

## Article

# Fine-Grained Cross-Modal Semantic Consistency in Natural Conservation Image Data from a Multi-Task Perspective

Rui Tao <sup>1,2</sup> , Meng Zhu <sup>3</sup>, Haiyan Cao <sup>2</sup> and Honge Ren <sup>1,4,\*</sup>

<sup>1</sup> College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China; trlx20@nefu.edu.cn

<sup>2</sup> College of Artificial Intelligence and Big Data, Hulunbuir University, Hulunbuir 021008, China; ske159@163.com

<sup>3</sup> College of Information Engineering, Harbin University, Harbin 150076, China; zhun913@163.com

<sup>4</sup> Heilongjiang Forestry Intelligent Equipment Engineering Research Center, Harbin 150040, China

\* Correspondence: nefu\_rhe@163.com

**Abstract:** Fine-grained representation is fundamental to species classification based on deep learning, and in this context, cross-modal contrastive learning is an effective method. The diversity of species coupled with the inherent contextual ambiguity of natural language poses a primary challenge in the cross-modal representation alignment of conservation area image data. Integrating cross-modal retrieval tasks with generation tasks contributes to cross-modal representation alignment based on contextual understanding. However, during the contrastive learning process, apart from learning the differences in the data itself, a pair of encoders inevitably learns the differences caused by encoder fluctuations. The latter leads to convergence shortcuts, resulting in poor representation quality and an inaccurate reflection of the similarity relationships between samples in the original dataset within the shared space of features. To achieve fine-grained cross-modal representation alignment, we first propose a residual attention network to enhance consistency during momentum updates in cross-modal encoders. Building upon this, we propose momentum encoding from a multi-task perspective as a bridge for cross-modal information, effectively improving cross-modal mutual information, representation quality, and optimizing the distribution of feature points within the cross-modal shared semantic space. By acquiring momentum encoding queues for cross-modal semantic understanding through multi-tasking, we align ambiguous natural language representations around the invariant image features of factual information, alleviating contextual ambiguity and enhancing model robustness. Experimental validation shows that our proposed multi-task perspective of cross-modal momentum encoders outperforms similar models on standardized image classification tasks and image–text cross-modal retrieval tasks on public datasets by up to 8% on the leaderboard, demonstrating the effectiveness of the proposed method. Qualitative experiments on our self-built conservation area image–text paired dataset show that our proposed method accurately performs cross-modal retrieval and generation tasks among 8142 species, proving its effectiveness on fine-grained cross-modal image–text conservation area image datasets.

**Keywords:** cross-modal; multi-task; image captioning; cross-modal retrieval; cross-modal alignment



**Citation:** Tao, R.; Zhu, M.; Cao, H.; Ren, H. Fine-Grained Cross-Modal Semantic Consistency in Natural Conservation Image Data from a Multi-Task Perspective. *Sensors* **2024**, *24*, 3130. <https://doi.org/10.3390/s24103130>

Academic Editor: Christoph M. Friedrich

Received: 11 March 2024

Revised: 3 May 2024

Accepted: 11 May 2024

Published: 14 May 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Neuro-networks function as parameterized databases, typically driven by specific tasks, with each network dedicated to fulfilling a corresponding task. However, there are instances where our requirements transcend single-task boundaries. Consider the context of rapidly accumulating natural conservation area image data. We seek not only to retrieve a single image but also to attach essential descriptions when summoning an image. Furthermore, we aspire to employ textual descriptions as queries to sift through our image repository, locating images that align with our specific needs. This scenario

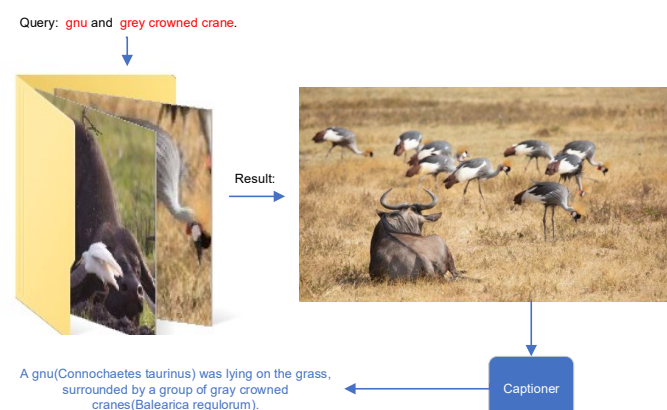
necessitates simultaneous engagement with two tasks: cross-modal image–text retrieval and image captioning.

As these data accumulate over time, the volume becomes formidable. For example, the Snapshot Serengeti Project at Serengeti National Park, Tanzania deployed hundreds of camera traps to understand the dynamics of African animal species. From 2010 to 2013, the project collected 3.2 million images from 225 camera traps [1]. And it was found to be very costly to manually process the images and add annotation labels, given such a large amount of data. The project carried out by Ref. [2] required thousands of technical volunteers to work for 2–3 months to annotate image data. With the improvement in camera manufacturing technology, each camera deployed in the field can record more than 40,000 photos per day due to a single trigger event [3], and many camera traps have been deployed in related projects. Refs. [4,5] deployed hundreds of camera traps in their project. Refs. [6,7] deployed about 50 cameras at water sources in natural conservation areas and recorded more than 800,000 wildlife images within a few weeks.

When we resort to two separate models to independently address these tasks, we encounter suboptimal outcomes. Specifically, the images retrieved through descriptive text queries may not align with the descriptive text generated by the model for the same image. In other words, these two models exhibit inconsistent encoding and decoding for the same data. Can we train a model that maintains consistency during both encoding and decoding, all while meeting the task requirements, thus mitigating semantic ambiguity within our cross-modal parameterized database?

To address this, we propose a multi-task model for joint training in cross-modal image–text retrieval and image captioning. Through the collaborative optimization of parameters, we achieve cross-module information sharing, thereby facilitating semantic-consistency encoding and decoding modeling. Post-training, the encoder and decoder can be independently employed to perform cross-modal image–text retrieval and image-captioning tasks while maintaining semantic consistency between the two tasks. This is made possible because our model is constructed upon a foundation of shared semantic-consistency representation space. Of course, the prerequisite is the construction of a dataset aligning with our specific needs and the judicious design of the model’s structure. For ease of exposition, we name the proposed method ReCap (Retrieval and Captioning).

As illustrated in Figure 1, we are able to retrieve corresponding images from the dataset using a customized textual input and subsequently generate descriptive text for the retrieved images. In this paper, our objective is to preserve semantic consistency in the context of fine-grained visual features and rich textual descriptions by jointly training a retriever and a captioner.



**Figure 1.** An application instance of the ReCap model.

The contributions of this work include (1) the creation of a dataset of image–text pairs for natural conservation; (2) proposing a combined offline and online training approach; (3) introducing a method for information transfer through collaborative parameter solving

within a multi-task module; and (4) presenting a technique for cross-modal alignment and semantic consistency preservation based on a shared representation space for cross-modal tasks.

## 2. Related Work

The cross-modal semantic consistency between images and text in our research is primarily achieved through the model design and joint training of two tasks: cross-modal retrieval and image captioning. The essence of this approach lies in the optimization of the cross-modal shared space embedding of images and text. On one hand, optimization is performed from the perspective of cross-modal alignment between image and text entities. On the other hand, the model needs to reorganize tokens related to the input image representation in the shared space in an autoregressive manner and output them in natural language, thereby achieving semantic consistency between image and text descriptions at a broader and deeper semantic level. The encoder and decoder constitute the core modules of our designed model, involving popular techniques in cross-modal alignment and cross-modal representation fusion. Subsequently, the literature review will delve into both cross-modal representation alignment and cross-modal representation fusion.

### 2.1. Cross-Modal Alignment

Currently, research on the cross-modal alignment of image and text representations is predominantly centered around contrastive learning methods. These studies achieve the embedding and alignment of image and text representations in a shared cross-modal space by training encoders separately for each modality using a contrastive learning loss. ConVIRT [8] demonstrates the potential of contrastive objectives to learn image representations from text. Inspired by ConVIRT, CLIP [9] performs pre-training on a dataset containing four billion image–text pairs and has become a milestone of vision–language models with excellent cross-modal representation. CLIP4Clip [10] demonstrates the CLIP model with high performance in cross-modal retrieval. ALIGN [11] performs pre-training on massive noisy web data. The above methods all use contrastive loss, which is the most effective loss for cross-modal alignment [12–15].

Intuitively, performing cross-modal contrastive learning by treating corresponding visual and textual entities as inputs to image and text encoders, respectively, can achieve better cross-modal alignment. Therefore, some research works in this domain utilize object detection models as visual unit extractors. The extracted target pixel regions are then fed to the image encoder for contrastive learning with the text encoder, enhancing the performance of cross-modal representations. Often, these studies require the integration of a pre-trained object detection model at the front end of the visual data input [16–18]. An intuitive approach is to align the visual features of the region where the object is located with the label. For example, Oscar [19] uses Faster R-CNN [20] to detect the object in the image and then aligns it with the word embeddings of the object tags. However, they are not suitable for fine-grained cross-modal alignment, as the object tags are too limited to align the vision features suitably. With a properly designed prompt, CLIP can be used for open-vocabulary classification, which solves the problem of limited object tags. ViLD [21] designed an open vocabulary object detection model by knowledge distillation from the CLIP. Ref. [22] achieved language-driven zero-shot semantic segmentation by directly using the representation of CLIP. Groupvit [23] implements unsupervised image segmentation by using the text representation of CLIP as a pseudo label.

Contrastive learning with dual encoders, while excelling in cross-modal retrieval tasks involving images and text, encounters challenges in adapting to fine-grained cross-modal retrieval tasks with natural conservation images due to the following reasons. First, certain species' visual features in natural conservation images exhibit high intra-class and inter-class similarities, resulting in dense distributions of these highly similar representations in the shared space. This necessitates encoders with finer discriminative capabilities. Second, these encoders, trained on image–text pair datasets using contrastive learning, are

often constrained by the representation of text descriptions alone and struggle to adapt well to cross-modal retrieval tasks where the semantics are similar but the expression methods differ.

## 2.2. Cross-Modal Fusion

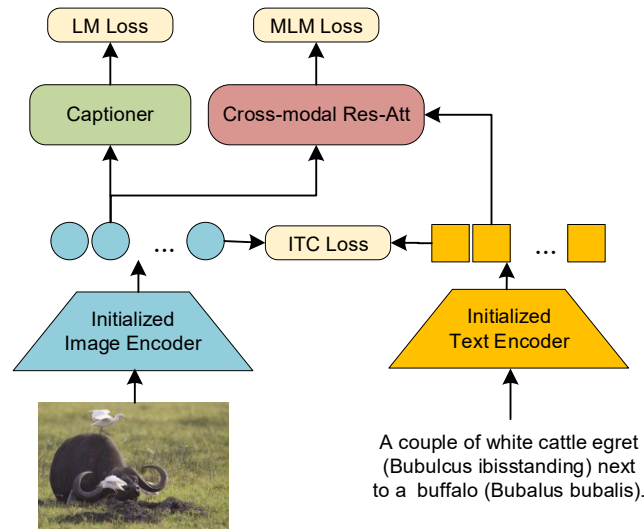
With the successful application of the transformer [24] architecture in the fields of natural language processing, computer vision, and multi-modal, ViLT [25] proposes a transformer-based multi-modal encoder which focuses on cross-modal feature fusion, and takes the masked language modeling loss [26] for visual embedding as future work. This work has been achieved by VL-BEiT [27] after ViT [28] and MAE [29]. From then on, a big convergence of language, vision, and multi-modal pretraining has emerged. BLIP [30] proposes a new vision–language pre-training framework that transfers flexibly to both vision–language understanding and generation tasks. The multi-way transformer proposed by BEiT-V3 [31] has achieved state-of-the-art transfer performance in both vision and vision–language tasks. FLIP [32], which is called Fast Language–Image Pre-training, presents a simple and more efficient method for training CLIP by dropping a part of masked tokens. VLMo [33] jointly learns a dual encoder and a fusion encoder with a modular Transformer network. Coca [34] is a minimalist design to pre-train an image–text encoder–decoder foundation model jointly with contrastive loss and captioning loss like CLIP and SimVLM [35], respectively.

Cross-modal feature fusion is not suitable for cross-modal retrieval tasks due to the lack of effective optimization for unimodal encoders. However, when applied to image-captioning tasks for the same input image, this method generates descriptions that share the same semantics but have different expressions. This indicates that such methods contribute to solving cross-modal semantic consistency. Our research goal is to explore the joint application of cross-modal feature fusion and cross-modal feature alignment, aiming to leverage their respective strengths and compensate for weaknesses, fostering mutual enhancement. This objective is emphasized in the Method section for in-depth discussion.

## 3. Design Concept and Proposed Methodology

The overarching design strategy is to develop and train a pair of image–text encoders that extract representations with cross-modal semantic consistency, and the feature point distribution in the shared space accurately reflects contextual relevance. Based on this strategy, we designed a pair of encoders for cross-modal contrastive learning, consisting of an image encoder and a text encoder. After considering computational costs and performance trade-offs, we chose to obtain a pair of encoders through distillation that can be freely modified according to the experimental requirements (refer to the Appendix A.1 for detailed information). To promote cross-modal semantic consistency, we introduced the method of momentum encoding. However, the input data for cross-modal momentum encoding come from different modalities and lack mutual information, making it challenging to maintain consistency. To address this issue, we adopted a multi-task perspective and utilized a residual attention network to fully integrate representations from both modalities before outputting the momentum encoding queue. Finally, we trained the cross-modal encoder using a contrastive learning approach with the obtained momentum encoding queue to achieve fine-grained cross-modal semantic consistent representations. The overall architecture of the proposed method is illustrated in Figure 2.

Before introducing the cross-modal momentum encoder, we first present the residual attention network and the design of the objective function.



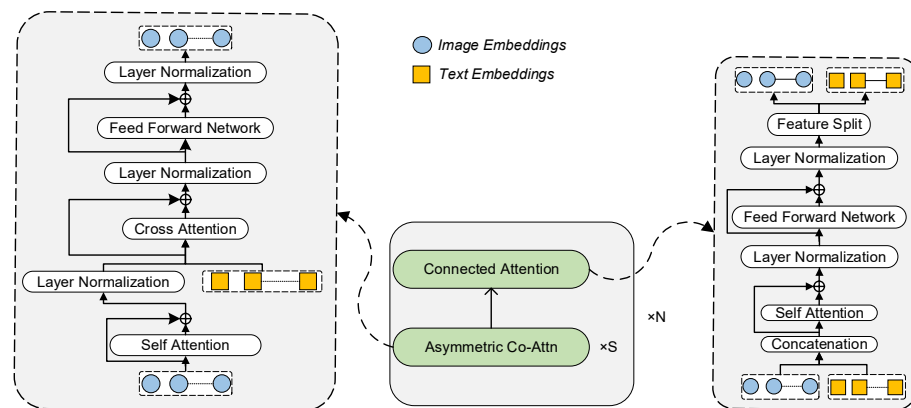
**Figure 2.** An overview of training ReCap for cross-modal semantics consistency.

### 3.1. Residual Attention Neuro-Network

Based on Reference [36], we designed a residual attention network as illustrated in Figure 3. For detailed derivation of its input and output, please refer to Appendix A.2. The primary training objective for Cross-modal Res-Att is masked language modeling (MLM). In this context, let us denote a caption as  $CnP$  and the set of randomly masked positions as  $M_{CnP}$ . The MLM loss can be formally defined as follows:

$$\mathcal{L}_{\text{MLM}} = - \sum_{i \in M_{CnP}} \log p(CnP_i | CnP_{\setminus M_{CnP}}), \quad (1)$$

where  $CnP_{\setminus M_{CnP}}$  is the masked version of the input caption, i.e., *Two [MASK] are [MASK] on [MASK] of a pond*. The cross-modal Res-Att module predicts the masked tokens based on the image and text context.



**Figure 3.** Residual Attention Network Architecture.

### 3.2. Captioner Training Objectives

The captioner is an autoregressive language generation model which operates on the principle of predicting the next token based on the input sequence and previously generated tokens. For example, given an initial input feature sequence  $F^k = F_1^k, \dots, F_u^k$  and the first token generated, denoted as  $C_1^k$ , the objective function is  $\log p_\theta(C_1^k | F^k)$ , and the

generation process for  $C_2^k$  is  $\log p_\theta(C_2^k|F^k, C_1^k)$ , and so on. Therefore, the objective function of an autoregressive language generation model is represented by Equation (2):

$$\mathcal{L}_{\text{LM}} = \max_{\theta} \sum_{k=1}^N \log p_\theta(C_1^k, \dots, C_m^k | F^k). \quad (2)$$

where  $\theta$  represents the trainable parameters, and the input sequence  $F^k$  can be visual features, language features, or a combination of both.

### 3.3. Image–Text Contrastive Loss Function

Following [8], the image–text contrastive learning (ITC) formulates the loss function according to InfoNCE [37]. Let  $T$  denote a certain species class embedding and  $V$  denote its visual embedding, then we have the embedding pair  $(V, T)$ . We use  $(V_i, T_i)$  to denote the  $i$ -th pair of positive samples and  $(V_i, T_j) j \neq i$  to denote a pair of negative samples. The ITC training objective of ReCap consists of two loss functions to make the distance of the positive pair closer than the negative one in the embedding space. Since ITC is asymmetric for each modality, it needs to be computed separately from both directions for images and text. The contrastive loss for the  $i$ -th pair in the image  $\rightarrow$  text direction:

$$\ell_i^{(V \rightarrow T)} = -\log \frac{\exp(\text{sim}(\mathbf{V}_i, \mathbf{T}_i)/\tau)}{\sum_{k=1}^N \exp(\text{sim}(\mathbf{V}_i, \mathbf{T}_k)/\tau)}, \quad (3)$$

where  $\text{sim}(\cdot)$  is the cosine similarity, i.e.,  $\text{sim}(a, b) = a^\top b / (\|a\| \|b\|)$ , and  $\tau$  is a temperature parameter. Similarly, we formulate the text  $\rightarrow$  image loss as:

$$\ell_i^{(T \rightarrow V)} = -\log \frac{\exp(\text{sim}(\mathbf{T}_i, \mathbf{V}_i)/\tau)}{\sum_{k=1}^N \exp(\text{sim}(\mathbf{T}_i, \mathbf{V}_k)/\tau)}. \quad (4)$$

Finally, the training objective is a weighted sum:

$$\mathcal{L}_{\text{ITC}} = \frac{1}{N} \sum_{i=1}^N (\lambda \ell_i^{(V \rightarrow T)} + (1 - \lambda) \ell_i^{(T \rightarrow V)}), \quad (5)$$

where  $\lambda \in [0, 1]$  is a hyperparameter weight, and  $N$  is the batch size.

### 3.4. Cross-Modal Momentum Encoder

In reference to the MoCo momentum encoding [12], we propose an offline encoder training method. Due to the high compression and ambiguity of textual information, compared to visual information, which is sparse, many detailed visual features are overwhelmed by dense textual information during cross-modal learning. To address this, we employ a residual attention network to repeatedly fuse visual features with textual features in a residual manner, increasing the proportion of visual information in the deep neuro-network’s forward channel. This enhances visual information redundancy to mitigate the drowning of sparse visual information during fusion with textual information. Additionally, because images contain factual information and exhibit invariance, aligning variable and ambiguous linguistic features around factual information contributes to eliminating linguistic feature ambiguity in context during cross-modal alignment. Consequently, this results in semantic consistency embedding, with visual information as the clustering center in the cross-modal representation space.

#### 3.4.1. Momentum Encoder

In brief, the principle of the momentum encoder is that the training of the encoder in unsupervised learning can be simplified as a look-up table problem. In other words, an encoded query should have high similarity to its corresponding key and low similarity to other keys. This simplifies the entire process to minimizing the contrastive loss. During the

solving process, contrastive learning requires a queue containing keys for both positive and negative samples to look up for queries. To maintain the consistency of encoding for positive and negative samples in the queue, the momentum encoder is employed.

The encoding update rule for the momentum encoder is shown in Equation (6), where the momentum parameter  $m \in [0, 1)$  is used. The query encoding  $\theta_q$  is updated based on gradient back propagation, while the key encoding  $\theta_k$  is updated using momentum. Typically,  $m$  takes a value greater than 0.9, which is equivalent to taking a moving average of the encoding updates. The slow-changing momentum encoder reduces the difference between the encoding of positive and negative samples in the queue, thereby improving the cross-task transfer performance of the encoder optimization process based on momentum in contrastive learning:

$$\theta \leftarrow m\theta_k + (1 - m)\theta_q \quad (6)$$

### 3.4.2. Offline Cross-Module Information Propagation

The cross-module joint solving of parameters constitutes the inter-module propagation of information. Deep learning models are essentially parameterized databases, with relationships among data implicitly encoded within the model's parameters. Therefore, cross-module operations on parameters represent the propagation of information across modules.

Firstly, as illustrated in Figure 4, we feed the image–text paired dataset to the unimodal encoders, obtaining image encodings ( $Ve_0, Ve_1, Ve_2 \dots$ ) through the ITC loss. Subsequently, as depicted in Figure 5, we feed ( $Ve_0, Ve_1, Ve_2 \dots$ ) to the Res-Att module. Based on the momentum encoding method proposed in Section 3.4.1, a cross-modal momentum encoding queue is obtained using the joint loss function shown in Equation (7). Specifically, we obtain a visual momentum encoding queue ( $Vm_0, Vm_1, Vm_2 \dots$ ) and a language momentum encoding queue ( $Tm_0, Tm_1, Tm_2 \dots$ ). Then, as shown in Figure 6, we feed back the momentum encodings to update the unimodal encoders:

$$\mathcal{L}_{\text{Res\&Cap}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{LM}}. \quad (7)$$

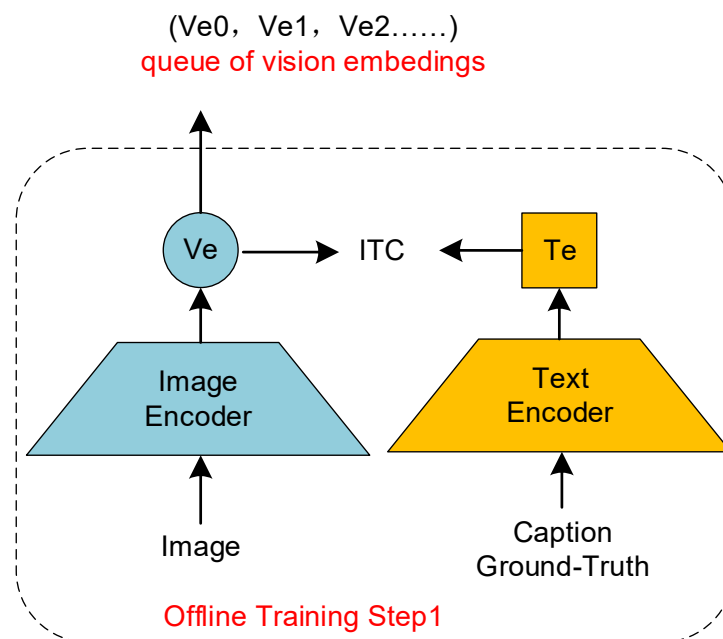


Figure 4. Initial visual encoding.

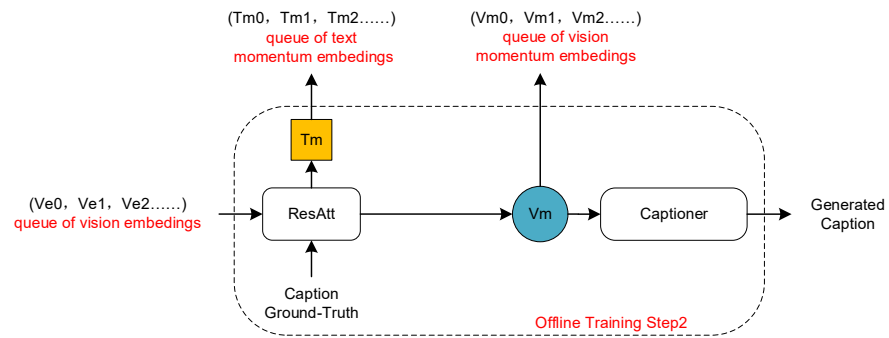


Figure 5. Redundant disambiguation momentum encoding.

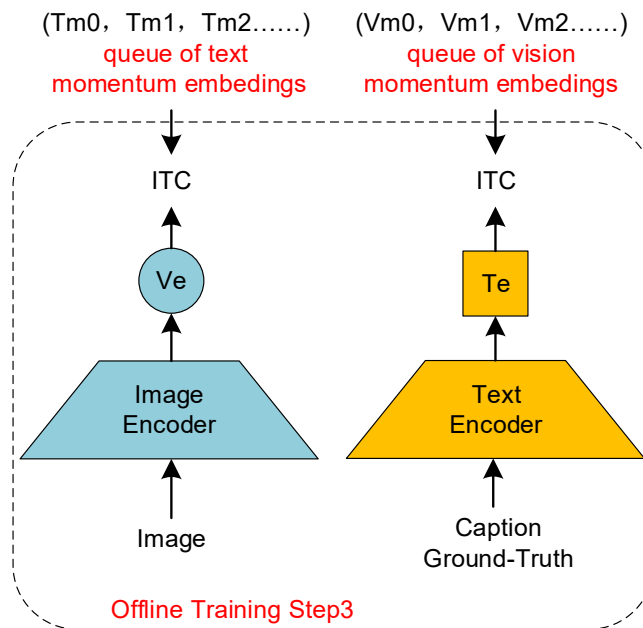


Figure 6. Unimodal encoding momentum update.

The loss function for the visual encoder in this training stage is represented by Equation (8), while the loss function for the language encoder is represented by Equation (9). The overall objective function is depicted by Equation (10):

$$\ell_i^{(V \rightarrow Tm)} = -\log \frac{\exp(\text{sim}(\mathbf{V}_i, \mathbf{Tm}_i) / \tau)}{\sum_{k=1}^N \exp(\text{sim}(\mathbf{V}_i, \mathbf{Tm}_k) / \tau)}, \quad (8)$$

$$\ell_i^{(T \rightarrow Vm)} = -\log \frac{\exp(\text{sim}(\mathbf{T}_i, \mathbf{Vm}_i) / \tau)}{\sum_{k=1}^N \exp(\text{sim}(\mathbf{T}_i, \mathbf{Vm}_k) / \tau)}, \quad (9)$$

$$\mathcal{L}_{\text{Momentum}} = \frac{1}{N} \sum_{i=1}^N \left( \lambda \ell_i^{(V \rightarrow Tm)} + (1 - \lambda) \ell_i^{(T \rightarrow Vm)} \right), \quad (10)$$

where  $\lambda \in [0, 1]$  is the hyperparameter weight, and  $N$  is the batch size. Repeating these steps forms a closed loop for cross-modal momentum encoder training, which can be conducted offline. It should be noted that our proposed offline training method needs to be accompanied by the decoupling of the momentum encoding queue we adopted. This decoupling allows for independent settings of batch size and the length of the momentum encoding queue. For instance, during training, we used a batch size of 32 and a queue length of 4096. This enabled us to contrast more negative samples, facilitating the model to learn representations closer to the domain distribution. The length of the queue can be



adjusted based on computational resources. In summary, the decoupling + offline strategy balances computational resources and model performance.

### 3.4.3. Why Contrastive Learning and Momentum Encoding

The objective of contrastive learning is specifically to distinguish between positive and negative samples. If the encodings of positive and negative samples come from different encoders or different training stages of the same encoder, the model may learn more about the differences between the encoders rather than the differences between the data. To ensure a fair comparison between positive and negative samples and optimize the features extracted by the encoder, consistency in the encoding of positive and negative samples needs to be maintained in a long queue. For example, in our model training, the length of the momentum encoding queue is set to 4096. Essentially, contrastive learning treats each sample as a multi-class classification task, thereby enhancing the flexibility of embedding cross-modal contextual information in a shared space. However, due to the inherent diversity and ambiguity of natural language expressions, ambiguity is inevitable. This challenge is particularly pronounced in image–text paired datasets, where the same image can be interpreted from various perspectives, leading to significantly different language descriptions with varying semantics. Therefore, there are significant challenges to achieving cross-modal semantic consistency in representation. From a model structure perspective, cross-modal representation is determined by the encoder, and the compression of data information by the encoder inevitably leads to information loss. This requires a balance between encoding efficiency and encoder performance.

Image captioning is a standardized task for cross-modal understanding, where the model generates corresponding language descriptions based on input image representations. The task inherently involves calculating the similarity between image and text representations, necessitating a shared semantic space for image–text cross-modal semantic alignment, similar to cross-modal retrieval tasks. In other words, sharing a semantic space is a fundamental prerequisite for both cross-modal generation and cross-modal retrieval tasks. When the factual information and diversity/ambiguity of natural language descriptions in images are projected into a shared semantic space, the goal is to enhance the mutual information between the two modal representations. As the mutual information between modal representations increases in this space, the performance of cross-modal retrieval and cross-modal generation models based on this representation space improves. To optimize cross-modal representation and shared space embedding for the captioner’s cross-modal understanding, we propose a multi-task perspective involving the joint training of image-captioning and cross-modal retrieval tasks. This approach ensures primary consistency in the shared representation space between the two tasks, thus facilitating improved cross-modal mutual information. If the two tasks are trained separately, although they may project into the same-dimensional space, they contain different information without an information-sharing process between the modalities, thus failing to effectively reduce discrepancies between the modalities. To establish an information channel, we employ dual momentum encoders. However, directly comparing the image momentum encoder and the text momentum encoder through contrastive learning faces challenges in ensuring cross-modal semantic consistency due to the different data properties between the two modalities. To synchronously and consistently update the momentum encodings of both modalities across modes, we propose using a residual attention network as a channel for cross-modal information exchange. Considering the sparsity of image data and the abstract nature of language data, we ensure that sparse data contribute proportionately to the information during the deep network’s feedforward process by using image features as residuals. Through cross-modal information fusion and momentum encoding, we obtain momentum encoding queues with higher cross-modal mutual information, resulting in better performance of the image–text encoder in cross-modal semantic consistency.



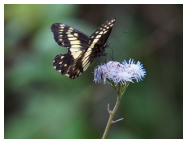

## 4. Experiments

Our method is named ReCap (Retrieval and Captioning). In this section, we primarily validate the effectiveness of our proposed method on standardization tasks using public datasets. Specifically, these tasks encompass image captioning and image–text retrieval on the COCO dataset, classification tasks on the iNaturalist2018 dataset, as well as image–captioning and image–text retrieval tasks on the iNaturalist2018 dataset.

### 4.1. Dataset Settings

We utilized the Karpathy split [38] of the MSCOCO dataset [39], comprising 123,000 images, with each image accompanied by five sentences as annotations. The iNaturalist2018 dataset comprises 8142 distinct species, each serving as an individual image classification category. It encompasses a total of 437,513 training images and 24,426 validation images. As this dataset initially lacked caption annotations, we conducted a comprehensive annotation effort, providing five sentences of description for each image. Furthermore, we annotated both the common name and the Latin name for each species. The specific process of enhancing the iNaturalist2018 dataset is detailed in Appendix A.3. In Table 1, we present some examples of our annotated data.

**Table 1.** Samples of nature conservation image–text pair dataset.

<b>Images</b>				
<b>Captions</b>	Two geese are walking on the shore of a pond.	A bunch of yellow flowers are sitting in a field.	A <i>Catantix nimbice</i> is sitting on an <i>Ageratum houstonianum</i> in the sun.	An <i>Aepyceros melampus</i> is grazing in a field.

### 4.2. Implementation Details

We utilized eight NVIDIA 3090 24G GPUs for the image–text encoder contrastive learning training process, with a queue length set to 4096 and a momentum parameter of 0.995. We employed the AdamW optimizer with a decay weight set to 0.02. The learning rate was the warm-up set to  $1 \times 10^{-4}$  for the first 1000 iterations and decayed in a cosine function manner to  $1 \times 10^{-5}$  for the subsequent iterations. The total training duration for the model was approximately 127 h.

### 4.3. Evaluation Metrics

The image caption model employs four widely recognized evaluation metrics, namely, BLEU (Bilingual Evaluation Understudy) [40], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [41], CIDEr (Consensus-based Image Description Evaluation) [42], and SPICE (Semantic Propositional Image-Captioning Evaluation) [43]. Among these, BLEU4 segments sentences into four-word chunks to gauge the descriptive accuracy of the model-generated captions. METEOR, building on the foundations of BLEU, addresses the issue of excessive word matching while emphasizing word recall and precision.

CIDEr, primarily applied in the domain of image description, employs TF-IDF (Term Frequency-Inverse Document Frequency) to weigh each sentence fragment. It encodes the frequency ( $E_r$ ) of a fragment in the reference description and the frequency ( $E_c$ ) in the generated description. Subsequently, it computes the similarity between  $E_r$  and  $E_c$  to generate an evaluation score for the model.

SPICE, on the other hand, is an evaluation metric based on scene graphs and semantic concepts. It assesses the extent to which the model-generated description aligns with the entities, attributes, and relationships present in the image.

The image classification task on iNaturalist has only one label for each picture, denoted as  $g_i$ . The result predicted by the model is denoted as  $p_i$ , and the error rate is

$$e_i = \min_i d(g_i, p_i), \quad (11)$$

where  $d(\cdot)$  is

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases}, \quad (12)$$

and the total score is

$$\text{score} = \frac{1}{N} \sum_i e_i. \quad (13)$$

#### 4.4. Experiment Project Selection

The core idea of our proposed method is briefly summarized as follows. Firstly, through the joint training of cross-modal retrieval and image-captioning tasks, we obtain a momentum-encoded queue with a contextual understanding of image–text pairs. This serves as an information bridge to train a cross-modal image encoder and a cross-modal text encoder using contrastive learning methods. This pair of encoders forms the basis for cross-modal fine-grained semantic consistency, as they determine the extraction and embedding of representations of various modal data into a shared cross-modal semantic space distribution. After training, our model yields an image encoder, a text encoder, and a captioner, which are the three key modules of ReCap. Due to the absence of a standardized task on a common dataset that can comprehensively evaluate our proposed method, we selected several standardized tasks on public datasets to individually test the performance of the three key modules of ReCap. Conducting experiments on standardized tasks on public datasets facilitates comparison with state-of-the-art (SOTA) methods on leaderboards, which, on the one hand, validates the effectiveness of the proposed method and, on the other hand, allows for a level measurement through comparison. Specifically, the experimental section validates the effectiveness of the captioner through the image-captioning task on the MSCOCO dataset as shown in Table 2. The effectiveness of the image encoder and text encoder’s cross-modal representations is verified through cross-modal retrieval tasks as shown in Table 3. The effectiveness of the image encoder is validated through the image classification task on the iNaturalist 2018 dataset as shown in Table 4. Additionally, Tables 2–4 in the experimental section reflect the proposed method’s multi-task perspective.

**Table 2.** Quantitative analysis of image captioning on MSCOCO dataset (%).

Method	B4	C	M	S
Oscar [19]	36.6	124.1	30.4	23.2
BUTD [44]	36.2	113.5	27.0	20.3
UnifiedVLP [45]	33.53	113.1	27.5	21.1
ClipCap [46]	33.5	113.1	27.5	21.1
<b>ReCap</b>	39.8	126.7	31.6	24.4

**Table 3.** Quantitative analysis of cross-modal retrieval on MSCOCO dataset (%).

Method	Retrieval I2T			Retrieval T2I		
	R@1	R@5	R@10	R@1	R@5	R@10
Oscar [19]	57.5	82.8	89.8	73.5	92.2	96.0
METER [47]	57.1	82.7	90.1	76.2	93.2	96.8
ViSTA [48]	52.6	79.6	87.6	68.9	90.1	95.4
ALADIN [49]	51.3	79.2	87.5	64.9	88.6	94.5
<b>ReCap</b>	65.5	89.2	92.9	77.1	92.6	96.3

**Table 4.** Comparison on Image Classification on iNaturalist 2018 (%).

Method	Top1 Accuracy
MetaFormer [50]	84.3
OMNIVORE [51]	84.1
RegNet-8GF [52]	81.2
VL-LTR [53]	81.0
$\mu$ 2Net+ [54]	81.0
MixMIM-L [55]	80.3
DeiT-B [56]	79.5
CeiT-s [57]	79.4
GPaCo [58]	78.1
<b>ReCap</b>	<b>85.1</b>

#### 4.5. Evaluation on the MSCOCO Dataset

We trained models on the MSCOCO dataset to perform image captioning and image–text retrieval tasks in order to validate the effectiveness of the proposed method. Table 2 presents the performance comparison of ReCap with state-of-the-art models in the context of image captioning. Here, B4 denotes BLEU-4, C represents CIDEr, M stands for METEOR, and S corresponds to SPICE. Further details are provided in Section 4.3. Table 3 illustrates the performance comparison of ReCap in image–text retrieval tasks against high-level models. Here, I2T denotes image-to-text retrieval, while T2I represents text-to-image retrieval. R@1, R@5, and R@10 respectively indicate recall rates for the top 1, top 5, and top 10 retrieval recommendations. The experimental results demonstrate that ReCap outperforms several state-of-the-art models, thereby validating the efficacy of the proposed method.

Based on the comparative data in Table 2, it is evident that ReCap demonstrates improved performance compared to others. Taking the scores in the B4 column as an example, the ReCap score is increased by nearly seven points. This improvement can be attributed to two main enhancements: firstly, the incorporation of an open vocabulary, meaning there is no restriction on the number of categories; and secondly, the Res-Att network excels in the fusion of cross-modal features, effectively emulating the representation style of the dataset. This results in a higher overlap between the generated captions and the ground truth.

As shown in Table 3, in the retrieval task of image to text, the R@1 score exhibits an improvement of approximately 8 to 14 percentage points compared to others. In the text-to-image retrieval task, there is an improvement of approximately 1 to 12 percentage points compared to others. This indicates a significant effect of the proposed method in the cross-modal alignment of image and text features. The improvement in text-to-image retrieval performance is relatively challenging due to the high information compression in textual data and the sparse nature of image data. When calculating mutual information, the same textual representation often exhibits similarity to a larger number of image representations. For instance, different models of cars appearing in images with similar backgrounds would have high similarity. To effectively differentiate between the brand and model of cars in the image, a finer-grained cross-modal alignment is required for text-to-image retrieval. Therefore, adopting an open vocabulary approach during the training of the image encoder is essential, as it avoids the limitations to a finite set of categories and proves crucial in the cross-modal modeling tasks involving image and text.

#### 4.6. Evaluation on the iNaturalist Dataset

In accordance with the introduction, the motivation behind this study is to address the need for the cross-modal processing of vast quantities of imagery data from natural conservation. In order to assess the cross-modal alignment of the model’s representations between images and text, we opted to employ the image classification task on the iNaturalist2018 dataset. This section’s experiments were conducted independently using the image encoder and text encoder. Notably, the image encoder was originally designed without a classification head. To achieve classification, we employed a method that involves compar-

ing the representations output by the image encoder with the prompt encodings generated by the text encoder.

The format of the prompts used is ‘a photo of <category>’, where ‘category’ corresponds to the category names in the dataset. In other words, for as many categories as there are in the dataset, there are corresponding prompts. In essence, our image classification approach assigns an image to the category with the highest similarity to its image representation. Specific experimental results are presented in Table 4. The experimental outcomes demonstrate that ReCap outperforms several state-of-the-art models, thereby confirming the the proposed method’s cross-modal alignment capability between image features and textual representations for species.

As shown in Table 4, ReCap demonstrates a performance improvement of approximately 1 to 7 percentage points compared to others. This indicates that our proposed method, employing an open vocabulary approach, is capable of handling image classification tasks on the iNaturalist Dataset. The experimental results not only affirm the effectiveness of our method in cross-modal representation alignment but also validate the feasibility of applying this approach to open vocabulary image classification tasks.

#### 4.7. Evaluation on the NACID Dataset

After verifying the effectiveness of the above, the model was trained on the NACID dataset Appendix A.3 and the two tasks of image captioning and image–text retrieval were evaluated. The model performance scores are shown in Tables 5 and 6. Through all the experimental results, it can be seen that the model has the ability to perform image captioning and image–text retrieval on the enhanced INaturalist2018 image–text pair dataset, which verifies the effectiveness of the ReCap model proposed in this paper.

**Table 5.** Quantitative analysis of image captioning on NACID (%).

Method	B4	C	M	S
ReCap	40.8	144.1	33.6	25.5

**Table 6.** Quantitative analysis of cross-modal retrieval on NACID (%).

Method	Text to Image			Image to Text		
	R@1	R@5	R@10	R@1	R@5	R@10
ReCap	72.8	89.1	93.2	82.0	96.6	98.3

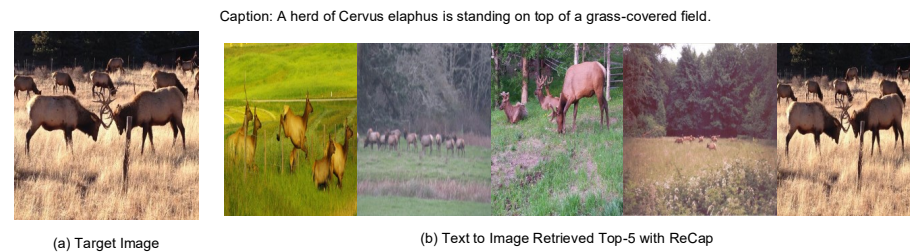
#### 4.8. Qualitative Evaluation

Next, we conducted qualitative experiments on cross-modal retrieval and generation using the NaCID test set. Additionally, to validate the effectiveness of the proposed method on natural protected area image datasets, we selected three image datasets from natural protected areas for zero-shot experiments.

The top 5 results for text-to-image retrieval are illustrated in Figure 7. Both non-target images and target images contain relevant content related to grassland and the target species. From the perspective of our application, we seek relatively open-ended retrieval results. This approach allows the model to continuously improve through small-sample learning in real-world applications. If the model were confined to strict one-to-one retrieval, it would lack practical utility.




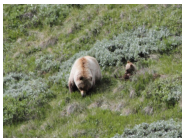
As shown in Table 7, the captions generated by the model align well with the content of the test images, and the species names are consistent with the Latin names used in the training set. This intuitively demonstrates the model’s learning capability in the domain of image–text cross-modal alignment. In the fourth prediction, the bear species (*Ursus arctos horribilis*) occurred 24 times in the training set, but there were no caption annotations for “cubs” in the training data prior to GPT-2 fine-tuning. This underscores the importance of pre-existing knowledge within NLP models for image-captioning tasks, as it can provide

additional information that is subsequently expressed in the form of generated descriptions. In the context of our approach, aligning image representations cross-modally in the pre-trained NLP decoder representation space leverages the rich knowledge of the NLP decoder for a deeper understanding of the images.






**Figure 7.** Examples of text-to-image retrieval on validation dataset.

**Table 7.** Examples sentences generated by ReCap for test images.

<b>Images</b>				
<b>Captions</b>	A few <i>Abudefduf saxatilis</i> swim in the stony water.	There are some red <i>Castilleja indivisa</i> in the grass.	A <i>Libellula quadrimaculata</i> is flying over the water.	A <i>Ursus arctos horribilis</i> and her cubs on a green field.

We conducted zero-shot experiments using three datasets related to natural conservations; refer to Appendix A Table A2. The experimental procedure was as follows: Firstly, we designed sentences resembling “A photo of <species>” based on the dataset content. Subsequently, we performed text-to-image retrieval with these sentences and provided the retrieved images to the captioner for generating descriptive text. The experimental results are presented in Table 8. The experimental results indicate that the species names on the retrieval side, the species within the images, and the species names on the generation side are all consistent. This observation underscores that the features extracted by the image encoder and text encoder are aligned, and the semantics of the encoder and decoder are in harmony, visually demonstrating the model’s capabilities in cross-modal alignment and semantic consistency between text and images. Examining the generated captions reveals the decoder’s capacity for systematic descriptions of foreground and background elements. This is a result of the combined influence of the model’s prior knowledge and fine-tuning.

**Table 8.** Examples of ReCap zero-shot retrieval and captioning.

Query	A photo of <i>Leopardus pardalis</i> .	A photo of <i>Phoenicopterus ruber</i> .	A photo of <i>Aglais io</i> .
Dataset	Wildlife Conservation Society	Birds 510 Species-Image Classification	Animals Detection Images Dataset
Result			
Caption	A small <i>Leopardus pardalis</i> walking through a forest at night.	A pink <i>Phoenicopterus ruber</i> standing in the water.	A close-up of an <i>Aglais io</i> is sitting on top of a flower.

#### 4.9. Ablation Study

The results of the ablation experiments are presented in Table 9. In the table, the term “C+C” indicates a direct connection between the encoder and captioner, where the visual representations generated by the encoder are used as input for the captioner. “C+R+C” signifies the bridging of encoder and captioner through the Res-Att module.

**Table 9.** Ablation study of ReCap on the MSCOCO and iNaturalist 2018 datasets.

Module Composition	MSCOCO			iNaturalist2018		
	I2T-R@1	T2I-R@1	Cap-B4	I2T-R@1	T2I-R@1	Cap-B4
C+C	51.5	75.2	31.9	54.1	68.9	32.3
C+R+C	51.3	75.7	35.3	53.7	69.5	36.1
ReCap	65.5	77.1	39.8	63.6	72.2	41.0

From the experimental results in the “C+C” row, it can be observed that the I2T and T2I performance on both datasets is relatively consistent, maintaining an average level. In comparison to the performance of ReCap, there is a slight decrease in T2I, while I2T and image captioning exhibit more substantial performance degradation. This suggests that when the encoder and decoder operate independently, the model’s performance heavily relies on the knowledge inherited from pre-trained models and the training process. However, without a channel for information transfer between them, they cannot leverage distinct task perspectives from each other to enhance each other’s performance.

Looking at the experimental results in the “C+R+C” row, there is a noticeable improvement in the performance of image captioning compared to the “C+C” row. This indicates that after a finer-grained cross-modal alignment of image and text representations at the micro-level, it becomes more favorable for the captioner to generate descriptions for images. It is evident that the Res-Att module significantly contributes to the optimization of cross-modal representation alignment and the refinement of shared semantic space embedding for text and images.

ReCap and the “C+R+C” configuration only differ in the presence of a momentum feedback loop in their model structures. From the experimental results, it is evident that there are overall performance improvements in the model, particularly in the I2T and image-captioning tasks. This suggests that the feedback information on the decoding side significantly aids in enhancing the performance of the encoder, resulting in substantial gains in the cross-modal alignment of image and text representations.

The improvement in image-captioning performance further illustrates that, after optimizing the encoder’s performance, it is possible to further enhance the decoder’s performance. From the perspective of data propagation, the encoder is at the front end, and the captioner is at the back end. With the addition of momentum feedback and Res-Att-based cross-modal fusion, the two form a feedback loop for mutual optimization.

## 5. Conclusions

The image–text representation initially undergoes coarse alignment through the encoder, followed by fine-grained alignment by the decoding side consisting of Res-Att and the captioner. Subsequently, the encoder is momentum updated based on the decoding side information, forming feedback from the decoding side to the encoding side, enhancing the quality of both the encoder and caption generation. The essence of this process lies in the sharing of a semantic space, where the decoder imparts its understanding of embedding similarities and categorization to the encoder. These insights are propagated to the encoder’s network parameters through momentum-based backpropagation. Furthermore, contrastive learning on the encoding side plays a crucial role. As mentioned earlier, the classification in contrastive learning is open-ended, with as many categories as there are samples. Such a classification method has no upper limit on granularity, compelling the encoder to learn subtle distinctions among samples as much as possible. Achieving this

solely from the encoding side would be information bottlenecked, and this is where feedback from the decoding side effectively bridges the information gap. Experimental results also confirm the contribution of prior knowledge in the decoder during this process. In summary, the feedback from the decoding side, the prior knowledge in the decoder, and momentum updates collectively enhance the quality of feature extraction in the encoder. All of this coalesces into a shared semantic space embedding for the encoder–decoder, where both entities possess a shared and aligned embedding space, embodying the essence of semantic consistency.

The performance of both cross-modal retrieval in image–text pairs and generative models fundamentally depends on the quality of shared space embeddings. The main contribution of our proposed method lies in the effective fusion of the advantages of both tasks in the cross-modal shared space embedding of images and text through thoughtful model design. This approach is particularly suitable for scenarios where there are strict alignment requirements between the objects in the image and the vocabulary in the text. Moreover, it demands that the model can further associate the input image representation with a more extensive and semantically rich textual description along a longer logical chain. Our proposed method is well suited for such scenarios.

**Author Contributions:** R.T. was responsible for model design, model training, dataset construction, and code writing and debugging. M.Z. was responsible for model design and code debugging. H.C. contributed to dataset annotation checking, article editing, and revision. H.R. contributed to conceptualization, methodology, draft writing, and reviewing, and provided the experimental conditions, including the artificial intelligence laboratory and experimental equipment. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Natural Science Foundation of Heilongjiang Province (LH2020F040), the Young Doctoral Research Initiation Fund Project of Harbin University “Research on Wood Recognition Methods Based on Deep Learning Fusion Model” Project (HUDF2022110), the Self-funded project of Harbin Science and Technology Plan Research on Computer Vision Recognition Technology of Wood Species Based on transfer learning Fusion Model Project (ZC2022ZJ010027), and the Fundamental Research Funds for the Central Universities (2572017PZ10).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

**Conflicts of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

## Appendix A

### *Appendix A.1. Encoder Design and Initialization*

We designed and initialized a pair of encoders, one for images and one for text, to extract representations of image and text data (the initialized image encoder and initialized text encoder as shown in Figure 2). These unimodal encoders serve as projectors that embed each modality into a shared semantic space.

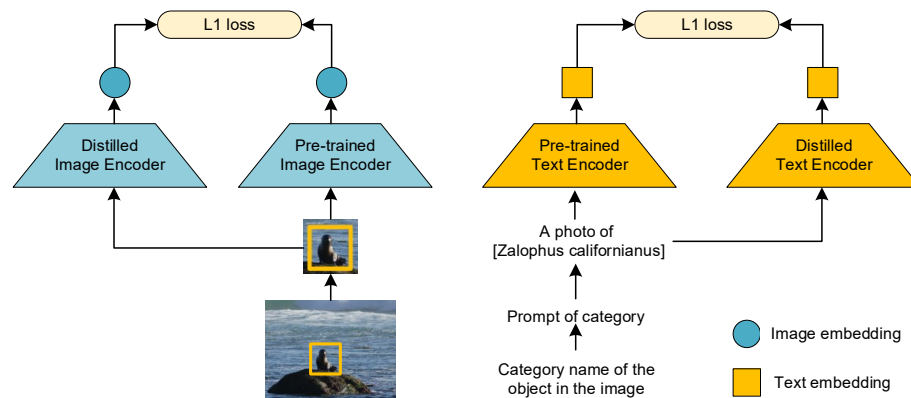
As illustrated in Figure A1, we leverage the knowledge from the pre-trained CLIP [9] model to initialize our lightweight transformer encoder. The encoder initialization is performed offline to reduce the computational requirements throughout the entire model training process. Distillation from the CLIP pre-trained encoder to the target encoder is achieved through the calculation of the L1 loss. Let the image encoder of the pre-trained CLIP model be denoted as  $V(\cdot)$ , and the text encoder as  $T(\cdot)$ . The distilled image encoder is denoted as  $D_V(\cdot)$ , and the distilled text encoder as  $D_T(\cdot)$ . The input image–text pairs are respectively denoted as  $i_n$  and  $t_n$ . The loss functions for distilling the image and text



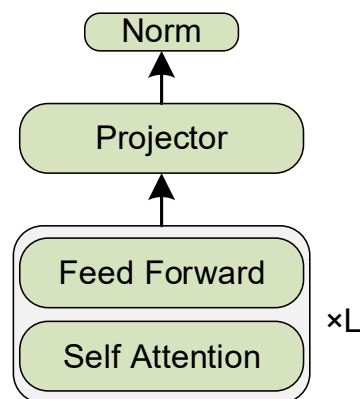
encoders are denoted as  $\mathcal{L}_V$  and  $\mathcal{L}_T$ , respectively. The  $\mathcal{L}_1$  loss for the model distillation is expressed as:

$$\begin{aligned}\mathcal{L}_V &= |D_V(i_n) - V(i_n)| \\ \mathcal{L}_T &= |D_T(t_n) - T(t_n)|.\end{aligned}\quad (\text{A1})$$

As illustrated in Figure A2, the unimodal encoder consists of  $L$  layers of stacked self-attention and feed-forward modules. The projector is employed to adjust the output dimensions of each module to ensure compatibility, while the normalization layer serves to balance the scale differences among various modal data, enhancing the model robustness and facilitating subsequent momentum calculations.



**Figure A1.** An example of knowledge distillation from a pre-trained model.



**Figure A2.** The structural details of the distilled encoder module.

#### Appendix A.2. Derivation Process of Input and Output of Residual Attention Network

In accordance with [36], we adopted the practice of concatenating every  $S$  Asymmetric Co-Attention (AC) block with a Connected Attention (CA) block, thereby creating a Cross-Modal Skip-Connection (CK) module. Furthermore, the Cross-modal Res-Att is concatenated with  $N$  CK modules. As visually represented in Figure 3, we represent the Self-Attention layer, Cross Attention layer, Feed Forward Network, Layer Normalization, and Concatenation layer as SA, CA, FFN, LN, and Cat, respectively. The image embedding is denoted as  $v = \{\mathcal{V}_{cls}, \mathcal{V}(I), \mathcal{V}(R_1), \dots, \mathcal{V}(R_i)\}$ , while the text embedding is represented as  $l = \{w_{cls}, w_1, \dots, w_n\}$ , consisting of word vectors corresponding to the input caption paired with  $I$ . Here, ' $I$ ' signifies the input image, ' $R_i$ ' refers to the  $i$ -th patch within it, and an additional [CLS] token is utilized to summarize the input sequence. Let  $I^{S-1}$ ,  $v^{S-1}$  and  $I^S$  represent the input word vectors, visual features, and output of the  $S$ -th AC layer, respectively. Then,

$$I_{SA}^S = LN\left(SA\left(I^{S-1}\right) + I^{S-1}\right), \quad (\text{A2})$$

$$l_{CA}^S = LN\left(CA\left(l_{SA}^S, v^{N-1}\right) + l_{SA}^S\right), \quad (A3)$$

$$l^S = LN(FFN(l_{CA}^S) + l_{CA}^S). \quad (A4)$$

Subsequently, we feed both  $l^S$  and  $v^{N-1}$  into a CA block to facilitate cross-modal information interaction. The computation  $[v^N; l^S]$  of the CK module's output is denoted as follows:

$$[v^{N-1}; l^{N-1}] = Cat([v^{N-1}, l^S]) \quad (A5)$$

$$[v_{SA}^N; l_{SA}^N] = LN(SA([v^{N-1}; l^{N-1}]) + [v^{N-1}; l^{N-1}]), \quad (A6)$$

$$[v^N; l^N] = LN(FFN([v_{SA}^N; l_{SA}^N]) + [v_{SA}^N; l_{SA}^N]). \quad (A7)$$

### Appendix A.3. NaCID Dataset

We utilized the iNaturalist2018 dataset [59], which consists of 8142 species, with 437,513 training set images and 24,426 validation set images. However, the dataset does not include text descriptions corresponding to the images. In order to generate text descriptions paired with images, we followed the pipeline of Laion COCO 600M [60] to curate our **Nature Conservation Image-text Pair Dataset (NaCID)** in four steps: (1) using BLIP L/14 to generate 40 captions for each image in iNaturalist dataset; (2) ranking them using Open AI CLIP L/14 to select the best five captions; (3) using Open AI RN50x64 CLIP model to select the best one; and (4) using a small, fine-tuned T0 [61] model to roughly repair the grammar and punctuation of the texts.

We obtained a dataset consisting of natural images and paired text descriptions which are called captions. After that, we used the spaCy [62] method to recognize the predefined span types related to the categories of animals and plants. Then, we followed the pipeline of entity name replacement [63] to further annotate the entities in captions with the fine-grained species names supported by the image classification ground truth of the iNaturalist dataset, such as *Heterotheca subaxillar*, *Ageratum houstonianum* etc. The entity definitions are shown in Table A1, where AML represents animals, and ANT represents plant classification.

**Table A1.** Applicable metadata for each entity type.

Entity Type	Applicable Types of Perturbable Spans
AML	<Animal-quantity> (e.g., a dog, two cats)
ANT	<Plant-quantity> (e.g., an apple, flowers)

### Appendix A.4. Three Datasets for Zero-Shot Experiments

Three conservation area datasets were used to test the cross-dataset robustness.

**Table A2.** Three conservation area datasets used for zero-shot experiments.

Dataset Names	Download URLs
birds 525 species	<a href="https://www.kaggle.com/datasets/gpiosenka/100-bird-species">https://www.kaggle.com/datasets/gpiosenka/100-bird-species</a>
Animals Detection Images Dataset	<a href="https://www.kaggle.com/datasets/antoreepjana/animals-detection-images-dataset">https://www.kaggle.com/datasets/antoreepjana/animals-detection-images-dataset</a>
Wildlife Conservation Society	<a href="https://library.wcs.org/Library/Science-Data/Datasets.aspx">https://library.wcs.org/Library/Science-Data/Datasets.aspx</a>

## References

1. Matin, M.; Shrestha, T.; Chitale, V.; Thomas, S. Exploring the potential of deep learning for classifying camera trap data of wildlife: A case study from Nepal. In Proceedings of the AGU Fall Meeting Abstracts, New Orleans, LA, USA, 13–17 December 2021; p. GC45I-0923.
2. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.S.; Packer, C.; Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E5716–E5725. [[CrossRef](#)]
3. Zett, T.; Stratford, K.J.; Weise, F. Inter-observer variance and agreement of wildlife information extracted from camera trap images. *Biodivers. Conserv.* **2022**, *31*, 3019–3037. [[CrossRef](#)]
4. Swanson, A.; Kosmala, M.; Lintott, C.; Simpson, R.; Smith, A.; Packer, C. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci. Data* **2015**, *2*, 1–14. [[CrossRef](#)]
5. McShea, W.J.; Forrester, T.; Costello, R.; He, Z.; Kays, R. Volunteer-run cameras as distributed sensors for macrosystem mammal research. *Landsc. Ecol.* **2016**, *31*, 55–66. [[CrossRef](#)]
6. Edwards, S.; Portas, R.; Hanssen, L.; Beytell, P.; Melzheimer, J.; Stratford, K. The spotted ghost: Density and distribution of serval *Leptailurus serval* in Namibia. *Afr. J. Ecol.* **2018**, *56*, 831–840. [[CrossRef](#)]
7. Stratford, K.; Stratford, S.; Périquet, S. Dyadic associations reveal clan size and social network structure in the fission–fusion society of spotted hyaenas. *Afr. J. Ecol.* **2020**, *58*, 182–192. [[CrossRef](#)]
8. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive learning of medical visual representations from paired images and text (2020). *arXiv* **2020**, arXiv:2010.00747. [[CrossRef](#)].
9. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Proceedings of Machine Learning Research, Virtual, 18–24 July 2021; pp. 8748–8763.
10. Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; Li, T. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* **2022**, *508*, 293–304. [[CrossRef](#)]
11. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, Proceedings of Machine Learning Research, Virtual, 18–24 July 2021; pp. 4904–4916.
12. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 16–18 June 2020; pp. 9729–9738.
13. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
14. Li, J.; Zhou, P.; Xiong, C.; Hoi, S.C. Prototypical Contrastive Learning of Unsupervised Representation. In Proceedings of the International Conference on Learning Representations, ICLR2021, Virtual, 3–7 May 2021.
15. Li, J.; Xiong, C.; Hoi, S. MoPro: Webly Supervised Learning with Momentum Prototypes. In Proceedings of the International Conference on Learning Representations, ICLR2021, Virtual, 3–7 May 2021.
16. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 104–120.
17. Xu, X.; Wang, T.; Yang, Y.; Zuo, L.; Shen, F.; Shen, H.T. Cross-modal attention with semantic consistence for image–text matching. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 5412–5425. [[CrossRef](#)]
18. Diao, H.; Zhang, Y.; Ma, L.; Lu, H. Similarity reasoning and filtration for image-text matching. In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Virtual, 2–9 February 2021; pp. 1218–1226.
19. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–137.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
21. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2021**, arXiv:2104.13921. [[CrossRef](#)].
22. Li, B.; Weinberger, K.Q.; Belongie, S.; Koltun, V.; Ranftl, R. Language-driven semantic segmentation. *arXiv* **2022**, arXiv:2201.03546. [[CrossRef](#)].
23. Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 18134–18144.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: San Jose, CA, USA, 2017.

25. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 5583–5594.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805. [[CrossRef](#)].
27. Bao, H.; Wang, W.; Dong, L.; Wei, F. Vi-beit: Generative vision-language pretraining. *arXiv* **2022**, arXiv:2206.01127. [[CrossRef](#)].
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [[CrossRef](#)].
29. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022, pp. 16000–16009.
30. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
31. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv* **2022**, arXiv:2208.10442. [[CrossRef](#)].
32. Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; He, K. Scaling Language-Image Pre-training via Masking. *arXiv* **2022**, arXiv:2212.00794. [[CrossRef](#)].
33. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Piao, S.; Wei, F. VLMO: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In *Proceedings of the Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: San Jose, CA, USA, 2022; pp. 32897–32912.
34. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv* **2022**, arXiv:2205.01917. [[CrossRef](#)].
35. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv* **2021**, arXiv:2108.10904. [[CrossRef](#)].
36. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv* **2022**, arXiv:2205.12005. [[CrossRef](#)].
37. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748. [[CrossRef](#)].
38. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
39. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
40. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
41. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
42. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
43. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 382–398.
44. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
45. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; Gao, J. Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 13041–13049.
46. Mokady, R.; Hertz, A.; Bermano, A.H. Clipcap: Clip prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734. [[CrossRef](#)].
47. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18166–18176.
48. Cheng, M.; Sun, Y.; Wang, L.; Zhu, X.; Yao, K.; Chen, J.; Song, G.; Han, J.; Liu, J.; Ding, E.; et al. ViSTA: Vision and scene text aggregation for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5184–5193.
49. Messina, N.; Stefanini, M.; Cornia, M.; Baraldi, L.; Falchi, F.; Amato, G.; Cucchiara, R. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. In Proceedings of the 19th International Conference on Content-Based Multimedia Indexing, Graz, Austria, 14–16 September 2022; pp. 64–70.

50. Diao, Q.; Jiang, Y.; Wen, B.; Sun, J.; Yuan, Z. Metaformer: A unified meta framework for fine-grained recognition. *arXiv* **2022**, arXiv:2203.02751. [[CrossRef](#)].
51. Girdhar, R.; Singh, M.; Ravi, N.; van der Maaten, L.; Joulin, A.; Misra, I. Omnivore: A single model for many visual modalities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16102–16112.
52. Touvron, H.; Sablayrolles, A.; Douze, M.; Cord, M.; Jégou, H. Grafit: Learning fine-grained image representations with coarse labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 874–884.
53. Tian, C.; Wang, W.; Zhu, X.; Dai, J.; Qiao, Y. VI-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 73–91.
54. Gesmundo, A. A Continual Development Methodology for Large-scale Multitask Dynamic ML Systems. *arXiv* **2022**, arXiv:2209.07326. [[CrossRef](#)].
55. Liu, J.; Huang, X.; Liu, Y.; Li, H. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv* **2022**, arXiv:2205.13137. [[CrossRef](#)].
56. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 10347–10357.
57. Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; Wu, W. Incorporating convolution designs into visual transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, Montreal, QC, Canada, 10–17 October 2021; pp. 579–588.
58. Cui, J.; Zhong, Z.; Tian, Z.; Liu, S.; Yu, B.; Jia, J. Generalized Parametric Contrastive Learning. *arXiv* **2022**, arXiv:2209.12400. [[CrossRef](#)].
59. Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S. The iNaturalist Species Classification and Detection Dataset. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 8769–8778.
60. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv* **2022**, arXiv:2210.08402. [[CrossRef](#)].
61. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T.L.; Raja, A.; et al. Multitask prompted training enables zero-shot task generalization. *arXiv* **2021**, arXiv:2110.08207. [[CrossRef](#)].
62. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. *SpaCy: INDUSTRIAL-Strength Natural Language Processing in Python*; Zenodo: Honolulu, HI, USA, 2020.
63. Yan, J.; Xiao, Y.; Mukherjee, S.; Lin, B.Y.; Jia, R.; Ren, X. On the Robustness of Reading Comprehension Models to Entity Renaming. *arXiv* **2021**, arXiv:2110.08555. [[CrossRef](#)].

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.