFPN (54.5 vs 55.4 for AP-box, and 46.8 vs 47.7 for AP-mask). It is verified through our ablation study that the simple and scale-aware attention mechanism contributes to improve the performance of the plain detector, SimPLR.

8. ***why scale-aware attention mechanism can learn scale equivariant features***

- By saying “scale equivariance”, we mean the ability of the detector to predict objects of multiple scales using only the single-scale feature input learnt by our adaptive-scale attention mechanism. Note that, Table 1 in the ViTDet paper shows that the feature pyramid yields much better performance than no feature pyramid (no feature pyramid: 51.2 AP-box and 45.4 AP-mask vs. simple feature pyramid: 54.6 AP-box and 48.6 AP-mask). Similarly, Table 2 in the Deformable DETR paper suggests that multi-scale inputs greatly improve the detection performance from 39.7 AP-box to 41.4 AP-box. However, our study demonstrates that the feature pyramid may not be necessary in a Transformer-based detector.

➔ *Replying to Rebuttal by Authors*

Official Comment by

Reviewer ovX2

Official Comment Reviewer ovX2 20 Aug 2023, 05:33 (modified: 28 Aug 2023, 21:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=Gj0PbYaDLM)

Comment:

Thank you for the rebuttal. I have read all the reviews and the rebuttal. The provided rebuttal partially addressed the my concerns. The reviewer some more question:

1. Is there an experiment that using multiple features with SimPLR. If the gain is limited, we can say that multiple features are unnecessary. Because the performance gain may come from other modules, the performance is comparable cannot say multiple features are unnecessary.
2. 'scale equivariant features' needs more visualization for feature map comparison to demonstrate the idea。

Replying to Official Comment by Reviewer ovX2

Official Comment by Authors

Official Comment

Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))

20 Aug 2023, 17:21 (modified: 28 Aug 2023, 21:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=FuTUancB5Y)

Comment:

We thank the reviewer for all feedbacks regarding our proposed method. Here, we answer questions raised by the reviewer:

1. ***An experiment that using multiple features with SimPLR***

- In Table 1, we use BoxeR+SimpleFPN as our strong baseline for multi-scale detector because we found that the setting of BoxeR is already an optimal setting to work with multi-scale features for object detection and instance segmentation. It is noted that the modules in Sec. 2.3 (i.e., masked instance-attention) are adopted for panoptic segmentation. As a result, we observed no performance gain between BoxeR and our SimPLR using multi-scale feature maps in object detection and instance segmentation.

	AP$_{\text{box}}$	**AP**$_{\text{mask}}$
BoxeR+SimpleFPN	55.4	47.7
SimPLR+SimpleFPN	55.3	47.7

We think the reason is that the original BoxeR utilizes the feature maps of 4 scales (1/8, 1/16, 1/32, and 1/64) each of which is already assigned an optimal window to predict objects of the corresponding size. Another reason is that feature vectors of one scale in the feature pyramid of BoxeR encoder work with 8 attention heads of its corresponding scale, leading to 32 attention heads in practice. Furthermore, as mentioned above, one downside of BoxeR+SimpleFPN is that it is less memory efficient compared to SimPLR (due to multi-scale features + transformer architecture in its detection head), making it harder when scaling up to larger ViT backbones (i.e., ViT-L or even ViT-H).

2. ***More visualization for feature map comparison***

Since the official comment limits us from posting the visualization, we perform a comprehensive analysis on COCO val set regarding the feature pyramid in the encoder of BoxeR+SimpleFPN. More specifically, we use COCO annotations to classify the detected objects into three categories: small (area 0 from to 32^2), medium (area from 32^2 to 96^2), large (area from 96^2 to $1e5^2$). Since the features in the last BoxeR encoder layer are picked as the object queries, we compute the distribution of object queries from 4-scale feature maps in the encoder corresponding to three categories above (i.e., *small*, *medium*, and *large*):

object size \ feature scale	1/8	1/16	1/32	1/64
small	71%	29%	0%	0%
medium	3%	65%	32%	0%
large	0%	0%	40%	60%

It can be seen that in BoxeR+SimpleFPN, all of object queries corresponding to *small* objects are learnt from the high-resolution feature maps. Notably, even the feature map of 1/32 scale is not used to predict *small* objects. The *medium* objects mostly come from the 1/16 and 1/32 feature maps with a large portion towards the 1/16 scale features. And, the feature maps with large receptive field are responsible for *large* objects.

To sum up, we can see that each feature map in the feature pyramid learns to deal with objects of a specific scale. Our visualization of the BoxeR encoder feature prediction is also consistent to this statistic and we will include them to our paper. In contrast, our SimPLR encoder learns to predict objects of all three categories using only a single-scale feature input.

Official Review of Submission1245 by Reviewer 4Lq8

Official Review Reviewer 4Lq8 03 Jul 2023, 18:15 (modified: 01 Sept 2023, 04:49)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 4Lq8

Revisions (/revisions?id=oFXoZDQwq1)

Summary:
The paper proposes a simple and plain transformer for universal detection and segmentation. The experiments show the effectiveness of the method. However, I am concerned about the motivation of the paper. Why design this kind of structure, and what's the advantage of using a similar strategy in the encoder? In Table 4, there is no obvious performance improvement to BoxeR.

Soundness: 2 fair
Presentation: 2 fair
Contribution: 2 fair
Strengths:
1, The paper is well-written and easy to understand.

2, The extensive experiments show the effectiveness of the method.

Weaknesses:
1, The novelty of this paper is limited. The main contribution is to make the multi-scale predictions into the decoder by the improved box attention.

2, Table 4 shows no obvious performance improvement between SimPLR and BoxeR.

3, Lacking the comparison of the inference time.

4, Lacking the quantitative comparison on small objects because the paper points out that SimPLR gives good predictions on small objects in the caption of Figure 3.

Questions:
1, In my understanding, the difference between fixed-scale and adaptive-scale attention is like the one between anchor settings in one-stage and two-stage detectors.

2, Could the author clarify that the 1/8 scale is better than the 1/16 scale of the input feature map?

Limitations:
N/A

Flag For Ethics Review: No ethics review needed.

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal

Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))

09 Aug 2023, 13:52 (modified: 23 Aug 2023, 17:38)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Ethics Reviewers Submitted, Authors

Revisions (/revisions?id=l81i8emE8F)

Rebuttal:

Thanks for the comments appreciating our writing `well-written and easy to understand`; along with `extensive experiments`. We address concerns below and will incorporate discussions into our updated paper.

1. ***Novelty of the paper***

- First, our goal is to explore whether it is possible for a Transformer-based detector to learn from only a single-scale feature map. Towards this goal, we utilize the plain ViT backbone and design a simple and adaptive-scale attention mechanism in the detection head which learns to choose suitable scale information in each of the attention heads. As a result, our detector (SimPLR) using only a single-scale feature is able to compete with highly-optimized multi-scale counterparts.
- Second, it is a common practice that multi-scale features contribute to improve the performance of detection systems. This observation is shown in various settings from hierarchical backbones (ResNet or SwinTransformer) with DeformableDETR and Mask2Former to plain backbone (ViT) with ViTDeT. For example, Table 1 in the ViTDet paper shows that the feature pyramid yields much better performance than no feature pyramid (no feature pyramid: 51.2 AP-box and 45.4 AP-mask vs. simple feature pyramid: 54.6 AP-box and 48.6 AP-mask). Similarly, Table 2 in the Deformable DETR paper suggests that multi-scale inputs greatly improve the detection performance from 39.7 to 41.4 AP-box. However, our study demonstrates that the feature pyramid may not be necessary in a Transformer-based detector. Specifically, our SimPLR, a plain Transformer-based detector, shows strong results while being more efficient in terms of FLOPs and training memory.
- We believe it is community interest that a pure Transformer-based architecture can remove the need of multi-scale feature input and simplify the detection framework.

2. ***Table 4 shows no obvious performance improvement between SimPLR and BoxeR.***

- Indeed, in Table 4, we indicate that SimPLR using only a single-scale feature map is as good as BoxeR which utilizes multi-scale feature maps. This is noteworthy, as most studies on modern object detectors suggest that multi-scale feature maps are important for detecting multi-scale objects. Specifically, even with the ViT backbone, ViTDet still shows that the feature pyramid is important, and Mask2Former also relies on multi-scale feature maps created from the hierarchical backbone. In addition, SimPLR is more efficient than BoxeR+SimpleFPN in terms of FLOPs and training memory.

3. ***Lacking the comparison of the inference time.***: Due to the character limit, please see the table in our response 1 to Reviewer hh4Z.
4. ***Lacking the quantitative comparison on small objects***

- We show in Table 4 that SimPLR using only a single-scale feature map shows better performance on small objects compared to Mask2Former. We also run the inference of checkpoint provided in ViTDet repo and give more breakdown comparison:

	Backbone	AP$_{\text{box}}$	AP$^{\text{small}}_{\text{box}}$	AP$_{\text{mask}}$	AP$^{\text{small}}_{\text{mask}}$
Mask2Former (5x)	Swin-B	n/a	n/a	46.7	26.1

	Backbone	AP$_{\text{box}}$	AP$_{\text{box}}^{\text{small}}$	AP$_{\text{mask}}$	AP$_{\text{mask}}^{\text{small}}$
SimPLR (5x)	ViT-B	55.4	36.1	47.6	26.8
ViTDet (100 epochs)	ViT-B	54.0	36.5	46.7	27.2
SimPLR (100 epochs)	ViT-B	55.6	37.1	48.0	27.5

- Compared to the plain-backbone detector, ViTDet, our approach shows better performance in detecting small objects. A similar trend is shown when compared to the end-to-end detector, Mask2Former. We will make it more clear in the experiment section.

5. ***the difference between fixed-scale and adaptive-scale attention is like the one between anchor settings in one-stage and two-stage detectors.***

- We would like to emphasize the difference between anchors and our proposed scale-aware attention: In modern detectors, the use of anchors is to generate object proposals and to improve the optimization of bounding box regression in the prediction layer. Given the object proposals, the detection head will perform region of interest pooling on the corresponding feature map in order to predict final bounding boxes and masks. As a result, in modern detectors, the anchors with the feature pyramid demonstrate to be more effective than ones with single-scale features (i.e., Feature Pyramid Network shows to be better than Faster R-CNN with the single-scale input).
- On the other hand, SimPLR employs scale-aware attention mechanisms (i.e., fixed-scale or adaptive-scale attention) to learn better features for detection and segmentation in the encoder. By encoding different scale information into multiple attention heads, we allow each feature to choose a suitable range of context during the training process. For example, the feature vector of a large object will focus on the long range dependency with larger scale information. This results in the fact that our SimPLR using only a single-scale feature map shows competitive performance compared to multi-scale counterparts. Moreover, SimPLR is a more efficient detector compared to multi-scale detectors.

6. ***Could the author clarify that the 1/8 scale is better than the 1/16 scale of the input feature map?***

- We show through our experiments in Table 2(b) that SimPLR with 1/8 scale performs much better than with 1/16 scale in both object detection (+1.2 AP-box) and instance segmentation (+1 AP-mask). This may come from the fact that the feature map of 1/8 scale provides better spatial information to learn fine-grained details of objects for our detection head.

Official Review of Submission1245 by Reviewer Fghh

Official Review ✎ Reviewer Fghh 📅 02 Jul 2023, 11:01 (modified: 01 Sept 2023, 04:49)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Fghh

Revisions (/revisions?id=iz9yXgYR0F)

Summary:
This paper presented a simple and plain framework that uses single-scale feature maps for detection and segmentation, called SimPLR. This paper demonstrated that a transformer-based detector, equipped with a scale-aware attention mechanism, can effectively learn scale-equivariant features. Experiments show good results.

Soundness: 2 fair
Presentation: 4 excellent
Contribution: 2 fair
Strengths:

1. This paper explored the using of single-scale feature maps instead of multi-scale feature maps for detection and segmentation. The idea is interesting.
2. The paper is well-written and easy to follow.
3. Experiments show plain framework can achieve competitive performance compared to multi-scale frameworks.

Weaknesses:

1. The effect of plain structures have been explored in ViTDet [24]. Therefore, the difference should be clearly discussed in the paper.

2. The design choice of the proposed framework is unclear.

Please see details in Questions and Limitations parts.

Questions:

1. Why are plain structures important in detection and segmentation tasks? If plain structures can improve the running speed of whole framework, the running speed (FPS) should be provided in Table 1 and Table 4.
2. From Table 2(b), it is still unclear that why the 1/8 scale is used in the final structure. The 1/4 scale has better mask performance.

Limitations:
As ViTDet [24] has demonstrated the effect of plain detectors, the contributions of this paper may be limited.

Flag For Ethics Review: No ethics review needed.
Rating: 5: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct: Yes

− = ≡

Rebuttal by Authors

Rebuttal
Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))
09 Aug 2023, 14:04 (modified: 23 Aug 2023, 17:38)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Ethics Reviewers Submitted, Authors
Revisions (/revisions?id=SgtNv0ViHI)

Rebuttal:
Thanks for the comments appreciating our writing `well−written and easy to follow`; our idea `interesting`; and our experiments show `competitive performance with plain framework`. We address **all** concerns below and will incorporate discussions into our updated paper.

1. ***The effect of plain structures have been explored in ViTDet***

- ViTDet only examines the plain ViT backbone and the feature pyramid still plays an important role in the detector (Table 1 in ViTDet paper shows that no feature pyramid: 51.2 AP-box and 45.4 AP-mask vs. simple feature pyramid: 54.6 AP-box and 48.6 AP-mask). In our work, we demonstrate for the first time that with a simple and plain design, a Transformer-based detector (SimPLR) can show competitive results with a single-scale feature map throughout its backbone and detection head. This is made possible by the capability of the transformer-based detection head in learning different scale information in different attention heads via the proposed adaptive-scale attention mechanism and the capability of capturing long-range information in the ViT backbone. We examine the finding and demonstrate the strong performance of SimPLR on object detection, instance segmentation, and panoptic segmentation.

2. ***The design choice of the proposed framework is unclear***

- Our aim is to verify whether a plain Transformer-based detector can learn to detect objects of multiple scales from only a single-scale feature map. The motivation of our work comes from (1) the transformer-based detection head with multi-head adaptive-scale attention allows us to encode context of different scales into each feature; (2) unlike local-attention operations in Swin or MViT, the global attention in ViT allows feature vector to learn the long range dependency within the feature map. Moreover, in Table 2, we show that simply using BoxeR with the single-scale feature of 1/8 scale (referred as “base”) causes a drop in the performance compared to BoxeR+SimpleFPN (54.5 vs 55.4 for AP-box, and 46.8 vs 47.7 for AP-mask). It is verified through our ablation study that the simple and scale-aware attention mechanism contributes to improve the performance of the plain detector, SimPLR. Our plain detector is also more efficient in terms of FLOPs and training memory.

3. ***Why are plain structures important in detection and segmentation tasks?***

- Due to the character limit, please see the table in our response 1 to Reviewer hh4Z for the FLOPs and inference time. Our SimPLR shows to be a more efficient detector with the single-scale input.

- Also, SimPLR shows strong results when scaling to ViT-Large backbone or when using ViT pre-trained with another self-supervised learning approach such as BEiTv2. Please see our response 2 to Reviewer qn8d.
- We believe it is community interest that a pure Transformer-based architecture can remove the need of multi-scale feature input and simplify the detection framework.

4. ***From Table 2(b), it is still unclear that why the 1/8 scale is used in the final structure***

- The Table 2(b) shows that the 1/8 scale gives very similar performance compared to the 1/4 scale (only 0.1 AP different in instance segmentation). We pick 1/8 scale as it is much more efficient in terms of both training memory (23.3 Gb with 1/8 scale vs. 35.8 Gb with 1/4 scale) and computational requirements (1/4 scale is ~1.7 times slower than 1/8 scale in terms of training speed). We will clarify our choice of the input feature scale in the paper.

5. ***As ViTDet [24] has demonstrated the effect of plain detectors, the contributions of this paper may be limited.***

- It is noted that ViTDet is a plain-backbone detector where its backbone is a plain ViT. ViTDet, however, still relies on the feature pyramid in order to reach competitive results in object detection and instance segmentation as shown in Table 1 of the ViTDet paper (no feature pyramid: 51.2 AP-box and 45.4 AP-mask vs. simple feature pyramid: 54.6 AP-box and 48.6 AP-mask). In this work, we push a step forward by asking whether the feature pyramid is really needed in a Transformer-based detector. Indeed, with a simple design of adaptive-scale attention mechanism, our detector (SimPLR) demonstrates strong results using only a single-scale feature map in both our backbone and detection head. We believe this is an interesting finding for the community where the feature pyramid is a common practice for object detection and segmentation.

Official Review of Submission1245 by Reviewer qn8d

Official Review Reviewer qn8d 02 Jul 2023, 06:22 (modified: 01 Sept 2023, 04:49)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qn8d

Revisions (/revisions?id=D9zNPJ4zg1)

Summary:
In this submission, the authors propose SimPLR, a simple and plain (i.e., utilizing a single-scale feature map) transformer-based architecture that can tackle bounding box detection, instance segmentation, and panoptic segmentation. To handle the variance of object scales (a critical challenge of object detection), the authors propose to incorporate scale information into the existing box attention [30], encouraging the network to learn scale-equivariant features (on top of the single-scale feature map). Specifically, the authors propose fixed-scale attention (where a few attention heads assigned to the same scale reference window) and adaptive-scale attention (where each attention head is in charge of 4-scale reference windows). The adaptive-scale attention is found to be more effective than fixed-scale attention. To extend SimPLR to panoptic segmentation, the authors propose the masked instance-attention, combining the existing instance-attention [30] and masked attention [8]. A comparable performance to current state-of-the-art methods is obtained on the COCO benchmark.

Soundness: 2 fair
Presentation: 3 good
Contribution: 2 fair
Strengths:

- The authors advocate the "simple and plain" architecture, and thus propose SimPLR for object detection and instance/panoptic segmentation, where most of the existing methods exploit the multi-scale architecture. As a result, the reviewer appreciates the effort to explore a different approach from most of the existing methods.
- The writing is reasonable, as most of the details are provided, including the preliminary background on box-attention.

Weaknesses:

The reviewer appreciates the efforts that the authors attempt to explore "simple and plain" architecture for detection and segmentation. However, the reviewer sees three issues in the paper: novelty, effectiveness, and missing related works. Particularly, the reviewer is not convinced by the effectiveness of the proposed method. More details are provided below:

Novelty:

- The proposed SimPLR is a simple extension of BoxeR [30] (box-attention) with the following modifications: (1) single-scale feature map incorporated with scale-aware box-attention, and (2) masked instance attention (combining instance attention [30] and masked attention [8]).

The novelty is not a big concern to the reviewer, as the reviewer thinks it is OK to extend the existing methods. However, the proposed SimPLR does not bring too much gain over BoxeR and other existing methods (see Effectiveness below), making the proposed method less effective and less appealing.

Effectiveness:

- When comparing to the other methods in Tab. 4, the proposed SimPLR does not bring any noticeable advantage. Specifically, (1) even with MAE pretraining, SimPLR attains the same performance as MViT (55.6 AP^b), (2) SimPLR is only marginally 0.2% PQ better than Mask2Former, and (3) in 5x schedule, SimPLR performs similarly to BoxeR.
- Line 211, the proposed method employs MAE-pretrained backbone, making it hard to fairly compare with other methods (e.g., Mask2Former) that only use ImageNet pretrained backbones, as MAE pretraining brings extra gains to the proposed method.
- In Tab. 4, it is better to (1) also include parameters (or even inference speed) for comparison, and (2) further split each method into "backbone" (plain vs. multi-scale backbone), "neck" (FPN, SimpleFPN, or N/A for single-scale feature map), and "head" (Cascaded R-CNN, DETR, Mask2Former, and so on). Tab. 5 in the supplementary Materials should be placed in the main paper and further refined.

Missing related works:

- There are a few related works that are not cited nor discussed in the paper:
  - Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation, CVPR 2023 (was on arXiv on 06/06/2022). Mask DINO also proposed a unified architecture for detection and segmentation.
  - MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers, CVPR 2021 (was on arXiv on 12/01/2020). MaX-DeepLab first proposed to use object queries to generate segmentation masks (via dot product with pixel features).

Typo:

- Fig. 1 caption 4th line: Where -> While

======= Post-rebuttal review ======

The reviewer appreciates the idea of using a “simple and plain” vision transformer architecture for object detection, and instance/panoptic segmentation. The provided rebuttal (along with the rebuttal for reviewer hh4z) addressed most of the reviewer’s concern. As a result, the reviewer raises the rating to Borderline Accept.

Why not higher?

The reviewer shares the same concern with other reviewers that current writing cannot well-reflect the advantage of proposed SimPLR over existing ViTDet. A careful analysis (e.g., the scale-equivariant property of SimPLR requested by the Reviewer hh4z) and more experimental results (e.g., requested in this review and other reviews) may help improve the presentation.

Why not lower?

The reviewer likes the idea of using a “simple and plain” ViT backbone for dense prediction tasks, which is not only simple (avoids the standard multi-scale feature network) but also enjoys the benefits of recent advancements for ViT (e.g., Masked Image Modeling pretraining).

Questions:

A few more questions about the experiments:

- Line 94: 2x2 feature grid is sampled. How about $m \times m$ feature grid, where m = 3, 5? Is the model sensitive to grid size?

- Line 165: top-scored feature vectors will be initialized as object queries. What if one simply samples feature vectors as object queries?
- The experiments can be more complete, if the authors experiment with more backbones (e.g., ViT-L) and on more datasets (e.g., Cityscapes). Currently, the authors only report ViT-B on COCO with limited experimental results.

Limitations:
The reviewer thinks the authors have adequately addressed the limitations.

Flag For Ethics Review: No ethics review needed.
Rating: 5: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal
Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))
09 Aug 2023, 14:18 (modified: 23 Aug 2023, 17:38)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Ethics Reviewers Submitted, Authors
Revisions (/revisions?id=ktbBZhVdEb)

Rebuttal:
Thanks for the careful review and detailed comments, noting our writing `reasonable`, our model `simple and plain`, and our motivation of exploring `different approach`. We address all concerns next and will update our paper accordingly.

Overall, we'd like to highlight that:

- Our goal is to not focus heavily on the architectural design but to explore whether it is possible for a Transformer-based detector to learn from only a single-scale feature map.
- Our simple design of adaptive-scale attention mechanism for the single-scale input leads to competitive performance compared to multi-scale counterparts, while being fast and memory efficient.

1. ***Novelty of the paper***: Due to the character limit, please see our response 1 to Reviewer 4Lq8.
2. ***Effectiveness of the proposed method***

- To have a better comparison, we report the FLOPs breakdown (backbone and detection head) along with the inference time. Please see the table in our response 1 to Reviewer hh4Z for the FLOPs and inference time. Our SimPLR shows to be a more efficient detector with the single-scale input.
- We would like to point out that the backbone in MViT and Swin are all pre-trained with ImageNet-21K which contains many more images and labels. In addition, MViT and Swin backbones are hierarchical architecture which is highly optimized for multi-scale detectors. Therefore, with only a single-scale input feature, it is still challenging for SimPLR to detect objects of multiple scales. To further confirm the effectiveness of SimPLR, we report the performance of several methods on COCO object detection and instance segmentation when scaling to the large backbone. The Mask2Former + Swin-Large is trained with the batch size of 16 and 100 epochs, resulting in a much larger number of iterations, 740K iterations. All other methods are trained with the batch size of 64 and 100 epochs. It is noted that SimPLR and ViTDet utilize ViT-Large backbone pre-trained only on ImageNet-1K. Our SimPLR continues to outperform other approaches in both AP-box and AP-mask metrics.

	Backbone	Training epoch,iteration	AP $_{box}$	AP $_{mask}$	fps
Swin	Swin-Large (21K, sup)	100 epochs, 180K iterations	54.8	47.3	10
MViT	MViT-Large (21K, sup)	100 epochs, 180K iterations	55.7	48.3	6

	Backbone	Training epoch,iteration	AP $_{\text{box}}$	AP $_{\text{mask}}$	fps
Mask2Former	Swin-Large (21K, sup)	100 epochs, 740K iterations	n/a	50.1	6
ViTDet	ViT-Large (1K, MAE)	100 epochs, 180K iterations	57.6	49.8	7
SimPLR	ViT-Large (1K, MAE)	100 epochs, 180K iterations	58.2	50.4	9 (13 with jit optimization)

- Another advantage of SimPLR is that it can benefit from the significant progress of self-supervised learning with ViT. Here, we show that SimPLR generalizes to other pre-training approach, such as BEiTv2:

	AP$_{\text{box}}$	AP$_{\text{mask}}$	PQ
Mask2Former (sup, 5x)	n/a	46.7	55.1
ViTDet (MAE, 100 epochs)	54.0	46.7	n/a
SimPLR (MAE, 5x)	55.4	47.6	55.3
SimPLR (BEiTv2, 5x)	55.7	48.2	56.3

3. ***Missing related works***

- We will cite and include the discussion about these works in the manuscript. Technically, Mask DINO incorporates queries denoising into the Mask2Former framework which is orthogonal to our method and could be complementary. Here, we focus on learning effectively from a single-scale input with a plain Transformer-based detector.

4. ***2x2 feature grid is sampled ... Is the model sensitive to grid size?***

- Our attention mechanism is not sensitive to the grid size. In the decoder, we sample 14x14 feature grids to capture more fine-grained information of objects. In the encoder, we choose the 2x2 feature grid to keep the training and inference efficient due to the large number of features in the feature map (1/8 scale of 1024 image contains 16384 features).

	AP$_{\text{box}}$	AP$_{\text{mask}}$	Training memory
m=2	55.4	47.6	23.3 Gb
m=3	55.5	47.6	27.5 Gb

5. ***top-scored feature vectors will be initialized as object queries. What if one simply samples feature vectors as object queries?***

- Our goal is to maximize the recall of object queries and top-scored feature vectors allow us to have the high recall (since the scoring function is just a simple projection, its computational complexity is small). We set up an experiment during the inference where we sample a center vector of 7x7 non-overlapping grids in a feature map of 128x128 features, resulting in ~324 object queries. The performance drops from 55.4 to 23.0 for AP-box and from 47.6 to 20.6 for AP-mask. One reason could be that the decoder is not trained for uniform spatial sampling, causing distribution shift in the decoder. Another reason is that when the area contains lots of objects, specially with small objects, the uniform sampling cannot give high recall compared to the sampling from a scoring function. This is also similar to two-stage detectors such as Mask R-CNN or Deformable DETR.

Replying to Rebuttal by Authors

Thanks for the rebuttal

Official Comment Reviewer qn8d 16 Aug 2023, 02:03 (modified: 28 Aug 2023, 21:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=aR7J1XUgiL)

Comment:
The reviewer thanks the authors for the rebuttal. The reviewer has read all the reviews and the provided rebuttal.

The provided rebuttal partially addressed the reviewer's concerns. The reviewer has one more question:

As detailed in the 1st and 2nd items of Effectiveness in **Weaknesses** section, is it possible that the authors can provide results using ImageNet-21K pretrained ViT backbone results to more fairly compare with other methods? Or, any other comment on the necessity of using MAE pretrained backbone would be very appreciated.

Replying to Thanks for the rebuttal

Thanks for the constructive feedbacks

Official Comment

Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))

16 Aug 2023, 11:51 (modified: 28 Aug 2023, 21:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=2I8cNsnDO4)

Comment:
We really appreciate all feedbacks from the reviewer to improve and verify our proposed approach. We are running the training with supervised pre-training ViT models and will show it as soon as we get the results.

About the comment on the necessity of using MAE pre-trained backbone, we demonstrate in our rebuttal that the proposed approach can also generalize well with other self-supervised pre-training methods such as BeiT. Regarding to the MAE pre-training, several works [1,2] have indicated that the MAE pre-trained ViT learns to focus on local region with more local attention pattern. While learning local pattern is easier for Swin Transformer or MViT due to the use of local attention and down sampling, it is more challenging for ViT with global self-attention and plain features.

[1] Park et al. What Do Self-Supervised Vision Transformers Learn? in ICLR 2023.

[2] Chen et al. Efficient Self-supervised Vision Pretraining with Local Masked Reconstruction. in arXiv:2206.00790

Replying to Thanks for the constructive feedbacks

Official Comment by Authors

Official Comment

Authors (Duy Kien Nguyen (/profile?id=~Duy_Kien_Nguyen1), Martin R. Oswald (/profile?id=~Martin_R._Oswald1), Cees G. M. Snoek (/profile?id=~Cees_G._M._Snoek1))

18 Aug 2023, 17:59 (modified: 28 Aug 2023, 21:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=sSbWbL9kGV)

Comment:
Here, we show the results with supervised pre-training ViT models using DeiT (Touvron et al.). Due to the time limit and resource constraints, we report the performance in object detection and instance segmentation with ViT-B backbone. We will update more results with ViT-L backbone in the next version of our paper. Here, we summarize our results with ViT-B and ViT-L so far:

	Backbone	Training schedule	AP$_{\text{box}}$	AP$_{\text{mask}}$	frame-per-seconds
Swin	Swin-B (21K, sup)	100 epochs, 180K iterations	54.0	46.5	13
MViT	MViT-B (21K, sup)	100 epochs, 180K iterations	55.6	48.1	11
Mask2Former	Swin-B (1K, sup)	50 epochs, 370K iterations	n/a	46.7	-
ViTDet	ViT-B (1K, MAE)	100 epochs, 180K iterations	54.0	46.7	11
BoxeR+SimpleFPN	ViT-B (1K, MAE)	50 epochs, 370K iterations	55.4	47.7	12
UViT	UViT-B (1K, self-training)	60 epochs, -	53.9	46.1	12
SimPLR	ViT-B (1K, sup)	50 epochs, 370K iterations	54.0	46.5	15 (25 w/ jit)
SimPLR	ViT-B (21K, sup)	100 epochs, 180K iterations	55.3	47.6	15 (25 w/ jit)
SimPLR	ViT-B (1K, BEiTv2)	50 epochs, 370K iterations	55.7	48.2	15 (25 w/ jit)
SimPLR	ViT-B (1K, MAE)	50 epochs, 370K iterations	55.4	47.6	15 (25 w/ jit)
SimPLR	ViT-B (1K, MAE)	100 epochs, 180K iterations	55.6	48.0	15 (25 w/ jit)

	Backbone	Training schedule	AP$_{\text{box}}$	AP$_{\text{mask}}$	frame-per-seconds
Swin	Swin-L (21K, sup)	100 epochs, 180K iterations	54.8	47.3	10
MViT	MViT-L (21K, sup)	100 epochs, 180K iterations	55.7	48.3	6
Mask2Former	Swin-L (21K, sup)	100 epochs, 740K iterations	n/a	50.1	6
ViTDet	ViT-L (1K, MAE)	100 epochs, 180K iterations	57.6	49.8	7
SimPLR	ViT-L (1K, MAE)	100 epochs, 180K iterations	58.2	50.4	9 (13 w/ jit)

To sum up, we see that the self-supervised learning methods for ViT such as MAE or BEiT help to improve the capability of the backbone, leading to better performance. The same behaviour has been observed in UViT and ViTDet where both papers need to adopt MAE pre-trained ViT or self-training UViT in order to enable plain-backbone detectors. As mentioned in the comment above, we think the self-supervised learning methods help ViT to learn better local pattern, which is challenging with plain features.

However, our single-scale detector, SimPLR, with supervised pre-training ViT-B backbone still shows very competitive performance compared to MViT, Swin, and Mask2Former. Furthermore, SimPLR shows a good scaling behaviour from ViT-B to ViT-L compared to Swin or MViT (especially, MViT improves only ~0.1 AP point from MViT-B to MViT-L). We also add more analysis regarding the scale-equivariant property of SimPLR (please see our reply to Reviewer hh4Z).

Replying to Official Comment by Authors

Thanks for the supervised-setting results

Official Comment Reviewer qn8d 19 Aug 2023, 06:51 (modified: 28 Aug 2023, 21:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers Submitted

Revisions (/revisions?id=dYsqFKTL2M)

Comment:

The reviewer sincerely thanks the authors for the extra efforts to provide the supervised-setting results. Additionally, the scale-equivariant property of SimPLR requested by the Reviewer hh4z is indeed interesting.

The reviewer has no more questions at the moment, and will carefully take into consideration the rebuttal and other reviewers' comments, when making the final decision.