Learning to Rank Patches for Unbiased Image Redundancy Reduction

Yang Luo[1], **Zhineng Chen**[1*], **Peng Zhou**[2], **Zuxuan Wu**[1], **Xieping Gao**[3], **Yu-Gang Jiang**[1]
[1]School of Computer Science, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University
[2]University of Maryland, College Park
[3]College of Information Science and Engineering, Hunan Normal University
yangluo21@m.fudan.edu.cn, {zhinchen, zxwu, ygj}@fudan.edu.cn,
pengzhou@terpmail.umd.edu, xpgao@hunnu.edu.cn

Abstract

*Images suffer from heavy spatial redundancy because pixels in neighboring regions are spatially correlated. Existing approaches strive to overcome this limitation by reducing less meaningful image regions. However, current leading methods rely on supervisory signals. They may compel models to preserve content that aligns with labeled categories and discard content belonging to unlabeled categories. This categorical inductive bias makes these methods less effective in real-world scenarios. To address this issue, we propose a self-supervised framework for image redundancy reduction called **L**earning **t**o **R**ank **P**atches (LTRP). We observe that image reconstruction of masked image modeling models is sensitive to the removal of visible patches when the masking ratio is high (e.g., 90%). Building upon it, we implement LTRP via two steps: inferring the semantic density score of each patch by quantifying variation between reconstructions with and without this patch, and learning to rank the patches with the pseudo score. The entire process is self-supervised, thus getting out of the dilemma of categorical inductive bias. We design extensive experiments on different datasets and tasks. The results demonstrate that LTRP outperforms both supervised and other self-supervised methods due to the fair assessment of image content. Code is available at https://github.com/irsLu/ltrp.*

1. Introduction

Image redundancy reduction refers to discarding the less informative regions of an image while preserving its essential semantics. It is a fundamental task in computer vision as compact image representation is an attractive property in various applications, especially in the era of vision Transformer (ViT) [14, 29]. The computational complexity of ViT grows quadratically with the number of image patches, making patch-level image redundancy reduction gaining increasing attention.

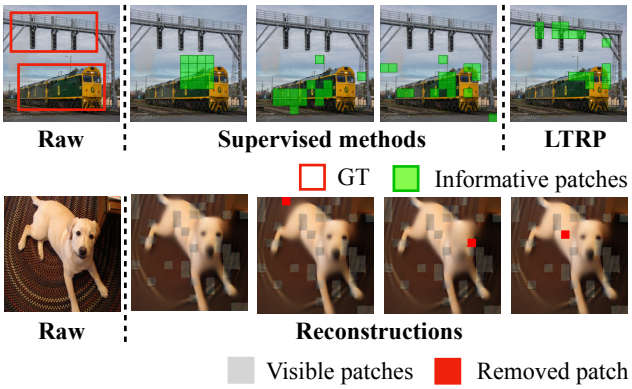


Figure 1. **Redundancy reduction** (upper): given an image, LTRP selects informative patches regardless of whether they are from the categories already learned. While for supervised methods (from left to right: GFNet [41], Grad-CAM [38] using ViT, EViT [24]) the preserved patches are mostly located on the learned category. **Patch removing** (bottom): an image and its reconstructions using MAE [20] with different sets of visible patches, columns 3-5 denote that the *red* patch is removed from the visible set. They generate reconstructions with different levels of semantic shift, e.g., without the dog tail, eyes, etc.

Current leading methods [12, 31, 41, 48] generally require supervised learning to endow the model with the capability to recognize informative image patches. Specifically, the model is simultaneously trained to estimate the patch's informativeness and classify the image to the right category. Thus, the learning process tends to select image patches that are less likely to degrade the classification accuracy, leading to categorical inductive bias. The bias may cause the model to remove image patches containing mean-

*Corresponding author: Zhineng Chen

ingful semantics that do not belong to the learned category. As shown in the upper of Fig. 1, supervised methods trained based on ImageNet-1K tend to select informative patches located on *train*, a category learned. While considering patches on unseen categories like *traffic light* as redundant and removing them. Note that such categorical inductive bias is unavoidable for models trained based on famous image datasets [13, 17, 25], which typically annotate one or a few categories for each image. On the other hand, activation mapping of popular unsupervised pre-training methods [6, 32] also reveal the informativeness of image patches to some extent. Nevertheless, they are less effective as the pre-training is not targeted for redundancy reduction.

Recently, masked autoencoder (MAE) [20] has gained increasing attention due to its great representative learning capability. Given an image with a portion of patches randomly masked, MAE learns the feature representation by reconstructing the masked patches. It can generate a reconstruction with meaningful semantics by using a small subset (e.g., 25%) of visible patches. We observe that when the masking ratio is high (e.g., 90%), further removing a visible patch may cause an obvious semantic shift in the reconstructed image. As shown in the bottom of Fig. 1, if the only visible patch on the dog's tail is masked, the reconstruction will not have the tail. Conversely, masking a patch on the dog's body does not induce perceptible semantic alteration. The result suggests that the informativeness of patches can be revealed from such an unsupervised perspective.

Motivated by this observation, we propose learning to rank patches (LTRP), a self-supervised framework to fairly reduce image redundancy. It firstly leverages a pre-trained MAE to generate a pseudo score for each visible patch by quantifying the semantic difference between reconstructions with and without that patch. A larger pseudo score indicates that the patch has a greater impact on the quality of the reconstructed image and is thus more valuable. Therefore, LTRP utilizes another ranking model to learn to rank the visible patches according to the pseudo scores, which successfully learns this scoring property. After pre-training, by ranking the patches descendingly according to their scores and keeping the top ones, LTRP can remain valuable patches for the subsequent downstream tasks like classification. As shown in the top right image of Fig. 1, patches containing meaningful semantics are unbiasedly preserved under the same keep ratio.

We conduct extensive experiments to validate the unbiasedness of LTRP in image redundancy reduction. For image classification at different keep ratios, LTRP not only achieves competitive accuracy compared to current leading methods in single-label classification, but also outperforms them by a clear margin in multi-label classification. When inspecting categories not included in ImageNet-1K, patches remained by LTRP are also more consistent with object detection and semantic segmentation labels, getting superior performance in all the evaluated metrics. Moreover, if a small ranking model is employed, we observe that an approximate linear speedup with the keep ratio is coincidentally obtained for image classification, which constitutes a promising solution for efficient ViT. Contributions of this paper are summarized as follows:

- We propose to treat the image redundancy reduction as a self-supervised patch-level ranking problem. It eliminates the reliance on supervisory signals, expanding the means to reduce image redundancy.
- We develop LTRP. It quantifies the variation between MAE-reconstructed images to create pseudo labels, then learns to rank these patches. It offers an elegant way to avoid categorical inductive bias.
- Extensive quantitative and qualitative evaluations demonstrate that LTRP can fairly and effectively score the learned and unseen categories, and provide highly competitive solutions for efficient ViT.

2. Related Work

2.1. Image Redundancy Reduction

While sharing similarities with tasks like salient object detection [16, 21] and image compression [33], image redundancy reduction has been studied mainly in the era of deep learning. Some CNN-based methods explored this task for inference efficiency [12, 41, 48]. Meanwhile, ViT models have been implicitly endowed with the redundancy reduction capability. Based on how this capability is built, we can broadly categorize existing studies into three classes.

Class activation mapping (CAM): Starting from a label, CAM [38, 50] was proposed to provide visual explanations for CNNs by combining feature maps from the penultimate layer using learned or gradient-based weights. CAM achieved success in various visual tasks, such as weakly-supervised object detection [11] and semantic segmentation [1, 18], which regards the activation as initial pseudo object labels. CAM can also be applied to ViT. However, it relies on supervisory signals to build the localization ability, thus suffering from the categorical inductive bias.

Self-supervised activation mapping: Recent studies in self-supervised learning (SSL) shows that the capability to perceive fine-grained objects can be obtained from contrastive learning-based pre-training. For example, object locations can be roughly inferred by using MoCo [9] to train a CNN and visualizing activation mapping of the last convolutional layer [32]. A similar capability is also observed from the heads of DINO [6]. Object localization is a natural property learned from SSL. However, as redundancy reduction is not the central goal for SSL, preserving image patches according to the generated activation mapping generally leads to worse performance in downstream tasks.

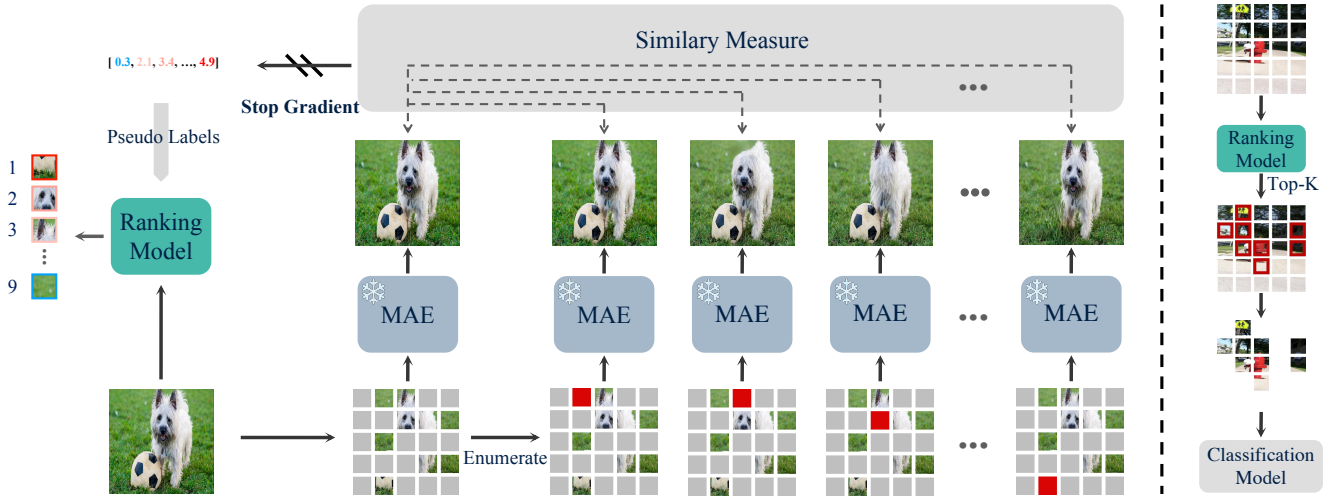


Figure 2. **LTRP training (left):** given an image, LTRP randomly selects a set of visible patches using a high masking ratio (e.g. 90%). A pre-trained MAE (parameter frozen) is applied to get its reconstruction, i.e., the anchor image. Then, the visible patches (red) are removed one by one with replacement, each time generating a new reconstruction and its semantic density score w.r.t the anchor image. The scores are treated as pseudo labels to train the ranking model using learning to rank. Once trained, the ranking model is preserved and the MAE is discarded. **LTRP inference (right):** Top-k patches are selected using the trained ranking model and fed into downstream tasks.

Token reduction: Since the computational complexity of ViT is quadratically linear to the number of tokens [14], reducing tokens, which can be viewed as a special form of image patches by using token-to-patch mapping, is an effective method for efficient ViT. Current studies can be classified into two categories. First, token pruning that discards redundant tokens. It involves retraining a network to dynamically filter tokens [34], using the policy gradient to select patches [31], reducing the number of tokens as inference proceeds [49], etc. Second, token merging that combines similar tokens. It involves leveraging attention scores between the class token and other tokens to gradually merge tokens [24], merging tokens between two subsets with equal size by using similarity of their key vectors from the self-attention layer [2] and others [35, 46]. Typically, the research focus of these methods is on efficient ViT. Unbiased redundancy reduction is rarely considered.

2.2. Learning to Rank

Learning to rank (LTR) [26] is a technique widely employed in information retrieval tasks such as search [8] and recommendation [22]. LTR primarily involves three types according to the utilization of document labels: the point-wise method [7, 23], which uses the rank of an individual document as label and formulates the learning as either classification or regression problem; the pair-wise method [3, 4], which utilizes the relative order of two documents as label for learning; and the list-wise method [5, 44], which considers the order relationship among all or a part of documents associated with a given query. Some studies [27, 47] have made meaningful attempts by leveraging learning to rank in computer vision tasks. To our knowledge, it has not been used for image redundancy reduction before.

LTRP is an application of LTR to image redundancy reduction. The query is the randomly masked image. Each document is associated with a visible patch, and whose pseudo score is the document label. Different from most existing solutions where the label is obtained from manual annotation or human-machine interaction, our label is obtained from an elegantly designed pre-task, thus enabling a novel self-supervised solution as described below.

3. Method

An overview of our learning to rank patches (LTRP) is depicted in Fig. 2. Its core elements can be summarized using two words: infer and rank. Given a pre-trained MAE model and an image with a portion of patches randomly masked, LTRP first *infers* the semantic density score for each visible patch by quantifying variation between the reconstructions with and without that patch. By repeating this process patch-by-patch, we obtain a series of scores. Then, LTRP *ranks* these patches by treating their scores as pseudo labels using learning-to-rank algorithms. When the training converges, we use the ranking model to select informative patches and only consider them in downstream tasks.

3.1. Preliminary

MAE [20] learns a robust image representation by first masking a portion of image patches and then recovering the original image, namely the mask-then-prediction task.

Specifically, MAE divides an image I into visible and masked patch sets according to a predefined masking ratio, where the visible set $\mathbf{P}_v = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n]$ has n patches. Then, the pre-training target can be written as:

$$\min_{\theta} H(\mathcal{T}(I_m), I_m^R), I_m^R = \mathcal{M}_\theta(\mathbf{P}_v) \quad (1)$$

where I_m, I_m^R and $\mathcal{M}_\theta$ denote the masked image, the reconstructed image and MAE parameterized by θ, respectively, $\mathcal{T}$ denotes a certain transformation (optional, e.g., HOG [42], normalized pixel values of each masked patch [20], etc.), and H is a distance measurement function.

3.2. Patch Informativeness Inference

To infer the patch informativeness, we adopt a higher masking ratio of 90% rather than 50% or 75% used in previous MAE studies [20, 45]. The reason is twofold. First, it enables reconstructions exhibiting perceptible change even when only one visible patch is further removed. Second, it decreases the likelihood of semantically similar patches being visible simultaneously, which may confuse the model regarding their informativeness.

Then, we use a pre-trained, parameter-frozen MAE to perform the reconstruction. For an image, the visible set $\mathbf{P}_v$ is stochastically selected and its reconstruction I_m^R, or saying, the anchor image, is obtained, we iteratively remove a visible patch $\mathbf{p}_i \in \mathbf{P}_v$ to obtain $\mathbf{P}_v^i = \mathbf{P}_v \backslash \mathbf{p}_i$. It can be simply achieved by using masked attention [40]. By reconstructing again, we get $I_m^{R_i}$, the image reconstruction without the i-th visible patch. A similarity-based function $\mathcal{F}$ runs over the pair $< I_m^R, I_m^{R_i} >$, and the quantitative similarity y_i is calculated by:

$$y_i = \mathcal{F}(I_m^R, I_m^{R_i}), I_m^{R_i} = \mathcal{M}_\theta(\mathbf{P}_v^i) \quad (2)$$

We term the inferred result as the semantic density score, as it quantizes the semantic variant between the two reconstructions. We simply use the ℓ_1 distance to measure the variant in this paper. By polling all the visible patches one-by-one, we get a pseudo score vector $\mathbf{y} = [y_1, y_2, ..., y_n]$. Note that the scores are obtained in a self-supervised manner, and Eq. (2) can be regarded as the pre-task to reveal the patch informativeness. In the implementation, the score inference can be parallelized for training acceleration.

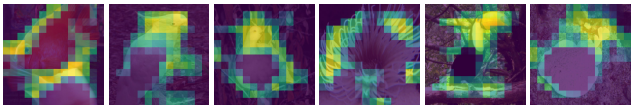


Figure 3. Normalized score maps generated by our ranking model.

3.3. Learning to Rank

With pseudo score $\mathbf{y}$ calculated from Eq. (2), we use learning-to-rank algorithms to train an independent ranking model to rank the patches. Formally, each image-partition instance $\mathbf{x} =< I, \mathbf{P}_v, \mathbf{y} >$ can be viewed as a training sample for the ranking model, which accepts $\mathbf{x}$ as the input and outputs a vector $\mathbf{s} = [s_1, s_2, ..., s_n]$ containing predicted scores for the patches. our LTRP utilizes a list-wise loss, ListMLE [44], to train the model as follows.

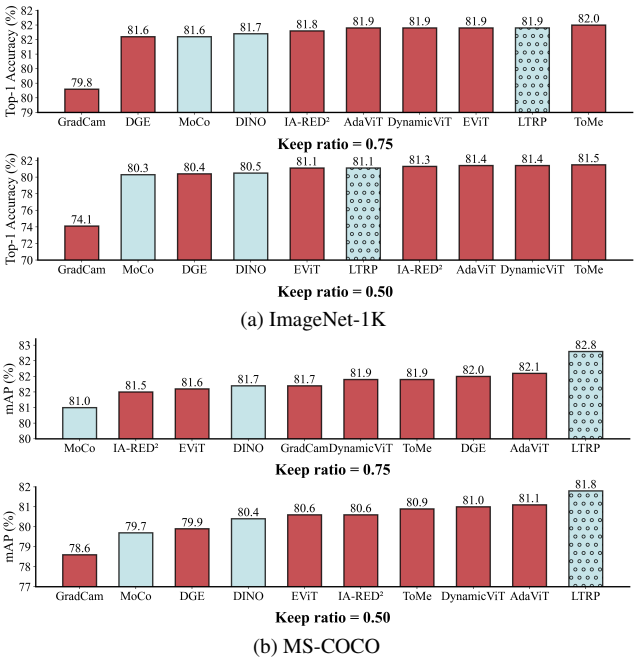


Figure 4. **Classification results** on single-label dataset (upper, ImageNet-1K) and multi-label dataset (bottom, MS-COCO) at different krs. All methods differ only in the patch selection step. Supervised and self-supervised methods are depicted using different colors. LTRP is annotated spots.

$$L = -logP_s(\pi), P_s(\pi) = \prod_{i=1}^{n} \frac{exp\left(s_{\pi(i)}\right)}{\sum_{k=i}^{n} exp\left(s_{\pi(k)}\right)} \quad (3)$$

where π denotes the descending permutation according to $\mathbf{y}$, and $\pi(k)$ is the patch ranked at the k-th position in π. The ranking model will optimize towards sorting patches in line with their semantic density scores. Compared with the other two kinds of ranking loss, i.e., point-wise and pair-wise losses, the list-wise loss more completely considers the overall relationships among patches. In the implementation, we apply a sparse design to the ranking model, which solely inputs the visible patches instead of the full set. We will verify this design and architecture choice of the ranking model in ablation studies.

Note that each training instance only contains a small portion of visible patches and scores. However, when going through a sufficient number of instances, the ranking model sees all patch locations of an image and learns to judge relative ranks among patches. In other words, the

ranking model knows how to score all the image patches regarding their informativeness. Consequently, we can solely use the ranking model for image redundancy reduction, as shown in the right part of Fig. 2.

3.4. Patch Selection

With the ranking model, a straightforward way for redundancy reduction is to select the top-k patches. Our observation suggests that the top patches favor object boundaries rather than homogeneous regions (e.g., dog's body in Fig. 1), as shown in Fig. 3. We attribute it to that changes in object boundaries more easily cause semantic shifts in the reconstructed image. However, a small proportion of patches from homogeneous regions are also essential for visual tasks [28]. Thus, we introduce DPC-KNN [15] to capture the representative homogeneous patches. Specifically, to select k informative patches including h ($h < k$) ones from homogeneous regions, we first select (k - h) patches as described above. Then, we use DPC-KNN to cluster all patches into k groups and select the largest h ones, given that larger groups are more likely to associate with homogeneous regions. For each selected group, we extract one patch that satisfies the following two conditions, i.e., not selected previously and closest to the cluster center in the remaining patches. The two sets of patches constitute the output of our LTRP. We will ablate h in experiments.

4. Experiments

4.1. Datasets and Experimental Setting

Datasets and evaluation metrics. We pre-train our LTRP on ImageNet-1K training set, and then assess it on two scenarios as follows. The first is image-level evaluation. We conduct image classification on ImageNet-1K, a single-label classification dataset, and on MS-COCO [25], a multi-label classification dataset. The classification model is trained only with the patches remained by different methods. Top-1 accuracy and mAP are employed as the evaluation metrics, respectively. The second is patch-level evaluation. It is conducted both on MS-COCO and PASCAL VOC [17] with their object detection labels, MS-COCO and ADE20K [51] with their semantic segmentation labels. In practice, we treat all the labeled bounding boxes or masks as foreground and evaluate their consistency with the retained patches by using IoU, F1-score, recall, and precision. Following EViT [24], IA-RED^2 [31], etc., we evaluate these methods at different keep ratios (krs). That is, the top-scored kr% patches are preserved. Meanwhile, all the ablation experiments are conducted on ImageNet-100, a subset of ImageNet-1K, given hardware resource constraints.

Model architectures. In LTRP training, we use MAE with ViT-B and choose ViT or CNN (e.g., ResNet50) as our ranking model. In image classification, we utilize ViT-B on ImageNet-1K and MS-COCO, ViT-S on ImgeNet-100 as the backbone for model training. They are both initialized from the MAE encoder [20]. Moreover, we add ML-Decoder [37] behind the classification backbone and use Asymmetric loss [36] for multi-label classification. No additional model is required for the patch-level evaluation, as the performance is obtained by comparing the preserved image patches with groundtruth. All experiments are conducted on a server with 8 NVIDIA GTX 3090 GPUs and Intel(R) Xeon(R) Gold 6226R CPU.

Comparing methods. We compare LTRP with off-the-shelf models from the three classes mentioned previously (see Sec. 2.1). For CAM, we choose Grad-Cam [38] with ViT. For self-supervised methods, we choose MoCo [9] with ResNet50 and DINO [6]. For token reduction, we choose EViT [24], IA-RED^2 [31], DynamicViT [34], AdaViT [30], DGE [39] and ToMe [2]. For fair assessment, these methods are all trained (supervised) or pre-trained (unsupervised) on ImageNet-1K.

4.2. Performance Evaluation

Image-level evaluation. Figure 4a shows the results of different methods at krs 0.7 and 0.5 on single-label classification. The unsupervised and supervised methods are depicted by different colors. LTRP performs better than the self-supervised ones, and slightly worse or on par with state-of-the-art supervised methods. Note that LTRP achieves this without utilizing any supervisory signals.

We then test LTRP on MS-COCO to validate the performance on multi-label classification. As shown in Fig. 4b, LTRP achieves mAP of 82.8% and 81.8% at krs 0.75 and 0.5, respectively. It outperforms the previous best unsupervised methods by 1.1% and 1.4%, and the state-of-the-art supervised methods by 0.7% at both krs. Note that these values correspond to clear accuracy gaps, as the performance differences between existing methods are small. Since different methods only differ in the patches selected, these improvements clearly indicate that LTRP enables a fairer and more meaningful redundancy reduction.

Patch-level evaluation. To further validate whether the patches are unbiasedly retained, we focus on the categories not included in ImageNet-1K, i.e., the unseen categories, for which all the methods have not been explicitly told to learn. We first evaluate the methods based on the object detection labels. The performance of different methods on PASCAL VOC and MS-COCO is shown in the first two rows in Fig. 5. Encouragingly, LTRP surpasses all the competitors in all the evaluated metrics at different krs. The significant improvement in precision highlights that patches selected by LTRP are more accurately located in bounding boxes of unseen categories. With the decrease of krs, the precision of many supervised methods goes up slowly or even goes down. It means that these methods tend to not re-

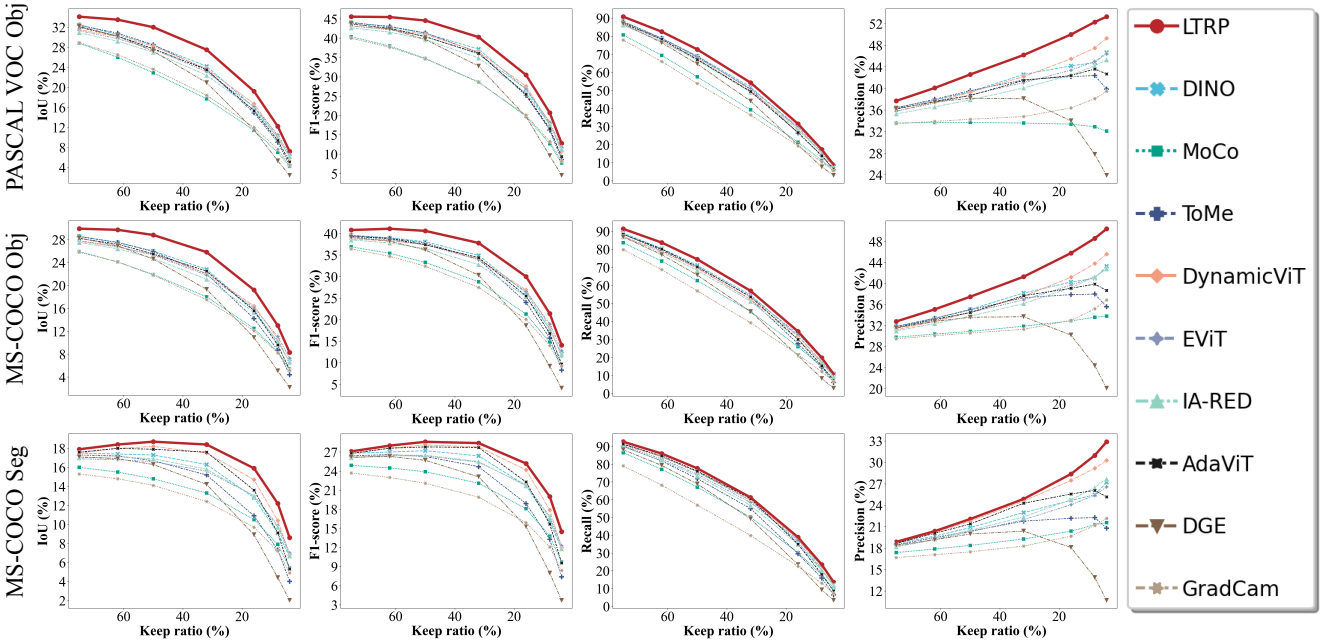


Figure 5. **Experimental results** on object detection and semantic segmentation datasets. For each dataset, we exclude the categories overlapped with ImageNet-1K (learned) and only compute the metrics on the remaining (unseen) categories, which all the methods have not been explicitly told to learn. The results illustrate the merit of LTRP in unbiased image redundancy reduction.

gard patches from unseen categories as the most important ones, while our LTRP still goes up quickly and successfully avoids this bias. The result implies that LTRP can be well generalized to datasets with quite different categories.

We then evaluate the methods based on the semantic segmentation labels. The third row in Fig. 5 shows the results on MS-COCO. LTRP also consistently gives the best results and exhibits a noticeable improvement compared to DynamicViT, the second-best method. We observe that the performance gap becomes larger as the kr decreases. It again demonstrates that LTRP preserves semantics corresponding to the unseen categories well.

4.3. Ablations

In Tab. 1, we provide detailed ablation experiments and introduce some empirical observations as follows.

Reconstruction model. Tab. 1a presents LTRP with different MAE as the image reconstruction model. The optimal model is surprisingly the lightweight model MAE-B. Utilizing MAE-B to generate pseudo label boosts the accuracy by 0.2% compared to larger MAE-L (21 M *v.s.* 81 M in #params). Moreover, MAE-L with GAN [19] loss can get even better reconstruction, but performs worse at kr 0.5. We observe that reconstructions of large models or utilizing GAN loss have finer low-level semantics such as texture details. It tends to pull these patches that excel in recovering low-level semantics ahead in the ranking, consequently reducing the proportion of patches with high-level semantics. This phenomenon causes the accuracy drop.

Masking ratio. Tab. 1b shows the influence of the ratio of removed patches. The accuracy of an excessively high masking ratio (0.95) drops by 0.1% compared with baseline. And a lower masking ratio (0.7) does not perform better than the masking ratio of 0.8. We note that extremely high and low masking ratios both result in the issue of *plateaus* in the pseudo score sequence, namely some patches have similar pseudo scores and occupy a continuous position in the ranking sequence. This causes the ranking model to be confused about the right order of these patches. Considering the efficiency issue, LTRP selects a masking ratio of 0.9.

Ranking model. We investigate the architecture, sparse design and similarity metric of ranking model in Tabs. 1c, 1d and 1h, respectively. We find out that our ranking model is model-agnostic, and even smaller models like ViT-T or ResNet50 can achieve performance similar to the baseline (ViT-S). However, to align with the comparing methods, we select ViT-S. The sparse design excludes the sort-irrelevant patches from the ranking model, it improves accuracy by 0.2% to 0.3% across different krs. Moreover, by skipping these patches, we significantly reduce computational costs.

Clustering. Tab. 1e presents our investigation of clustering. Interestingly, selecting all patches (with clustering ratio 1) through clustering at the pixel level is not extremely bad compared with our baseline (82.7% *v.s.* 83.7%). A small subset (e.g., 20%) of patches from homogeneous regions gets an improved LTRP.

model	kr=0.75	kr=0.5
MAE-B	**84.4**	**83.3**
MAE-L	84.2	83.2
MAE-L-GAN	84.2	83.0

(a) **Reconstruction model.** Lightweight MAE performs better.

ratio	speedup	kr=0.75	kr=0.5
0.95	**6.4x**	84.3	83.2
0.9	3.1x	84.4	83.3
0.8	1.5x	**84.8**	**83.6**
0.7	1.0x	84.6	83.4

(b) **Masking ratio.** A ratio of 0.9 gets a better trade-off between speedup and accuracy.

metrics	kr=0.75	kr=0.5
ℓ_1	**84.4**	**83.3**
PSNR	83.8	82.5
SSIM	83.6	79.9

(c) **Distance Metric.** The ℓ_1 distance is more simple and effective.

model	params	kr=0.75	kr=0.5
Mobile-Former[10]	3.2M	84.3	83.1
ViT-T	5.5M	**84.5**	**83.4**
ViT-S	22M	84.4	83.3
ResNet50	24M	**84.5**	**83.4**

(d) **Ranking Model.** All models exhibit similar accuracy. For a fair comparison, we employ ViT-S.

ratio	kr=0.75	kr=0.5
0	84.4	83.3
0.1	**84.7**	83.4
0.2	84.6	**83.7**
0.3	84.6	83.6
1	83.6	82.7

(e) **Clustering ratio.** A few patches with homogeneous regions can improve classification accuracy.

loss	kr=0.75	kr=0.5
Regression	84.1	83.0
RankNet	82.9	79.5
ListNet	84.1	82.5
ListMLE	**84.4**	**83.3**

(f) **Ranking loss.** List-wise methods perform better than point-wise and pair-wise ranking losses.

phase	speedup	kr=0.75	kr=0.5
encoder	1	**84.4**	**83.7**
decoder	**1.3x**	**84.4**	83.3

(g) **Removal phase.** Removing visible patches before the decoder well balances accuracy and speed.

input	kr=0.75	kr=0.5
visible patches	**84.7**	**83.5**
all patches	84.4	83.3

(h) **Sparse design.** Inputting only visible patches into the ranking model yields better results.

case	kr=0.75	kr=0.5
pixel (w/o norm)	**84.4**	**83.3**
pixel (w norm)	81.5	75.7

(i) **Reconstruction target.** LTRP favors MIM models based on raw pixels.

Table 1. **LTRP ablation experiments** on ImageNet-100. We report accuracy (%) results for krs **0.75** and **0.5**, respectively. The selected configurations are marked in gray.

Ranking loss. Tab. 1f studies the choice of ranking loss. Interestingly, the point-wise approach also has decent performance in our LTRP. Compared to ListMLE, its accuracy drops by only 0.3% at both krs. Point-wise methods train <query, document> pair independently, ignoring the relationship among documents. In contrast, our ranking model considers all documents (patches), i.e., <query, documents>, enabling the acquisition of holistic knowledge. For pair-wise methods, the accuracy drops obviously at krs of 0.75 and 0.5 (82.9 *v.s.* 84.4 and 79.5 *v.s.* 83.3, respectively). Overall, ListMLE performs the best and is also the most appropriate method for utilizing our pseudo-labels.

Removal phase. In Tab. 1g, although the results of removing visible patches before the lightweight decoder show a decrease in performance compared to the approach of removing before the encoder (a reduction of 0.4% at kr 0.5), we can achieve a noteworthy 1.3x speedup in terms of FLOPs (7.07G *vs.* 5.38G). So we employ the former.

Reconstruction target. Pre-patch normalization is a trick that normalizes a patch using the mean and standard deviation of its pixels. In Tab. 1i, we observe that utilizing pre-trained MAE based on pre-patch normalization for generating pseudo-labels performs exceedingly poorly, which is 7.6% worse than using unnormalized pixels. We suspect this drop partially because LTRP requires quantifying the global semantic changes between two images. Conversely, pre-patch normalization appears to reducing global coherence, as it focuses on enhancing the contrast locally [20].

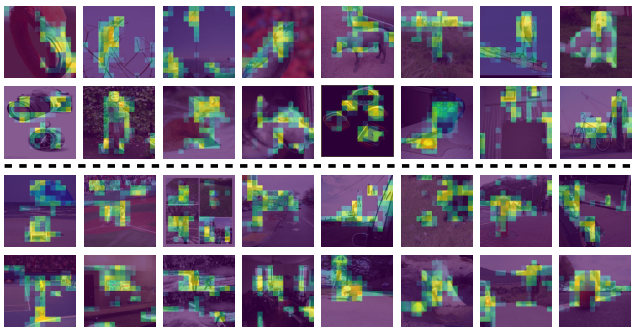


Figure 6. Normalized score maps generated by LTRP on *validation* set of ImageNet-1K (top) and MS-COCO (bottom). Only the top-scored 64 patches are displayed.

4.4. LTRP for Efficient ViT

LTRP allows the classification model to be trained only on a subset of image patches. We verify whether this LTRP-based classification would be more efficient for ViT.

We elucidate this issue from both inference and training perspectives. For inference, time for both patch ranking and classification should be counted. As shown in Tab. 1d, even ViT-T has commendable performance. So we choose small ranking model and large classification model as LTRP-based solutions. Table 2 enumerates several combinations on ImageNet-1K at kr 0.75. Using ViT-T as the ranking model shows 1.19×, 1.36×, or 1.33× acceleration on GPU

cls. model	ranking model	params (M)	FLOPs ↓ (G)	inference speed↑ GPU im/s	CPU im/s	top-1 acc.
	-	86	17.6	1095.1	4.3	83.3†)
ViT-B	ViT-S	108(+25.6%)	13.1(-25.6%)	1034.9(-5.5%)	4.1(-4.7%)	83.2(-0.1%)
	ViT-T	91.5(+6.4%)	13.1(-25.6%)	1302.2(+18.9%)	5.1(+18.6%)	83.1(-0.2%)
	-	307	61.62	357.7	1.3	85.8†)
ViT-L	ViT-S	329(+7.2%)	46.0(-25.3%)	445.6(+24.6%)	1.5(+15.4%)	85.7(-0.1%)
	ViT-T	312.5(+1.8%)	45.9(-25.5%)	485.5(+35.8%)	1.7(+23.1%)	85.6(-0.2%)
	-	632	167.4	140.0	0.5	86.6†)
ViT-H	ViT-S	654(+3.5%)	124.8(-25.4%)	181.0(+29.3%)	0.6(+20.0%)	86.3(-0.3%)
	ViT-T	637.5(+0.9%)	124.7(-25.5%)	185.8(+32.7%)	0.6(+20.0%)	86.1(-0.5%)

Table 2. **Experiments on efficient ViT** on ImageNet-1K at kr 0.75. Different combinations of the ranking and classification (cls.) models show different balances between FLOPs (or inference speed) and accuracy. "-" means do not incorporate the ranking model and select all patches. † means run by us. When using ViT-T as the ranking model, LTRP accelerates the inference ranging from 18.9% to 32.7% with accuracy sacrifice of 0.1%-0.5%.

over the scheme that directly uses ViT-B, ViT-L or ViT-H as the classification model, respectively (ViT-H uses a 16*16 patch grid while ViT-B and ViT-L use 14*14). Meanwhile, the accuracy drops only range from 0.2% to 0.5%. The results imply that LTRP-based ones are more appropriate choices in scenarios requiring both efficiency and accuracy. With the popularity of large visual and multi-modal models, large ViT models are more frequently utilized and we believe that these solutions would receive more attention. While for training, LTRP-based solutions include running the ranking model once for patch selection and training the classification model N epochs. The cost raised by the ranking model would be negligible as N increases. Meanwhile, training the classification model would also be accelerated by nearly $1/kr^2\times$ as fewer patches are involved.

In sum, compared with the vanilla classification, LTRP-based solutions not only accelerate training but also can improve the inference speed up to over 30% on GPU with a possible neglectable accuracy degradation.

4.5. Visualization

Figure 6 displays the informativeness-aware capability of LTRP on ImageNet-1k and MS-COCO. LTRP well identifies patches located in meaningful object regions in ImageNet-1K images. Meanwhile, it also shows superior unbiasedness in selecting multiple objects from MS-COCO images. We also visualize a process of hierarchical patch selection in Fig. 7. With the decrease of retained patches, less important patches are gradually removed. LTRP is consistent in favour of patches located on semantic objects. This removal process is mostly in line with human perception. The results again demonstrate the effectiveness of LTRP.

To explore why LTRP performs better than popular self-supervised methods, we plot the average attention distance of each head of DINO and our LTRP ranking model (both are ViT-S) in Fig. 8. A more scattered distance distribution is observed for LTRP especially in the last few layers. As

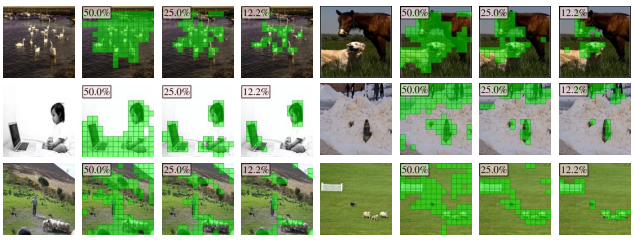
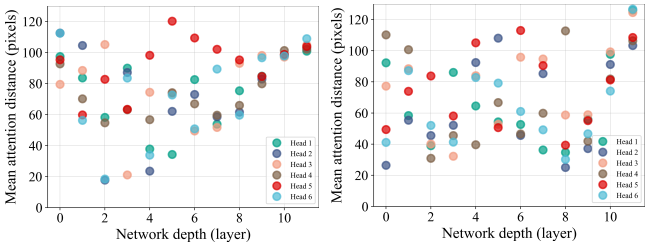


Figure 7. Hierarchical redundancy reduction process of LTRP on MS-COCO *validation* set. Keep ratios are shown in the top left of each image. Patches with green boxes are the retained ones.



(a) DINO (b) LTRP

Figure 8. Average attention distances organized according to layer and head for DINO and our LTRP ranking model.

stated in [43], such a distribution suggests that more diverse representation learning is achieved, which explains in part the superiority of LTRP.

5. Conclusion

We have introduced LTRP, the first self-supervised method for image redundancy reduction. By leveraging the characteristic of self-supervised learning to create patch-level pseudo labels, and then learning to rank patches, LTRP elegantly eliminates the issue of categorical inductive bias commonly observed in leading supervised methods. Extensive experiments on different datasets and tasks basically validate our proposal. LTRP achieves competitive accuracy in single-label classification and obviously better accuracy in multi-label classification. Meanwhile, the retained patches align more closely with object detection and semantic segmentation labels, especially for categories not explicitly learned by supervised methods. In addition, incorporating LTRP into image classification can lead to a notable speedup in inference with neglectable accuracy degradation if applied appropriately. These results convincingly indicate the effectiveness of LTRP and its unbiased patch selection. We hope that LTRP will provide valuable insights into image redundancy reduction and related visual tasks.

Acknowledgement This work was supported by National Science and Technology Major Project (No. 2021ZD0112805) and in part by National Natural Science Foundation of China (No. 62372170, 62172103).

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. 2

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *International Conference on Learning Representations*, 2023. 3, 5

[3] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005. 3

[4] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. Adapting ranking svm to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193, 2006. 3

[5] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 5

[7] Rich Caruana, Shumeet Baluja, and Tom Mitchell. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in neural information processing systems*, 8, 1995. 3

[8] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the learning to rank challenge*, pages 1–24. PMLR, 2011. 3

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 5

[10] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2022. 7

[11] Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han. High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, 29: 5794–5804, 2020. 2

[12] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2351–2360, 2021. 1, 2

[13] 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*, pages 248–255. Ieee, 2009. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 3

[15] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99: 135–145, 2016. 5

[16] Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, and Yu Fu. Visual saliency detection by spatially weighted dissimilarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 473–480. IEEE, 2011. 2

[17] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 2, 5

[18] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 6

[20] 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*, pages 16000–16009, 2022. 1, 2, 3, 4, 5, 7

[21] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE international conference on computer vision*, pages 1665–1672, 2013. 2

[22] Alexandros Karatzoglou, Linas Baltrunas, and Yue Shi. Learning to rank for recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 493–494, 2013. 3

[23] Ping Li, Qiang Wu, and Christopher Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. *Advances in neural information processing systems*, 20, 2007. 3

[24] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 1, 3, 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 5

[26] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009. 3

[27] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 1040–1049, 2017. 3

[28] Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5361–5372, 2023. 5

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[30] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, et al. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022. 5

[31] Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. Ia-red^2: Interpretability-aware redundancy reduction for vision transformers. *Advances in Neural Information Processing Systems*, 34:24898–24911, 2021. 1, 3, 5

[32] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16031–16040, 2022. 2

[33] Majid Rabbani and Paul W Jones. *Digital image compression techniques*. SPIE press, 1991. 2

[34] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 3, 5

[35] Cedric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. Learning to merge tokens in vision transformers. *arXiv preprint arXiv:2202.12015*, 2022. 3

[36] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 5

[37] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–41, 2023. 5

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2, 5

[39] Lin Song, Songyang Zhang, Songtao Liu, Zeming Li, Xuming He, Hongbin Sun, Jian Sun, and Nanning Zheng. Dynamic grained encoder for vision transformers. *Advances in Neural Information Processing Systems*, 34:5770–5783, 2021. 5

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[41] Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, and Gao Huang. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. *Advances in Neural Information Processing Systems*, 33:2432–2444, 2020. 1, 2

[42] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14668–14678, 2022. 4

[43] Yixuan Wei, Han Hu, Zhenda Xie, et al. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 8

[44] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, 2008. 3, 4

[45] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 4

[46] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3

[47] Long Xu, Jia Li, Weisi Lin, Yun Zhang, Yongbing Zhang, and Yihua Yan. Pairwise comparison and rank learning for image quality assessment. *Displays*, 44:21–26, 2016. 3

[48] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2369–2378, 2020. 1, 2

[49] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10809–10818, 2022. 3

[50] 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. 2

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5