

# A Patch-Based Architecture for Multi-Label Classification from Single Positive Annotations

Warren Jouanneau<sup>1,2</sup>, Aurélie Bugeau<sup>2,3</sup>, Marc Palyart<sup>1</sup>, Nicolas Papadakis<sup>4</sup> and Laurent Vézard<sup>1</sup>

<sup>1</sup>*Lectra, F-33610 Cestas, France*

<sup>2</sup>*Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR 5800, F-33400 Talence, France*

<sup>3</sup>*Institut Universitaire de France (IUF), France*

<sup>4</sup>*Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France*

**Keywords:** Partial and Unlabeled Learning, Patch-Based Method, Classification.

**Abstract:** Supervised methods rely on correctly curated and annotated datasets. However, data annotation can be a cumbersome step needing costly hand labeling. In this paper, we tackle multi-label classification problems where only a single positive label is available in images of the dataset. This weakly supervised setting aims at simplifying datasets assembly by collecting only positive image examples for each label without further annotation refinement. Our contributions are twofold. First, we introduce a light patch architecture based on the attention mechanism. Next, leveraging on patch embedding self-similarities, we provide a novel strategy for estimating negative examples and deal with positive and unlabeled learning problems. Experiments demonstrate that our architecture can be trained from scratch, whereas pre-training on similar databases is required for related methods from the literature.

## 1 INTRODUCTION

Data annotation, or labelling, is at the core of supervised learning approaches. In image classification, state-of-the-art methods rely on an ever-increasing amount of data, which makes the annotation collection a major issue. Training datasets are made of image-label associations obtained from an expensive manual annotation, an automatic collection, or filtering and mapping of existing image descriptions. Assembling a dataset is difficult, especially when: experts must annotate unlabelled data; rare events have to be recognized; or a dataset is created from different sources with inconsistent label taxonomies.

The accurate characterization of most images requires multi-label classification. Image content is indeed rich in information, as it includes multiple structured components. Associating a single label to an image is a too restricted setting. It is preferable to annotate an image with several and non-exclusive labels (e.g. presence/absence of wood, metal, fabric, etc.).

For supervised learning purpose, positive and negative examples for each label are needed in the training dataset. The image-labels associations need to be exhaustive. Any missing or incorrect annotation for the label  $l$  on a given image  $X$  leads to a wrong example for this label. Such errors may have an impact on the complete labelling of the image  $X$ , and

on the characterization of the label  $l$  on the remaining images of the dataset. Naturally, obtaining error free multi-label annotations makes the dataset creation more complex. To alleviate the task of dataset creation, weakly supervised learning methods only rely on partial data annotations. Such methods combine approaches ranging from fully supervised to unsupervised learning, e.g. few shot learning where only a small set of examples is available for each label.

A special case of weak supervision for classification is positive and unlabeled (PU) learning (Bekker and Davis, 2020). In PU learning, only partial positive labeling is available. In the multi-label PU setting, the training data contain only a single label for each image, whereas several labels can be present in each image. Hence, only a subset of images of the training data set are annotated for each label. For the remaining images, we have no information, which means that we do not know if a label is present or not in the image. In the multi-label positive and unlabeled learning context, obtaining annotations is greatly simplified. PU learning is indeed adapted to automatic data collection and dataset merging. Positive examples for each label can be collected independently of what they represent for other labels. More generally, as it does not require negative examples and annotation completeness, PU is well suited to handle heterogeneous labelings coming from different datasets.

In PU learning, one difficulty is to deal with the absence of negative examples. The situation is even worse in the multi-label context, where each image is actually both positive for certain labels and negative for other ones.

### 1.1 Problem Setting and Methodology

We face the multi-label PU learning which is a multi-label classification problem where each image may contain zero, one or several objects from one or several different categories. Moreover, in our use cases, only one positive label per image is used during training, and no negative labels are available.

In this work, we propose to address the absence of negative examples in training datasets with a patch-based approach. Following (Trockman and Kolter, 2022), we argue that the patch level can be better suited than the image level to achieve multi-label characterization in the case of PU learning. More precisely, we use the fact that when an image is a positive example regarding a given label, then some of its patches can be considered positives for this label, and other ones negatives. When training examples only contain positive labeling at the image level, the patch-label association is nevertheless unknown.

### 1.2 Contributions and Outline

Our main contribution, presented in section 3 is a light patch-based architecture adapted to multi-label PU learning. By considering an image as a set of patches, we build multi-label image representations with a patch attention mechanism. Assuming that a small image patch mostly contains a single class label, our architecture allows the estimation of negative examples by leveraging on patch-based image representation self-similarities. This is a main novelty, as the negative examples estimated by existing approaches for PU learning (Cole et al., 2021; Verelst et al., 2022) are based on the prediction score at the image level without relying on the local data content.

In section 4, experiments demonstrate that our patch-based framework is adapted to multi-label classification problems with single positive annotations, while providing an explicit spatial localization of labels. When training models from scratch, our architecture generalizes faster and better than Resnet-5, while being significantly lighter (reduction  $\times 100$  of the number of parameters).

## 2 RELATED WORKS

In this section, we first review multi-label classification models using set of patches. Next, we discuss existing strategies to obtain a global image representation from the information contained in a set of patches and their embeddings. Finally, we present state-of-the-art methods for positive and unlabeled learning.

### 2.1 Patch Embeddings for Multi-Label Image Classification

Multi-label classification of images can be split into multiple single label classification tasks (Read et al., 2009), where each classifier is in charge of predicting a specific label. Recent works demonstrate the interest of tackling the joint multi-label problem (Wei et al., 2014). In image data, there is indeed a correlation between localization and labels. As a consequence, detection methods not only intend to infer the image labels, but they also aim at estimating their localization through bounding boxes (Ren et al., 2015; Redmon et al., 2016) or segmentation masks (He et al., 2017).

Considering an image as a set of patches is an appropriate model for multi-label classification. In practical applications, unless a hierarchical label taxonomy is considered (hairs, head, body), labels are indeed related to a subpart of the image only (e.g. wood, brick, metal...). With a patch-based approach, each patch specializes and contributes to the classification with respect to a single label. Patch-based approaches have been introduced for texture synthesis problems (Efros and Leung, 1999). Their interest has been demonstrated for image level tasks such as denoising (Buades et al., 2005), super-resolution (Freeman et al., 2002), segmentation, labeling (Coupé et al., 2011), classification (Varma and Zisserman, 2008), etc. In particular, the representation of an image from its extracted patches has been thoroughly studied, and we refer the reader to (Liu et al., 2019) for a review of methods from the bag-of-word framework to recent deep models.

The Transformer architecture (Vaswani et al., 2017), originally proposed for natural language processing, has soon been transposed to image tasks. In order to adapt Transformer architecture to images, the ViT method (Dosovitskiy et al., 2020) considers an image as a set of patches or as a sequence (if patch position is encoded) of patches. The ViT achieves impressive results without convolution layers. Many extensions of the image Transformer focus on the construction of a relevant set of elements to feed to the network. As an example, CrossViT (Chen et al.,

2021) relies on patches of different size, which brings multi-scale information. In other works, a convolution neural network (CNN) can be used as a stem (Xiao et al., 2021) or as a feature pyramid (Zhang et al., 2020) to feed a Transformer network.

Relying on sets of patch embeddings proved to be a powerful methodology for image classification with modern learning methods. The ConvMixer model (Trockman and Kolter, 2022) shows for instance that Transformers’ performances are mostly due to the representation of images as a set of patches, rather than to the architecture itself. When facing image classification problems, one has to come back from patch level to image level at some point.

## 2.2 From Set of Patch Embeddings to Image Representation

We now present methods providing a relevant image representation from the embeddings of image patches. Image representation must handle multi-label classification and deal with the presence of multiple instances of these labels in the images. To that end, the information contained in the set of embedded patches has to be aggregated. In the literature, the aggregation of elements in a set is mainly done using pooling operator such as average, max, min or sum of the element representations (Zaheer et al., 2017). The pooling of feature maps related to receptive fields of different sizes is for instance usually done with global average (Lin et al., 2013).

The raise in popularity of the attention mechanism has led to new pooling methods (Ilse et al., 2018) realizing weighted sums of element representations. Transformers (Vaswani et al., 2017), relying on multi-head attention, can also operate as element pooling. The transformer uses cross-attention with weights given by a similarity score between the element representations and “queries”. In the case of patches, these “queries” can be seen as codebooks (Zhao et al., 2022). Queries can be designed beforehand, or they can be parameters learned by the model. A main limitation of the transformer is its quadratic complexity with respect to the dimension of input data. In order to reduce the computational burden of transformers, the perceiver (Jaegle et al., 2021) architecture computes intermediate latent representations of reduced dimension before realizing cross-attentions with the queries. Hence, a task dependent element pooling can be considered for each query. The queries can therefore be defined as label embeddings (Lanchantin et al., 2021), with an independent pooling for each label.

In computer vision, multiple instance learning consists in both detecting the presence of a label at

the image level and localizing accurately the corresponding instances (Carbonneau et al., 2018). The attention mechanism provides a joint solution to these problems (Ilse et al., 2018), as element pooling naturally deals with multiple instances of a label in an image. The detection of a label is indeed given by the pooling of the patch embeddings with the corresponding query. The similarity weight between a patch and a query indicates the degree of participation of the patch in the label decision. If a patch has an important weight in the pooling of a label predicted as positive, then this patch should contain relevant information relative to this label. As a consequence, the similarity weights can help to localize distinct subareas of the image corresponding to a single label.

In a supervised setting, patch attention mechanism is adapted to both multi-label and multiple instance cases. However, its application to weakly supervised problems with positive only annotations at the image level requires the development of new methods.

## 2.3 Positive and Unlabeled Learning for Multi-Label Classification

Positive and unlabeled (PU) classification is a weakly supervised classification problem where only positive examples are available. As studied in the review (Bekker and Davis, 2020), many methods have addressed this problem in the single-label setting.

However, only few methods dedicated to PU learning for multi-label classification problems exist. As detailed in (Cole et al., 2021), most works consider the many (and not single) positive (Kanehira and Harada, 2016) case, the single positive or negative one (Huang and Yan, 2018) or the existence of negatives for each label (Ishida et al., 2017).

In this paper, we focus on the complex multi-label application case where only a single positive label is known for each element of the training set. In (Mac Aodha et al., 2019), all but the known positive examples are considered as negatives and included as groundtruth negative examples in the loss function optimized during training. This corresponds to a uniform penalization of the positive predictions, that can be enhanced with a dedicated spatial consistency loss (Verelst et al., 2022). Cole et al. (Cole et al., 2021) propose to enhance this kind of methods with the Regularized Online Label Estimation (ROLE) strategy. The ROLE model first improves the loss function with a term penalizing the distance between the number of positive label predictions for an image and a hyperparameter. This hyperparameter corresponds to the mean number of positive labels per image that is nevertheless unknown in general.

Then, an online label estimation strategy is considered for unobserved labels. It consists in learning new parameters that should correspond to the groundtruth value of unknown labels, that can be either positive or negative. These estimations are realized in a separate branch of the model and are compared with the actual predictions of the model in a dedicated loss.

The ROLE model (Cole et al., 2021) makes an interesting proposition with the online estimation of negative examples. Nevertheless, it does not leverage on data content to provide negative examples and tackle the multi-label aspect of the problem. The online strategy mainly stabilizes the learning through memorization of former label predictions. In practice, it reinforces the tendencies (positive or negative predictions) provided by the current model, by making the predicted weight values closer and closer to 0 or 1. We argue that the patch-based approach is a suitable strategy to estimate negative examples in the context of multi-label PU learning.

### 3 METHODOLOGY

We start this section with a formal statement of the multi-label classification problem addressed in this paper. Given a set of labels  $l \in L$ , the objective is to determine  $\mathbb{P}(x_n|l)$ , the probability of presence of the label  $l$  in an image  $x_n$ . We denote as  $\mathbf{y}_n = \{y_{n,l}\}_{l \in L}$  the ground truth labels that indicate if a class  $l \in L$  is present ( $y_{n,l} = 1$ ) or not ( $y_{n,l} = 0$ ) in an image  $x_n$ . Hence, the multi-label problem can be formulated as the estimation of a labeling score  $\hat{\mathbf{y}}_n = \{\hat{y}_{n,l}\}_{l \in L}$ , where  $\hat{y}_{n,l} \in [0, 1]$  indicates the presence ( $\hat{y}_{n,l} \rightarrow 1$ ) or absence ( $\hat{y}_{n,l} \rightarrow 0$ ) of the label  $l$  in the image  $x_n$ .

Training multi-label classification in a fully-supervised manner requires entirely annotated data, for which the collection and annotation are costly. In our case, this complete annotation is not available. We have partial ground truth annotations on an image dataset containing partial positive  $\mathbf{z}_n^+$  and negative  $\mathbf{z}_n^-$  examples for the image  $x_n$ . We consider that  $z_{n,l}^+ = 1$  (resp.  $z_{n,l}^- = 1$ ) means that the label  $l$  is present (resp. absent) in the annotated image  $x_n$ . If  $z_{n,l}^+ = z_{n,l}^- = 0$  then we have no information on the label  $l$ . Finally, the two annotated sets are assumed compatible, so that the case  $z_{n,l}^+ = z_{n,l}^- = 1$  is impossible.

In the Positive and Unlabeled (PU) context considered in this paper, nothing is known about the negatives ( $z_{n,l}^- = 0$  for all  $n$  and  $l$ ) and the positive labels are only partially observed. In our applications, we nevertheless consider that one single positive label  $z_{n,l}^+ = 1$  is available per image  $x_n$ .

To solve the multi-label PU problem, we propose

a weakly supervised method that relies on the use of image patches. It is built on top of a key hypothesis: a small enough image patch is mostly characteristic of only one single label.

In section 3.1, we describe our patch-based architecture for multi-label classification. Section 3.2 presents the loss proposed to train the model. In section 3.3, we build on our patch-based architecture to provide negative examples and deal with PU learning.

#### Why a New Patch Architecture for PU Learning?

In section 3.1, we propose a new patch architecture adapted to multi-label learning problems, where only (partial) positive labels are available in the training set. Classical models such as ViT (Dosovitskiy et al., 2020) rely on one global image representation for the classification task. The embedding obtained for a given image and a given label thus contains information coming from all areas of the image, including areas that are relevant for other labels. Hence it is difficult to characterize negative examples for a single label (i.e. images that are negative detections for one label) using these global image embeddings.

In this work, as in ConvMixer (Trockman and Kolter, 2022), we rather represent an image with a subset of its patches. Contrary to ConvMixer that considers a global image representation, we build different image representations, one per class label. Moreover, all embeddings lives in the same latent space, that are composed of embedded features extracted from different subsets of patches. Thanks to this common latent space, subsets of patches can be easily obtained by only selecting image patches that are relevant for each label. As we later demonstrate in section 3.2, such a model leads to a new and simple scheme for estimating negative examples at the patch level. As a by-product, it also allows for the localization of detected labels within images

### 3.1 Patch-Based Architecture

We now provide a detailed description of our patch-based architecture. As illustrated in Figure 1, the architecture is composed of five blocks based on the bag-of-word framework. A subset of patch is first selected. The patches are then embedded in a general representation space, in which also live label codebooks. A pooling of the patch embedding for each label is then performed using an attention mechanism. It is followed by the final multi-label classification.

#### 3.1.1 Patch Extraction

(Block 1 of figure 1): An image is first converted into a set of patches extracted at different resolutions.

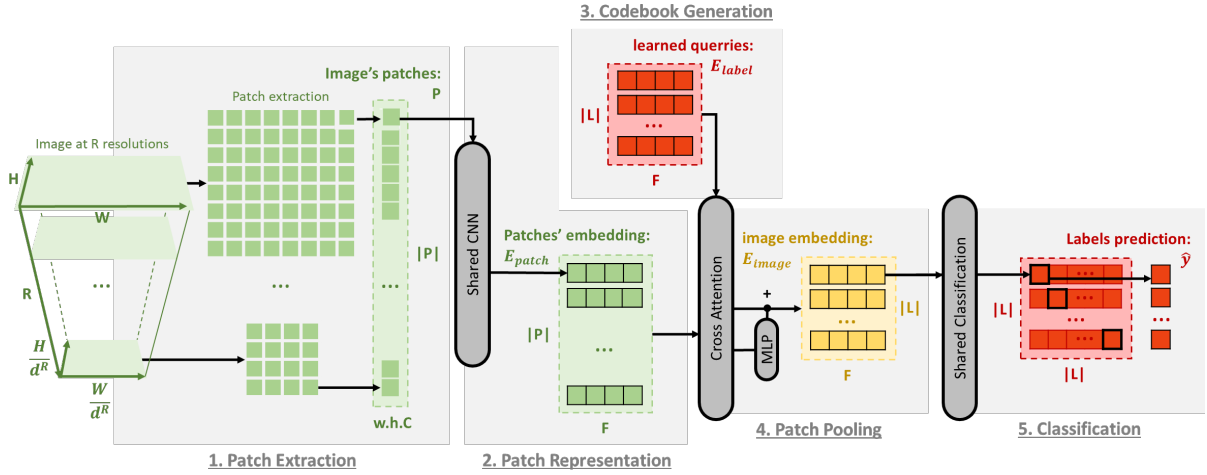


Figure 1: Proposed multi-resolution patch deep neural network architecture.

Each image  $I \in \mathbb{R}^{H \times W \times C}$  of height  $H$ , width  $W$  and with  $C$  channels is downsampled spatially  $R$  times with a constant uniform ratio  $d$ , using bilinear interpolation. For each resulting image, patches  $P$  are extracted using a sliding window looping through the image both horizontally ( $W$ ) and vertically ( $H$ ) with a uniform window size  $h \times w$  and stride. At each resolution, we denote a patch  $p \in \mathbb{R}^{h \times w \times C}$ . The patch size being the same across all resolutions, the set of patches contain fine to coarse information with decreasing resolution. In this work, we consider the same size for the stride and the sliding windows, so that patches form a perfect grid of the image. When the image size is too large, patches can be randomly subsampled at each resolution level to limit the computational burden.

### 3.1.2 Patch Representation

(Block 2 of figure 1): Each patch is fed into a CNN architecture. The weights of this model are shared across all patches. With this backbone model, all  $m$  patches are projected into the same latent space of dimension  $F$ ,  $E_{patch} = \{e_{patch_i} \mid e_{patch_i} \in \mathbb{R}^F, i = 1 \dots m\}$ , resulting in a vector of embedded features  $e_{patch_i} \in E_{patch}$  that play the role of patch descriptors. The CNN architecture consists of multiple efficient-Net blocks (Tan and Le, 2019).

### 3.1.3 Codebook Embedding

(Block 3 of figure 1): Each label is associated with its own representative patch,  $e_{label_l}$ , that we call codebook. We assume that each representation  $e_{label_l}$  contains the embedded feature that a patch should contain to be discriminated as positive in regard to the corresponding label  $l$ . The set of all codebooks, denoted,  $E_{label} = \{e_{label_l} \mid e_{label_l} \in \mathbb{R}^F, l \in L\}$ , is obtained

through back-propagation.

### 3.1.4 Image Representations with Patch Pooling

(Block 4 of figure 1): An image representation is created from the embedded patches. To that end, we propose to consider attention pooling. The pooling is done using cross-attention between the set  $E_{patch}$  of embedded patches and the learned label codebooks  $E_{label}$ . This attention mechanism can be seen as a two steps approach. The first step consists in evaluating the relevance of all selected patches with respect to each possible label. To do so, we consider the scalar product between vectors to define a score matrix  $A$  of weights  $\alpha_{l,i}$  between the representation  $e_{patch_i}$  of patch  $i$  and the representation  $e_{label_l}$  of label  $l$ :

$$\alpha_{l,i} = \frac{\exp(e_{label_l} \cdot e_{patch_i})}{\sum_{j=1 \dots n} \exp(e_{label_l} \cdot e_{patch_j})}. \quad (1)$$

The second step then realizes a weighted sum of the patch representations through the matrix product  $AE_{patch}$ . Inspired from (Vaswani et al., 2017), we finally define the image representation as:

$$E_{image} = f(AE_{patch}) + AE_{patch}, \quad (2)$$

where  $f$  is a feed forward MLP. With this attention framework, we get multiple image representations  $E_{image} = \{e_{image_l} \mid e_{image_l} \in \mathbb{R}^F, l \in L\}$ . Hence, one image representation  $e_{image_l}$  embeds features of a subset of patches matching the global patch representation of a label  $l \in L$ .

### 3.1.5 Classifier

(Block 5 of figure 1): We propose to realize the single multi-label classification with a shared classifier.

Given an image representation  $e_{\text{image}_l}$ , a classifier provides a prediction  $\hat{y}_l$  relative to the presence of labels  $l \in L$  in the image. In practice, the predictions are obtained with a softmax operator

$$\hat{y}_l = \frac{\exp(W_l e_{\text{image}_l})}{\sum_{k \in L} \exp(W_k e_{\text{image}_l})}, \quad (3)$$

where  $W_l$  are weight matrices that are learned for each label. The weights being shared with the softmax operator, the classifier operates a partition of the latent space. Our objective with this classifier model is to get a label specialization of the embedding space. In other words, this classification architecture is designed to enforce each learned label codebook  $e_{\text{label}_l}$  to be the centroid of the patch embeddings relative to the label  $l$ .

### 3.1.6 Discussion on the Proposed Architecture

Overall we suggest that our architecture enforces intra cluster consistency with the patch pooling and inter cluster dissimilarity with the classifier. Indeed, while the scalar product between patches representation relative to the same label is maximized by the attention pooling (see relation (1)), the classifier block implicitly aims at minimizing the scalar product between label image representations  $e_{\text{image}_l}$ .

There are several differences and novelties in our patch embedding strategy with respect to the ViT one (Dosovitskiy et al., 2020) and its extensions (Chen et al., 2021; Xiao et al., 2021; Zhang et al., 2020). As in (Chen et al., 2021), the initial patch extraction is performed from multiple resolution resulting in coarse to fine patches. Next, all patches are processed independently of their positions. Compared to ViT, we do not need positional encoding. Finally, we use a CNN to extract high level features used as patch embeddings. A patch is therefore considered as an image and not as a token as in ViT.

## 3.2 Training Loss

We now present the multi-label loss used in our framework to achieve a prediction  $\hat{\mathbf{y}}$  of the ground truth  $\mathbf{y}$  from positive and negative examples  $\mathbf{z}^+$  and  $\mathbf{z}^-$ . First, to introduce the different losses and avoid possible confusions, we review the differences between multi-class and multi-label problems. We recall that we omit the image index  $n$  to simplify the notations:  $\hat{\mathbf{y}} = \{\hat{y}_l\}_{l \in L}$  is the prediction of the probability of presence of labels  $l$  for any single image.

### 3.2.1 Supervised Multi-Class Learning

In multi-class problems, classes are mutually exclusive: for each image, there exists a single label  $l$  such that  $y_l = 1$  in the ground truth, while  $y_k = 0$  for all  $k \neq l$ . As a consequence, the predictions  $\hat{y}_l$  are constrained to belong to the simplex (i.e.  $\hat{y}_l \geq 0$  and  $\sum_{l \in L} \hat{y}_l = 1$ ). The standard loss function is then the Cross Entropy  $\mathcal{L}_{\text{CE}}$  between the available positive examples  $\mathbf{z}^+$  and the normalized predictions  $\hat{\mathbf{y}}$ :

$$\mathcal{L}_{\text{CE}}(\mathbf{z}^+, \hat{\mathbf{y}}) = - \sum_{l \in L} z_l^+ \log(\hat{y}_l). \quad (4)$$

### 3.2.2 Supervised Multi-Label Learning

On the other hand, in multi-label classification, multiple positive groundtruth are possible for a single image ( $\sum_{l \in L} y_l \geq 1$ ). Hence, the simplex constraint can not be considered anymore, and each label prediction is an independent score  $\hat{y}_l \in [0, 1]$ . Negative predictions must also be taken into account and compared with negative examples contained in  $\mathbf{z}^-$ . In this work, we use the Binary Cross Entropy loss  $\mathcal{L}_{\text{BCE}}$ , that is a standard loss for multi-label classification. For each class  $l \in L$ , BCE is the sum of two cross entropy (4) terms between positive  $z_l^+ = 1$  (resp. negative  $z_l^- = 1$ ) ground truths observations and positive  $\hat{y}_l$  (resp. negative  $1 - \hat{y}_l$ ) predictions:

$$\mathcal{L}_{\text{BCE}}(\hat{\mathbf{y}}) = \mathcal{L}_{\text{CE}}(\mathbf{z}^+, \hat{\mathbf{y}}) + \mathcal{L}_{\text{CE}}(\mathbf{z}^-, 1 - \hat{\mathbf{y}}). \quad (5)$$

### 3.2.3 Positive and Unlabeled Learning (PU)

PU learning involves two main difficulties: (1) the groundtruth is partially labeled and (2) we only have access to positive examples ( $z_l^- = 0$  for all  $l \in L$ ).

To handle the partial labeling of positive examples, we make the assumption that, using the loss  $\mathcal{L}_{\text{BCE}}$ , the features learned on one image for a given label will transpose globally to others. To deal with unknown negative labels, two possibilities can be distinguished. The first one consists in using the available positive labels only and train the model with the loss  $\mathcal{L}_{\text{CE}}(\mathbf{z}^+, \hat{\mathbf{y}})$ . However, this loss function is globally minimized with the trivial solution predicting all labels as positive for all images. The second option is to consider all labels except the observed one as negative examples (Mac Aodha et al., 2019), i.e. training with the loss  $\mathcal{L}_{\text{CE}}(\mathbf{z}^+, \hat{\mathbf{y}}) + \lambda \mathcal{L}_{\text{CE}}(1 - \mathbf{z}^+, 1 - \hat{\mathbf{y}})$ . The penalization parameter  $\lambda \geq 0$  is difficult to tune in general. This model realizes a blind homogeneous penalization of negative examples, independently of the image content, which encourages predicting only one positive label ( $\sum_{l \in L} \hat{y}_l \approx 1$ ).

We propose a trade-off between both approaches, with the weak negative loss

$$\mathcal{L}_{\text{WN}} = \mathcal{L}_{\text{CE}}(\mathbf{z}^+, \hat{\mathbf{y}}) + \mathcal{L}_{\text{CE}}(\tilde{\mathbf{z}}^-, 1 - \hat{\mathbf{y}}), \quad (6)$$

where  $\tilde{\mathbf{z}}^-$  is a weak estimation of the unknown ground truth negative examples. We highlight a main difference between our model and the one introduced in (Cole et al., 2021). In (Cole et al., 2021), all unobserved ground truth labels are learned, i.e. the unknown value of  $y_l$  is estimated online for all  $z_l^+ = 0$ . On the other hand, we only aim at estimating partial weak negative examples  $\tilde{z}_l^- = 1$ , corresponding to a subset of ground truth labels  $y_l = 0$ . In the next section, we detail how leveraging our image representation to obtain these negative examples.

### 3.3 Negative Example Estimation with Embedding Self-Similarities

In order to tackle the positive and unlabeled problem, we propose to estimate negative examples  $\tilde{\mathbf{z}} = \{\tilde{z}_l^-\}_{l \in L}$  for each image. This estimation is done using the information contained in image representations  $E_{\text{image}}$ , together with our initial postulate hypothesis: at most one label can be observed in a patch.

#### 3.3.1 Image Representations Self-Similarities

We first show that two image representations,  $e_{\text{image}_l}$  and  $e_{\text{image}_k}$  for labels  $l$  and  $k$  respectively, are similar if they come from similar image patches. To measure patch closeness, we consider the cosine similarity metric  $\text{sim}(u, v) = \frac{u \cdot v}{\|u\|_2 \times \|v\|_2}$  between two vectors  $u$  and  $v$ . As defined in (2), the image representation  $e_{\text{image}_l} \in E_{\text{image}}$  for label  $l$  is designed to select patch embeddings  $e_{\text{patch}_l} \in E_{\text{patch}}$  that positively correlate with label embeddings  $e_{\text{label}_l}$ . We also highlight that with our multi-class classifier, label embeddings  $E_{\text{label}} = \{e_{\text{label}_l}, l \in L\}$  are assumed to cluster the embedding space. As a consequence, if the cosine similarity  $\text{sim}(e_{\text{image}_l}, e_{\text{image}_k})$  between the image representations for two labels is large, it is most likely that these representations are based on similar subsets of patch embeddings  $e_{\text{patch}_l}$ . Therefore, they come from similar image patches.

Next, as we assume that one label is mainly observed in a patch, if a set of patches is really representative of a label  $l^*$ , this set can not be relevant for characterizing other labels  $k \neq l^*$ . Thus, the model should not return a positive classification score for a label  $k$  different from  $l^*$ .

Combining these observations, we conclude that when the cosine similarity  $\text{sim}(e_{\text{image}_l}, e_{\text{image}_k})$  is large, the image representations  $e_{\text{image}_l}$  and  $e_{\text{image}_k}$  are

based on the same set of patches and at least one of the prediction for the labels  $l$  and  $k$  should be negative.

#### 3.3.2 Estimating Negative Labels

We propose to exploit self cosine similarities between image representations to estimate negative examples  $\tilde{\mathbf{z}}^-$ . We recall that at least one positive example, say  $z_{l^*}^+ = 1$ , is observed for any image. Hence, we rely on the value of the cosine similarity with observed labels,  $\text{sim}(e_{\text{image}_{l^*}}, e_{\text{image}_k}) \in [-1, 1]$ , to determine if unobserved labels  $k$  can be considered as negative examples. To that end, we first define the weights

$$\beta_{l,k} = \varphi(\text{sim}(e_{\text{image}_l}, e_{\text{image}_k}), \theta), \quad (7)$$

where  $\varphi(x, \theta) = \mathbb{1}_{[x > \theta]} x$  is a thresholded Relu operator of parameter  $\theta$ . This parameter  $\theta \in [-1, 1]$  is the value at which the cosine similarity is small enough to consider the two embeddings  $e_{\text{image}_l}$  and  $e_{\text{image}_k}$  as different. Choosing  $\theta \geq 0$  guarantees that  $\varphi(\text{sim}(e_{\text{image}_l}, e_{\text{image}_k}), \theta) \in [0, 1]$ .

Then, as illustrated in Fig. 2, we estimate for each image a negative example score  $\tilde{z}_l^-$  for all unobserved labels  $l \in L$  (i.e. when  $z_l^+ = 0$ ):

$$\tilde{z}_l^- = \max_{\substack{k \in L \\ z_k^+ = 1}} \beta_{l,k}. \quad (8)$$

This states that the label  $l$  can be considered as a (weak) negative example if its embedding is similar enough to the embedding of one of the observed labels  $z_k^+ = 1$ . Negative scores  $\tilde{z}_l^-$  take continuous values in the range  $[0, 1]$  thus giving weak negative labels for  $0 < \tilde{z}_l^- < 1$ . The score is 0 if the cosine similarity value is smaller than the threshold  $\theta$  (see  $\tilde{z}_{l'}^-$  in Fig. 2).

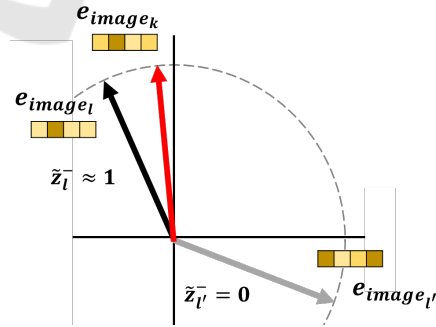


Figure 2: Estimation of weak negative example scores  $\tilde{z}_l^-$  and  $\tilde{z}_{l'}^-$  for unobserved labels and  $l$  and  $l'$  from the single observed label  $k$ .

Reintroducing the image index  $n$ , and recalling the definition (4) of the Cross Entropy, our weak negative

loss function (6) can be rewritten as

$$\begin{aligned} \mathcal{L}_{\text{WN}} &= \sum_n \mathcal{L}_{\text{CE}}(\mathbf{z}_n^+, \hat{\mathbf{y}}_n) + \mathcal{L}_{\text{CE}}(\tilde{\mathbf{z}}_n^-, 1 - \hat{\mathbf{y}}_n) \\ &= - \sum_n \sum_{l \in L} z_{n,l}^+ \log(\hat{y}_{n,l}) - \sum_n \sum_{l \in L} \tilde{z}_{n,l}^- \log(1 - \hat{y}_{n,l}). \end{aligned} \quad (9)$$

This model thus provides negative scores  $z_{n,k}^-$  which values depend on the embedding similarity between unobserved labels  $z_{n,k}^+ = 0$  and observed ones  $z_{n,l}^+ = 1$ .

## 4 EXPERIMENTS

In this section, we first describe the experimental setting in section 4.1. In section 4.2, we validate numerically our proposed architecture and our framework for negative estimation from multiple image representations self-similarities. Comparisons and discussions are finally provided in section 4.3.

### 4.1 Settings

#### 4.1.1 Datasets

Only few datasets propose multi-label annotated images. Datasets traditionally used for detection or segmentation can nevertheless be adapted to fit the multi-label learning problem. With such datasets, the ground truth bounding boxes or masks are discarded, and only the associated labels are conserved. All the instances of a label are considered as one label in the resulting annotation. Hence, in the ground truth annotation,  $y_{n,l} = 1$  means that at least one instance of label  $l$  is observed in the image  $x_n$ . In our work, the single positive label is obtained with a uniform random selection between all ground truth labels.

As in (Cole et al., 2021; Verelst et al., 2022), we consider two datasets that are adapted to the multi-label Positive and Unlabeled problem, since different labels are present in an image: COCO (Lin et al., 2014) and Pascal VOC (Everingham et al., 2010). The 2012 version of VOC contains 5,717 images for model learning and 5,823 images for validation. Objects can be of 20 different classes, giving 20 different labels. Each color image can be at a resolution of up to  $640 \times 640$ . The 2014 version of COCO contains 82,081 images for model learning and 40,137 images for validation. The dataset is annotated with 80 different labels. The dataset images are also in colors and have a resolution of up to  $500 \times 500$ .

#### 4.1.2 Model Architecture

The patch extraction was performed on  $R = 3$  down-sized levels of resolution, with a factor of  $d = 2$  between each level. The patches are squared sub images of size  $w = h = 64$ , extracted with a stride of 64. For COCO and VOC, it results in a set of around 130 patches per images. All embeddings  $e_{\text{patch}_i}$ ,  $e_{\text{label}_l}$  and  $e_{\text{image}_j}$  are of size  $F = 256$ . The  $64 \times 64$  patches are processed with a CNN backbone. This patch embedder contains the first five EfficientNet blocks with the same hyperparameters than (Tan and Le, 2019). The last two layers are composed of an average pooling layer and a fully connected layer of size  $F = 256$  in order to obtain the patch representations  $e_{\text{patch}_i}$ .

Label embeddings  $E_{\text{label}}$  of size  $|L| \times F$  are learned by the model, the number of labels being  $|L| = 80$  for COCO and  $|L| = 20$  for VOC. Attention pooling is performed with a regular cross attention and processed by a MLP of two layers with 256 neurons, resulting in  $L$  image representations  $e_{\text{image}_j}$  of size  $F = 256$ . All weights are initialized with the unit variance scaling method. The GELU activation is used for MLPs and the swish activation for all convolutions.

The experiments were conducted on one Tesla P40 GPU through an Azure virtual machine. Our model has been trained for 25 epochs with batches of 16 images. The starting learning rate is set to  $lr = 0.001$  and scheduled to decrease every 5 epochs, taking the different values  $lr = [0.001, 0.0005, 0.00025, 0.000125]$ . We use the optimizer LAMB (You et al., 2019) with a weight decay of 0.0001. For the proposed negative estimation framework, we used  $\theta = 0$  to compute the weight  $\beta$  in (7), resulting in a standard ReLU for the similarity normalization.

#### 4.1.3 Compared Methods

We consider as reference the model trained using full supervision (i.e. all positives and negatives labels are known) with  $\mathcal{L}_{\text{BCE}}$  (5). We recall that all other models are trained using only one positive label per image.

We also compare our method with two closely related methods (Cole et al., 2021) and (Verelst et al., 2022) addressing the problem of multi-label learning with single positive examples. All results are obtained using the mean average precision (mAP) as the evaluation metric on the same validation sets of images and from models trained on the same dataset versions. We highlight that (Cole et al., 2021) presents results obtained with a pre-trained (and fine-tuned) Resnet-50. Our architecture being trained from scratch, we consider the setting of (Verelst et al., 2022), obtained with conditions similar to ours. Hence, for comparisons with (Cole et al., 2021), we report the results



from (Verelst et al., 2022), where Resnet-50 architectures are trained from scratch with different losses during 100 epochs.

We provide comparisons with the models obtained after 100 epochs with the losses  $\mathcal{L}_{AN}$ ,  $\mathcal{L}_{EPR}$ ,  $\mathcal{L}_{ROLE}$  proposed in (Cole et al., 2021) and  $\mathcal{L}_{EN+CL}$ ,  $\mathcal{L}_{EN+SCL}$  from (Verelst et al., 2022). In brief,  $\mathcal{L}_{AN}$  is the "assume negative" loss considering all unobserved labels as negatives;  $\mathcal{L}_{EPR}$  is the expected positive regularization applying a penalization to make the sum of positively predicted labels close to the average number of positive labels per image; and  $\mathcal{L}_{ROLE}$  is the regularized online label estimation which considers the same penalization and also estimates ground truth unobserved labels as parameters of the model for all images. Next,  $\mathcal{L}_{EN+CL}$  is the expected negative with consistency loss that uses augmented versions of an image; and  $\mathcal{L}_{EN+SCL}$  is the expected negative with spatial consistency loss that predicts spatial classification scores on an augmented image feature map.

When training our architecture with  $\mathcal{L}_{EPR}$ , the average label per image is set to  $k = 2.92$  for COCO and  $k = 1.38$  for VOC. As recommended in (Cole et al., 2021), for  $\mathcal{L}_{ROLE}$ ,  $lr$  is multiplied by 10 in the branch estimating the value of all unobserved labels.

## 4.2 Results and Comparisons

We now present experiments to validate both the architecture and the framework for negative example estimation. In Table 1, we provide mAP obtained with our architecture (top part) and the Resnet-50 one (bottom part). For fair comparison of architectures, both models have been trained from scratch with the loss  $\mathcal{L}_{BCE}$ . This provides upper bounds for the architectures, as this experimental setting corresponds to the "ideal" case where all positive and negative are known. Globally, our patch-based architecture seems adapted to multiple label learning. In full supervision, our architecture achieved 65.8 on COCO, which is better than the 64.8 reported in (Verelst et al., 2022) for the Resnet-50. On VOC, the difference in favor of our patch model is significant (61.6 vs 53.4).

To validate our negative example estimation framework, we present results obtained in the single positive case. We recall that our strategy consists in using image representations self-similarities to estimate weak negative examples that are plugged into the loss  $\mathcal{L}_{WN}$ . As illustrated in Table 1, for the same number of epochs, mAP results are close to the ones obtained with full supervision (63.2 vs 65.8 for COCO and 60.4 vs 61.6 for VOC).

We also trained our patch-based architecture with the competing  $\mathcal{L}_{AN}$ ,  $\mathcal{L}_{ROLE}$  and  $\mathcal{L}_{EPR}$  losses. Our ar-

chitecture trained with our negative estimation proposition gives the best results both on COCO and VOC. With  $\mathcal{L}_{AN}$ , the sum of predicted labels is always close to 1, which better fits the VOC dataset (which true mean number of labels per image is  $k = 1.38$ ), than the COCO one ( $k = 2.92$ ). It should be noticed that  $\mathcal{L}_{ROLE}$  underperforms with the considered learning setting. We suggest this is due to the complexity of the model, which intends to estimate all unobserved labels with a dedicated branch in the loss function.

For completeness, we reproduce in Table 1 (bottom part) the mAP results reported in (Verelst et al., 2022), when training a Resnet-50 architecture from scratch with the competing losses. Contrary to our architecture, the decrease of mAP performance with single positive examples is significant with respect to the full supervision.

All these results indicate that the proposed framework is adapted to the complex multi-label PU problem including only single positive examples.

## 4.3 Discussion

### 4.3.1 Computational Burden

The computational burden for training our patch-based architecture is significantly reduced with respect to the Resnet-50 model of (Cole et al., 2021) and (Verelst et al., 2022). First, the model size is reduced by a factor 100. Our model has approximately 250K parameters to learn, whereas the Resnet-50 architecture is composed of 23M parameters. With our light patch-based architecture, better results are also obtained with only 25 epochs, instead of 100 epochs for the Resnet-50 architectures. This suggests that our architecture generalizes better.

### 4.3.2 Hyper-Parameters

Our weak negative loss does not include any hyper-parameter. Excepting the network architecture, the only extra hyper-parameter of our full model is the threshold of the cosine similarity in (7) that we simply fix to  $\theta = 0$ . On the other hand, the losses  $\mathcal{L}_{EPR}$  and  $\mathcal{L}_{ROLE}$  penalize the sum of positive predictions with respect to the average number of label per image. This is a strong assumption, as the variance of positive labels on all image of the dataset can be large. Moreover, the prior knowledge of the mean number of labels is often not available in real use cases.

The loss  $\mathcal{L}_{ROLE}$  also relies on an online estimation of all unobserved ground truth annotations from current predictions. This model is thus greatly influenced by the initialization and the first few epochs, while increasing the memory requirements.

Table 1: mAP results for our patch approach trained for 25 epochs and different losses (top) and comparisons (bottom) with the results reported by (Verelst et al., 2022) after training a Resnet-50 architecture for 100 epochs. Best results in bold.

Model	Loss/method	COCO-14	VOC-12
Patch based architecture (ours)	$\mathcal{L}_{BCE}$ (fully-annotated)	65.8	61.6
	$\mathcal{L}_{AN}$ (Cole et al., 2021)	62.6	60.0
	$\mathcal{L}_{EPR}$ (Cole et al., 2021)	61.4	58.8
	$\mathcal{L}_{ROLE}$ (Cole et al., 2021)	33.3	49.7
	$\mathcal{L}_{WN}$ (ours)	<b>63.2</b>	<b>60.4</b>
Resnet-50 (reported in (Verelst et al., 2022))	$\mathcal{L}_{BCE}$ (fully-annotated)	64.8	53.4
	$\mathcal{L}_{AN}$ (Cole et al., 2021)	50.2	45.7
	$\mathcal{L}_{ROLE}$ (Cole et al., 2021)	51.9	45.0
	$\mathcal{L}_{EN+CL}$ (Verelst et al., 2022)	54.3	47.0
	$\mathcal{L}_{EN+SCL}$ (Verelst et al., 2022)	54.0	50.4

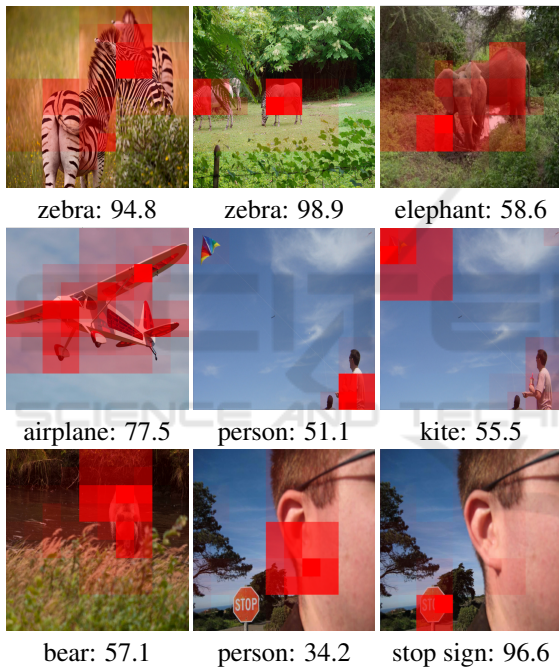


Figure 3: Examples of patch attention scores for true positives. Patches are filled with their attention score values, that vary from 0 (transparent) to 1 (red). The second line present the prediction score for the given labels..

#### 4.3.3 Label Localization

Our architecture has the potential to locate patch examples corresponding to a detected label (Fig. 3). Indeed, the attention scores  $\alpha_{l,i}$  computed in the attention patch pooling with relation (1) (see block 4 of Fig. 1) allows determining the level of patch  $i$  participation to the classification decision relative to label  $l$ . The patch-based model thus offers a natural framework to interpret the obtained results, without relying on advanced gradient backpropagation mech-

anisms (Selvaraju et al., 2017).

#### 4.3.4 Pre-Training and Fine-Tuning

It is important to underline the current limitation of our approach with respect to Resnet-50 architectures. The models (Cole et al., 2021) and (Verelst et al., 2022) provide significant better mAP results (72 for COCO and even 88 for VOC), when considering a Resnet-50 pre-trained on Imagenet, with potential fine-tuning refinements. We postulate that our results could also be improved by pretraining either our full patch-based architecture on Imagenet, or the patch embedder on bounding boxes of a detection dataset. The performance could also be increased by conducting an extensive hyper-parameter search (dimension of the representation space  $F$ , number of attention layers, ...) and training tuning.

## 5 CONCLUSION

In this work, we proposed a light patch-based architecture for multi-label classification problems. Leveraging on patch embedding self-similarities, we provide a strategy for estimating negative examples when facing the challenging problem of positive and unlabeled learning. The patch-based attention strategy also gives a natural framework to localize detected labels within images.

Numerical experiments demonstrate the interest of the approach when no dedicated pre-trained network is available. Our model is able to generalize fast from few labels, as it provides relevant results when trained from scratch during a few epochs. In the related literature, the best performances are obtained with pre-trained Resnet-50 architectures having a number of parameters 100 times greater than ours.

In order to improve our model and reach state-of-the-art results, the main directions we draw are the online estimation of positive and negative example at batch level, the pre-training of the patch embedder and an improved model to cluster the patch embedding space with respect to the labels.

## REFERENCES

- Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760.
- Buades, A., Coll, B., and Morel, J.-M. (2005). A non-local algorithm for image denoising. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 60–65 vol. 2.
- Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.
- Chen, C.-F. R., Fan, Q., and Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *IEEE Int. Conf. on Computer Vision*, pages 357–366.
- Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., and Jojic, N. (2021). Multi-label learning from single positive labels. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 933–942.
- Coupé, P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., and Collins, D. L. (2011). Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Efros, A. A. and Leung, T. K. (1999). Texture synthesis by non-parametric sampling. In *IEEE Int. Conf. on Computer Vision*, volume 2, pages 1033–1038.
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.
- Freeman, W. T., Jones, T. R., and Pasztor, E. C. (2002). Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *IEEE Int. Conf. on Computer Vision*, pages 2961–2969.
- Huang, X. and Yan, M. (2018). Nonconvex penalties with analytical solutions for one-bit compressive sensing. *Signal Processing*, 144:341–351.
- Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. In *Int. Conf. on Machine Learning*, pages 2127–2136.
- Ishida, T., Niu, G., Hu, W., and Sugiyama, M. (2017). Learning from complementary labels. *Advances in neural information processing systems*, 30.
- Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021). Perceiver: General perception with iterative attention. In *Int. Conf. on Machine Learning*, pages 4651–4664.
- Kanehira, A. and Harada, T. (2016). Multi-label ranking from positive and unlabeled data. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 5138–5146.
- Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y. (2021). General multi-label image classification with transformers. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 16478–16488.
- Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conf. on Computer Vision*, pages 740–755.
- Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., and Pietikäinen, M. (2019). From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109.
- Mac Aodha, O., Cole, E., and Perona, P. (2019). Presence-only geographical priors for fine-grained image classification. In *IEEE Int. Conf. on Computer Vision*, pages 9596–9606.
- Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2009). Classifier chains for multi-label classification. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 254–269.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 779–788.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE Int. Conf. on Computer Vision*, pages 618–626.
- Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. on Machine Learning*, pages 6105–6114.
- Trockman, A. and Kolter, J. Z. (2022). Patches are all you need? *arXiv preprint arXiv:2201.09792*.
- Varma, M. and Zisserman, A. (2008). A statistical approach to material classification using image patch exemplars. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(11):2032–2047.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

- Verelst, T., Rubenstein, P. K., Eichner, M., Tuytelaars, T., and Berman, M. (2022). Spatial consistency loss for training multi-label classifiers from single-label annotations. *arXiv preprint arXiv:2203.06127*.
- Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., and Yan, S. (2014). Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*.
- Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., and Girshick, R. (2021). Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400.
- You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. (2019). Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.
- Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. *Advances in Neural Information Processing Systems*, 30.
- Zhang, D., Zhang, H., Tang, J., Wang, M., Hua, X., and Sun, Q. (2020). Feature pyramid transformer. In *European Conf. on Computer Vision*, pages 323–339.
- Zhao, T., Zhang, N., Ning, X., Wang, H., Yi, L., and Wang, Y. (2022). Codedvtr: Codebook-based sparse voxel transformer with geometric guidance. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 1435–1444.

