KEY PATCH PROPOSER: KEY PATCHES CONTAIN RICH INFORMATION

Jing Xu [1*], **Beiwen Tian**[2] **& Hao Zhao**[2]
Nanjing University of Aeronautics And Astronautics[1], Tsinghua University[2]
jing.xu@nuaa.edu.cn, tbw23@mails.tsinghua.edu.cn, zhaohao@air.tsinghua.edu.cn

ABSTRACT

In this paper, we introduce a novel algorithm named Key Patch Proposer (KPP) designed to select key patches in an image without additional training. Our experiments showcase KPP's robust capacity to capture semantic information by both reconstruction and classification tasks. The efficacy of KPP suggests its potential application in active learning for semantic segmentation. Our source code is publicly available at: https://github.com/CA-TT-AC/key-patch-proposer.

1 INTRODUCTION AND RELATED WORK

Masked Auto-encoder (MAE) (He et al., 2022) has emerged as a potent strategy for representation learning. The strategy boosts the encoder's ability to produce non-trivial representations for images by dividing an image into non-overlapping patches, randomly masking the majority of the patches, and optimize the weights of the encoder so as to reconstruct the masked patches. MAE effectively primes the backbone network by learning textures within patches and interconnection between patches, improving the adaptability for a variety of downstream tasks.

One particular downstream task is active learning (Mittal et al., 2023). In the history of active learning for 2D images, the acquisition of annotations is performed at various scale, e.g., the full image (Sinha et al., 2019), superpixels (Cai et al., 2021), polygons (Mittal et al., 2019) (Golestaneh & Kitani, 2019), or even pixels (Shin et al., 2021).

Following the patch division in MAE, we consider the active learning at the patch scale for images and we propose a novel non-learning based patch proposal method, referred to as *Key Patch Proposer* (KPP). The proposed patches, i.e., the key patches, hold significant potential for enhancing active learning in semantic segmentation tasks. The detailed task formulation is in Sec. 2.

Patch proposal fundamentally represents a submodular function maximization problem (SFMP) (Lovász, 1983) (Feige et al., 2011), which is NP-hard even in the absence of constraints. Details on the correlation between KPP and SFMP can be found in Appendix A. Deep learning methods are considered as solutions to this NP-hard problem, but the experiments suggest otherwise. In this paper, the proposed KPP is inspired by greedy search, a common polynomial-time approximation algorithm for SFMP, to obtain a sub-optimal solution for this inherently challenging issue. Buschjäger et al. (2021) has proved that greedy algorithm is one of the best solutions of SFMP in terms of performance level.

2 METHOD

Firstly, we resize an image into a fixed size (224×224) and then divide it into patches, following He et al. (2022). We define the problem of patch proposal as follows: given a set of patches P, we aim to find a subset of patches $P^* \subseteq P$ with $|P^*| = r|P|$ to minimize the L2 error L of the image reconstructed using patch set P^*. Formally, we aim for a patch set P^* that satisfies:

$$P^* = \underset{\{P_s \subseteq P, |P_s| = r|P|\}}{\arg\min} L(P_s) \tag{1}$$

where $L(P_s)$ represents the reconstruction error between the ground-truth masked patches $P \setminus P_s$ and the masked patches predicted by a pretrained MAE with the unmasked patches P_s. $r \in (0, 1) \subset \mathbb{R}$ is a hyper parameter that controls the ratio of proposed patches, and $|P|$ denotes the cardinality of P.

[*]This work is done during Jing Xu's internship at Institute for AI Industry Research, Tsinghua University.

We propose a greedy search method to solve this problem. The algorithm starts with an initial set P_0 which contains only the central patch. Then, the algorithm iteratively adds the patch that minimizes the reconstruction error. With i being the iteration index and also number of patches in P_i^*,

$$\begin{aligned} p_i^* &= \underset{p \in P \setminus P_{i-1}^*}{\arg\min}\ L(P_{i-1}^* \cup \{p\}) \\ P_i^* &= P_{i-1}^* \cup \{p_i^*\} \end{aligned} \tag{2}$$

The algorithm terminates when the number of patches in P_i^* reaches $r|P|$. Pseudo codes are provided in Appendix B.

3 Experiments

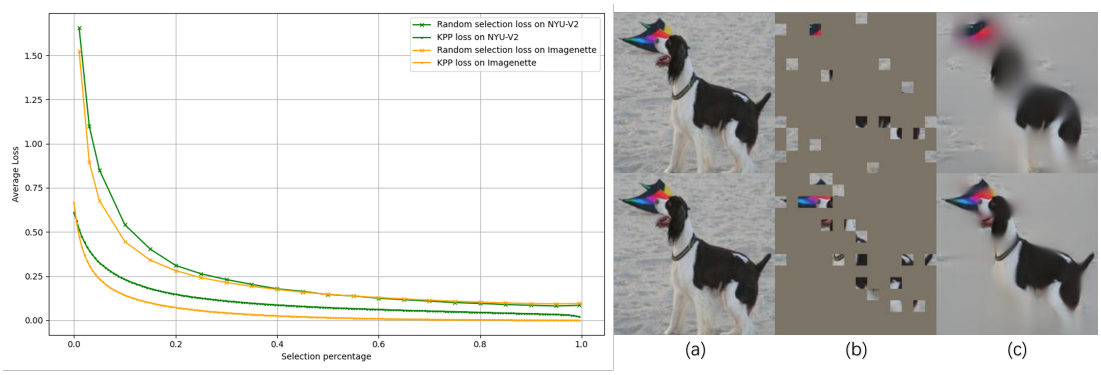

Figure 1: Left: Comparative analysis of reconstruction loss between KPP and random patch selection across various selection percentages. Right Section: The upper row demonstrates the random selection method, while the lower row depicts the KPP algorithm's approach. (a): ground truth; (b): selected patches; (c): reconstructed images utilizing the chosen patches.

We conduct experiments on ImageNette dataset (Howard, 2019), and NYU Depth V2 dataset (Nathan Silberman & Fergus, 2012) (See Appendix C for detail). We use the pretrained ViT-B/16 model provided by He et al. (2022) and the reconstruction loss is the mean squared error between the original masked patches and the reconstructed masked patches. We compare the performance of KPP with random patch selection. As demonstrated in the left chart of Figure 1, we can see that KPP achieves lower reconstruction loss than random patch selection across various selection percentages. In the right section of Figure 1, we visulize one case of random patch selection and KPP, with more shown in Appendix E. Qualitative results demonstrate that KPP selects informative and representative patches of the original image.

Selection percentage	5%	10%	25%	50%	100%
KPP acc	92.26	96.25	98.47	98.93	99.11
Random acc	89.53	95.08	98.32	98.88	

Table 1: Classification accuracy comparison of ViT-B/16 model on imagenette dataset.

We also conduct quantitative experiments on ImageNette (Howard, 2019). During both training and validation stages, we use KPP or random strategy to select input patches, and then finetune the pretrained ViT-B/16 on the selected patches only. Performances with and without KPP are reported in Table 1, indicating that KPP surpasses random patch selection in terms of classification accuracy across various selection percentages, especially with small percentage of unmasked patches.

4 Conclusion

In this work, we propose a novel algorithm, KPP, to select key patches within an image. Through our experiments focused on reconstruction and classification, we demonstrate the potential of KPP to identify key patches and quantify the semantic information they contain. We envisage KPP's application in active learning, a possibility we aim to further explore in future research.

URM STATEMENT

All authors meet the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Sebastian Buschjäger, Philipp-Jan Honysz, Lukas Pfahler, and Katharina Morik. Very fast streaming submodular function maximization. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 151–166, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86523-8.

Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10983–10992, 2021. doi: 10.1109/CVPR46437.2021.01084.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Uriel Feige, Vahab S. Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011. doi: 10.1137/090779346. URL https://doi.org/10.1137/090779346.

S.Alireza Golestaneh and KrisM. Kitani. Importance of self-consistency in active learning for semantic segmentation. *British Machine Vision Conference,British Machine Vision Conference*, Dec 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jeremy Howard. Imagenette, 2019. URL https://github.com/fastai/imagenette. Accessed: 2023-12-01.

L. Lovász. *Submodular functions and convexity*, pp. 235–257. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983. ISBN 978-3-642-68874-4. doi: 10.1007/978-3-642-68874-4_10. URL https://doi.org/10.1007/978-3-642-68874-4_10.

Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with illusions about deep active learning. *Cornell University - arXiv,Cornell University - arXiv*, Dec 2019.

Sudhanshu Mittal, Joshua Niemeijer, JörgP. Schäfer, and Thomas Brox. Best practices in active learning for semantic segmentation. Feb 2023.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: semantic segmentation with pixelpick. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1687–1697, 2021. doi: 10.1109/ICCVW54120.2021.00194.

Samrath Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Sep 2019. doi: 10.1109/iccv.2019.00607. URL http://dx.doi.org/10.1109/iccv.2019.00607.

A EXPLAINING ON REDUCTION OF PATCH PROPOSAL PROBLEM TO SUBMODULAR FUNCTION MAXIMIZATION PROBLEM

In mathematics, a submodular set function (also known as a submodular function) is a set function that, informally, describes the relationship between a set of inputs and an output, where adding another input has a decreasing additional benefit.

If Ω is a finite set, a submodular function is a set function $f : 2^{\Omega} \rightarrow \mathbb{R}$, where 2^{Ω} denotes the power set of Ω, which satisfies one of several conditions. One of the conditions is that for every $X, Y \subseteq \Omega$ with $X \subseteq Y$ and every $x \in \Omega \setminus Y$, we have that $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$. The purpose of this section is to prove that we can define a submodular function f in the reconstruction task that satisfies the condition above.

In the reconstruction task, we define set Ω as all the patches in a image with positional information. X is some selected patches from Ω. We define f as the mean squared error of the whole image but not of the masked patches. Formally, we define f as follows:

$$f(X) = (\text{MAE}(X) - Sort(\Omega))^2 \tag{3}$$

where $Sort(\Omega)$ is the sorted patches in Ω by their positional information. It can be easily verified that f satisfies the condition above. Suppose $X \subseteq Y$ and $x \in \Omega \setminus Y$ is a new patch towards both X and Y. x always provides more semantics for the reconstruction of Y than X, since Y is a superset of X. Therefore, $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$.

By doing such verification, we can use polynomial-time approximation algorithms for submodular function maximization problem to solve the patch proposal problem.

B ALGORITHM OVERVIEW

In this section, we provide pseudo codes to explain our proposed KPP algorithm briefly. n_{keep} in line 1 represents the number of patches to be kept. We initialize L_{min} with a large constant to ensure that it will be updated. In the for loop between line 5 and 10, we figure out the next best patch p_i^* towards P_{i-1}^* and then update P_i^* accordingly.

Require: set P with all patches in, proposal ratio r, empty set P_1^*
1: $n_{keep} \leftarrow r|P|$
2: **for** i from 2 to n **do**
3: $\quad L_{\text{min}} \leftarrow$ MAXN
4: $\quad$ **for** p in $P \setminus P_{i-1}^*$ **do**
5: $\quad\quad L_{cur} \leftarrow \text{Loss}(P_{i-1}^* \cup \{p\})$
6: $\quad\quad$ **if** $L_{cur} < L_{min}$ **then**
7: $\quad\quad\quad p_i^* \leftarrow p$
8: $\quad\quad\quad L_{min} \leftarrow L_{cur}$
9: $\quad\quad$ **end if**
10: $\quad$ **end for**
11: $\quad P_i^* \leftarrow P_{i-1}^* \cup p_i^*$
12: **end for**
13: **return** P_n^*

Algorithm 1: Pseudo-code of KPP algorithm.

C DATASET OVERVIEW

ImageNette dataset (Howard, 2019) is a subset of ImageNet (Deng et al., 2009) with 10 classes distributed in 9469/3925 training/validation images.

The NYU-Depth V2 dataset (Nathan Silberman & Fergus, 2012) is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. It contains 1,449 RGB-D samples where 795 image-depth pairs are used to train the RGB-D model, and the remaining 654 are utilized for testing.

In the scope of our experiment, we exclusively utilize the RGB data from the NYU-Depth V2 dataset and restrict our experimental activities to its training set. This limitation is guided by the requirement in active learning systems for patch selection exclusively during the training phase. Our model, having been pretrained on ImageNet-1k, does not demonstrate zero-shot learning capabilities when deployed on ImageNette. Therefore, to validate its zero-shot learning potential, we employ the NYU-Depth V2 dataset. This approach substantiates the pretrained model's capability for direct application to other downstream datasets without necessitating further training.

D Ablation Study on Initial Patch

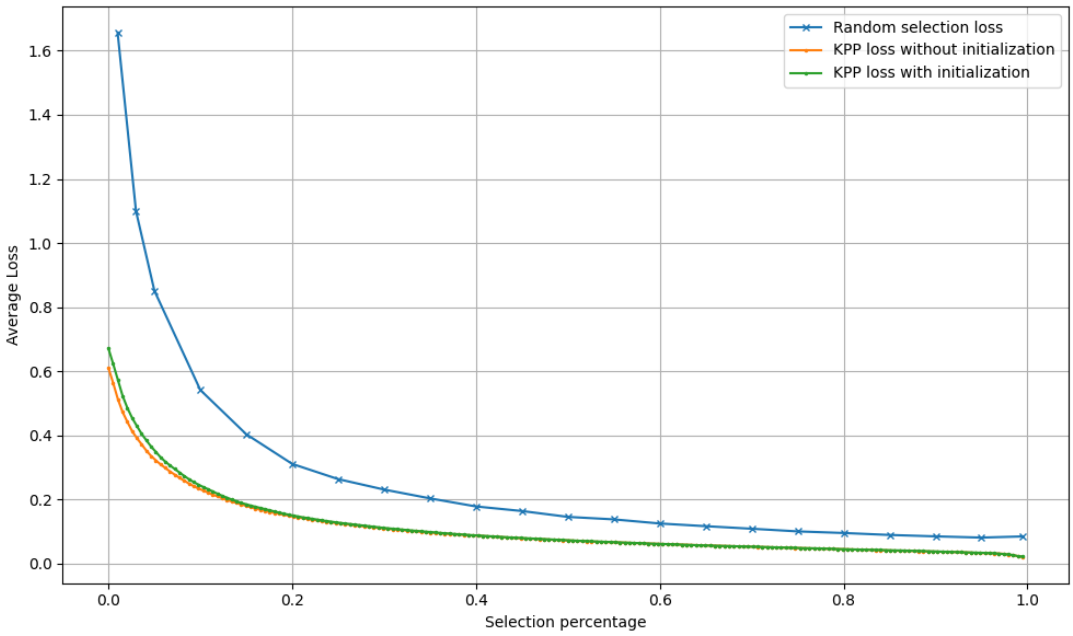

Figure 2: Ablation study of reconstruction loss with/without initial patch.

We present an ablation study on the choice of the initial patch, depicted in Figure 2. This study was conducted using the training set of NYU-V2, akin to the scenario illustrated in the left chart of Figure 1. Notably, our results demonstrate a clear distinction: the loss incurred without this initial patch selection is obviously lower compared to when this initialization step is included, particularly noticeable when the selection percentage is below 0.1.

E Qualitative Results for KPP and random selection

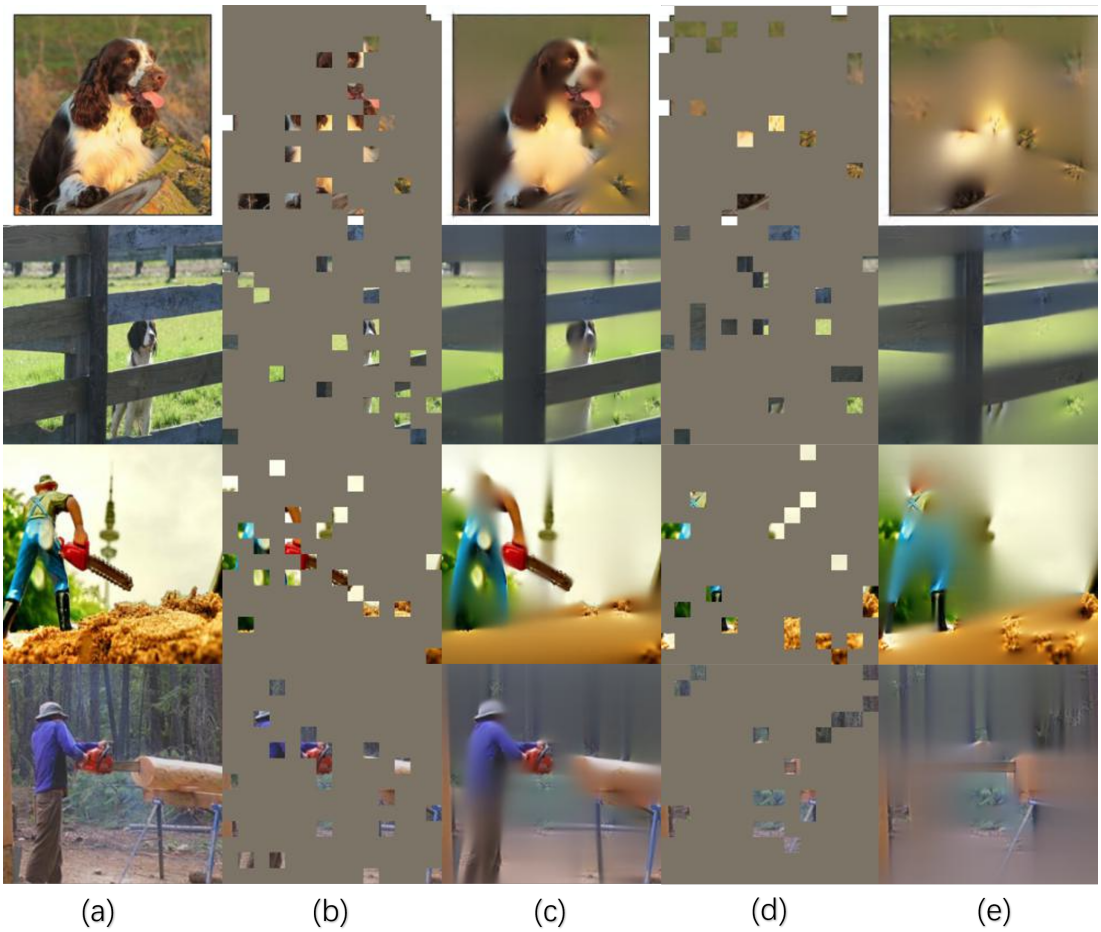

Figure 3: (a): Original images; (b): 10% patches selected by KPP; (c): Images reconstructed by KPP patches. (d): 10% patches selected randomly; (e): Images reconstructed by random selected patches.