applied sciences

MDPI

Article

Long-Tailed Metrics and Object Detection in Camera Trap Datasets

Wentong He [1,2], Ze Luo [1,*], Xinyu Tong [1,2], Xiaoyi Hu [1,2], Can Chen [1] and Zufei Shu [3]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; hwt0316@cnic.cn (W.H.);xytong@cnic.cn (X.T.); huxiaoyi@cnic.cn (X.H.); chencan@cnic.cn (C.C.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Guangdong Chebaling National Nature Reserve, Shaoguan 512528, China; szfcbl@163.com
* Correspondence: luoze@cnic.cn

Abstract: With their advantages in wildlife surveys and biodiversity monitoring, camera traps are widely used, and have been used to gather massive amounts of animal images and videos. The application of deep learning techniques has greatly promoted the analysis and utilization of camera trap data in biodiversity management and conservation. However, the long-tailed distribution of the camera trap dataset can degrade the deep learning performance. In this study, for the first time, we quantified the long-tailedness of class and object/box-level scale imbalance of camera trap datasets. In the camera trap dataset, the imbalance problem is prevalent and severe, in terms of class and object/box-level scale. The camera trap dataset has worse object/box-level scale imbalance, and too few samples of small objects, making deep learning more challenging. Furthermore, we used the BatchFormer module to exploit sample relationships, and improved the performance of the general object detection model, DINO, by up to 2.9% and up to 3.3% in terms of class imbalance and object/box-level scale imbalance. The experimental results showed that the sample relationship was simple and effective, improving detection performance in terms of class and object/box-level scale imbalance, but that it could not make up for the low number of small objects in the camera trap dataset.

Keywords: camera trap; long-tailed metrics; class imbalance; object/box-level scale imbalance; deep learning; object detection; sample relationship

Citation: He, W.; Luo, Z.; Tong, X.; Hu, X.; Chen, C.; Shu, Z. Long-Tailed Metrics and Object Detection in Camera Trap Datasets. *Appl. Sci.* **2023**, *13*, 6029. https://doi.org/10.3390/app13106029

Academic Editor: Antonio Fernández-Caballero

Received: 7 April 2023
Revised: 10 May 2023
Accepted: 12 May 2023
Published: 14 May 2023

1. Introduction

Compared to traditional wildlife monitoring methods, camera traps offer advantages such as low cost and high concealment, allowing wild animals to be monitored and surveyed automatically and without disturbance [1–4]. Equipped with motion or infrared sensors, such cameras automatically capture images or videos of passing animals [5,6], resulting in billions of images/videos being collected annually worldwide [7]. The massive amount of raw camera trap data contains valuable information on the species, sex, age, health, number, behaviors, and locations of animals, making camera traps an indispensable tool for conservation and management efforts [8,9]. However, manual extraction of this information is a laborious, knowledge-intensive, time-consuming, and expensive endeavor. To address these challenges, researchers have turned to deep learning—a technique that enables computers to learn multiple levels of abstraction from images [10]. Many researchers have attempted to use deep learning techniques to detect and classify animals in camera trap images [11,12], with some yielding promising results that showcase the potential of these methods, which can speed up complex tasks, such as species recognition and individual counting [13].

Deep learning models learn the intrinsic features of the training data by updating the model parameter weights, and subsequently outputting inference results on unseen data;

therefore, the distribution of data has a crucial impact on the model performance. As a real-world dataset, the camera trap dataset typically exhibits imbalanced distribution, in which certain classes contain numerous images, while others contain only a few, resulting in a long-tailed distribution [14]. Deep learning models often perform well for species that are frequently captured by the camera trap, but may perform poorly for species that are rarely captured, especially for rare or endangered species with inherently low population sizes. Such a class imbalance makes the training of deep learning models in long-tailed camera trap datasets highly challenging [15]. In recent years, several datasets have been proposed for different long-tailed learning tasks. For long-tailed image classification, ImageNet-LT [16], CIFAR-10/100-LT [17], Places-LT [16], and iNaturalist 2018 [18] are four benchmark datasets. Meanwhile, LVIS0.5/1.0 [19] is the most widely used benchmark dataset for long-tailed object detection and instance segmentation. To investigate the long-tailed problem more effectively, researchers have used quantitative metrics, such as Imbalance Factor [16–18] and the Gini Coefficient [20], to precisely assess the degree of long-tailedness in the aforementioned datasets. Lu et al. [20] demonstrated the reasonability and effectiveness of the Gini coefficient, and revealed significant variations in long-tailedness among the existing long-tailed datasets. This brings us to the following questions: what is the degree of class imbalance in camera trap datasets; is the class imbalance prevalent and severe? Until now, there have been no studies investigating the quantitative metrics of long-tailedness in camera trap datasets.

Object detection is the task of locating and classifying objects in an image, using bounding boxes [21]. It is a critical problem in computer vision, and has many applications, such as autonomous driving and pedestrian detection [22–24]. Compared to object detection in a balanced dataset, long-tailed object detection is more complex and challenging, primarily due to the presence of an extreme class imbalance [25]. This imbalance often results in the detection loss or incorrect classification of rare classes, leading to poor overall detection performance. Thus, long-tailed object detection requires the development of appropriate strategies to address the imbalance issue. It is common knowledge that species with similar characteristics—particularly those within the same class, family or genus—tend to share similar body parts: for instance, squirrels generally share similar body and tail shapes. Transferring this shared knowledge between images of different species, to mitigate the effects of class imbalance, can help improve long-tailed object detection ability, especially for rare classes. Exploiting the invariant features between images belonging to the same species is also helpful. The diverse and firm sample relationships in camera trap datasets can be utilized to alleviate the issue of insufficient images for rare classes. Hou et al. [26,27] devised a simple yet effective batch transformer block that enables deep recognition and object detection models to explore sample relationships from the perspective of the batch dimension, facilitating the transfer of knowledge from frequent classes to rare ones. Though they extensively evaluated the effectiveness of this approach on multiple benchmark long-tailed datasets, it has not been tested on any camera trap datasets to date.

The distribution of object sizes can also impact the performance of object detection models, particularly with respect to small objects that usually have indistinguishable features and limited context information [28–30]. Oksuz et al. [25] observed a skewness in the distribution in favor of small objects in the MS-COCO dataset, and defined certain sizes of objects or input bounding boxes over-represented as an object/box-level scale imbalance. From the perspective of object/box-level scale imbalance, what is the distribution of the animal objects sizes in camera trap datasets? Is the distribution similar to the MS-COCO dataset or the long-tailed dataset, in which certain sizes of animal objects have a significant number of images, while others have very few? Is exploiting sample relationships also effective at improving the detection performance for different sizes of animal objects? Thus far, no research has been conducted to answer these questions.

In this work, we focused on the long-tailed metrics and object detection in camera trap datasets. The main contributions of this paper are as follows: (1) We utilized four commonly used metrics to quantitatively analyze the class imbalance in several camera

trap datasets, revealing that it was more severe than in the benchmark long-tailed dataset. (2) We also quantitatively analyzed the object/box-level scale imbalance in camera trap datasets, revealing that it too was more severe than in benchmark datasets. Moreover, the camera trap datasets contained too few small objects, making the object detection more challenging. (3) For the first time, we utilized a simple yet effective module, named BatchFormer (Batch transFormer), to explore the effectiveness of sample relationships. The results demonstrated that exploiting sample relationships can improve object detection performance in long-tailed camera trap datasets, in terms of class and object/box-level scale imbalance.

The structure of this paper is as follows. In Section 2, we present the Materials and Methods used in this study. The camera trap dataset details, several metrics, and object detection experiments results for multiple camera trap datasets are presented in Section 3. In Section 4, we discuss our findings, and look ahead to future work. In Section 5, we present the conclusions.

2. Materials and Methods

2.1. Camera Trap Datasets

In this study, multiple camera trap datasets were used: these were obtained from the public data of the LILA BC website and the collation of the data collected by camera traps situated at Chebaling National Nature Reserve in GuangDong, China.

LILA BC, also known as the Labeled Information Library of Alexandria: Biology and Conservation, is a data repository website that aims to provide rich data, especially labeled images, for Biology and Conversation-related research, using machine learning algorithms. The website currently hosts over ten million labeled images, including massive camera trap labeled images [31]. We selected 11 public camera trap datasets from this website, including Orinoquia Camera Traps (Orinoquia) [32], SWG Camera Traps 2018-2020 (SWG) [33], Island Conservation Camera Traps (Island) [34], Snapshot Karoo (Karoo) [35], Snapshot Kgalagadi (Kgalagadi) [36], Snapshot Enonkishu (Enonkishu) [37], Snapshot Camdeboo (Camdeboo) [38], Snapshot Mountain Zebra (Zebra) [39], Snapshot Kruger (Kruger) [40], Snapshot Serengeti (Serengeti) [41], and WCS Camera Traps (WCS) [42]. These datasets were collected from conservation projects conducted worldwide. The number of images ranged from thousands to hundreds of thousands, and the number of species varied from a few to hundreds. The annotations were organized in COCO Camera Traps format [43]. For the object detection experiments, we chose SWG Camera Traps 2018–2020 and WCS Camera Traps, based on whether they were labeled with the bounding box, the number of species or the number of images. SWG Camera Traps 2018–2020 were collected from 982 locations in Vietnam and Lao. Labels were provided for 120 categories, containing more than 80,000 bounding box annotations. WCS Camera Traps were collected from 12 countries. Labels were provided for 675 categories, and contained approximately 360,000 bounding box annotations.

In this study, we constructed one camera trap dataset, named CBL Camera Traps (CBL): this was obtained from the Chebaling national reserve, and included about 40,000 images of 69 species. The pipeline of constructing this dataset was as follows: firstly, each image captured by the camera trap was identified by zoological experts; then, the MegaDetector pre-trained model was used to obtain the border coordinates of the animal object in the image; finally, the images with inaccurate border coordinates were manually eliminated. This dataset was in COCO Camera Traps format. Some examples of some species from the 12 camera trap datasets are shown in Figure 1.

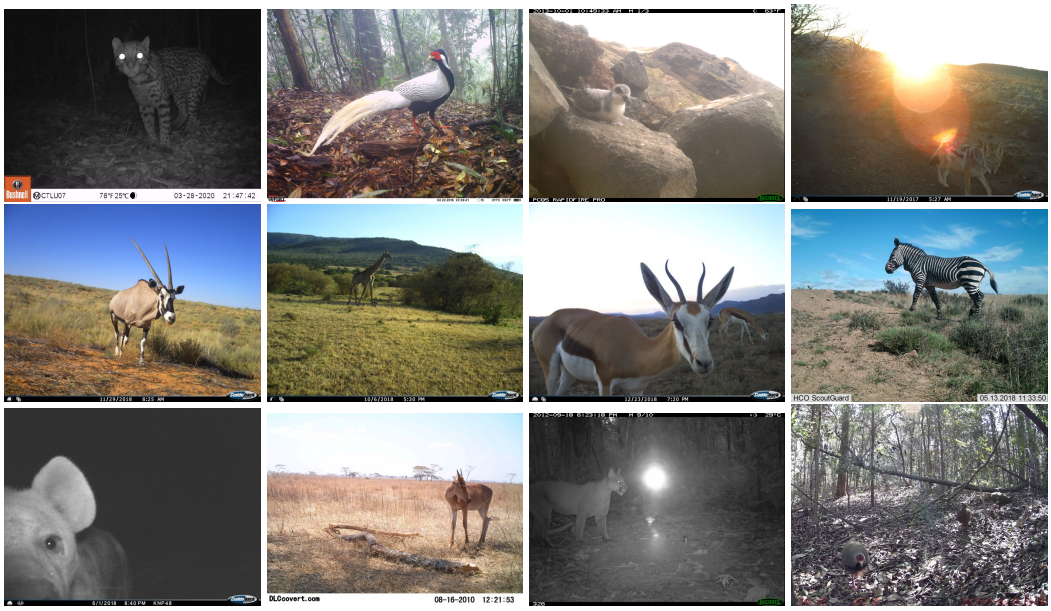

Figure 1. Examples of some species from the 12 Camera Trap Datasets.

2.2. Long-Tailedness Metrics

Deep learning models need to learn abstract features from large amounts of camera trap images, for animal classification and detection; therefore, the model performance heavily depends on the distribution of the dataset. Accurately and objectively quantifying the degree of long-tailedness in camera trap datasets is an important prerequisite for research and practical applications. By referring to existing studies [16–18,20], we exploited four commonly used metrics: the Gini Coefficient (denoted as GC); the Imbalance Factor (denoted as IF); Standard Deviation (denoted as SD); and the Mean/Median (denoted as MM), to measure the long-tailedness of multiple camera trap datasets.

2.2.1. Gini Coefficient

As the long-tailed distribution of data is similar to income distribution inequality, Yang et al. used the Gini coefficient—which was first proposed by Gini [44], and is often used to evaluate the degree of income imbalance—to quantify long-tailedness. The most important step in the calculation of the Gini coefficient is to obtain the Lorentz curve according to the number of samples of each class. Firstly, we sorted the number of samples of K classes dataset C_i, ($i = 1, 2, \ldots, K$) in ascending order, and calculated the normalized cumulative distribution D_i, as follows:

$$D_i = \frac{1}{K}\sum_{j=1}^{i}(C_i) \tag{1}$$

where K represents the number of species, and C_i represents the sampled number of species i, and is synonymous in Equations (4)–(6). Next, we normalized the x-axis as the ratio of the category index to the total categories, and obtained the Lorenz curve $L(x)$ through linear interpolation, as shown in Figure 2. The Lorenz curve $L(x)$ can be expressed as:

$$L(x) = \begin{cases} D_i, & x = \frac{i}{K} \\ D_i + (D_{i+1} - D_i)(Kx - i), & \frac{i}{K} < x < \frac{i+1}{K} \end{cases} \tag{2}$$

where $i = 1, 2, \ldots, K$. For the balanced dataset, the Lorenz curve is the line of equality. Finally, the Gini coefficient can be intuitively calculated as follows:

$$GC = \frac{A}{B} \tag{3}$$

where A represents the inequality between the line of equality and the Lorenz curve, and B is the triangle area representing the equality marked by blue. The Gini Coeffient has a bounded distribution, which ranges from 0 to 1. Usually, the greater the Gini Coefficient of one dataset, the more imbalanced the dataset, and vice versa.

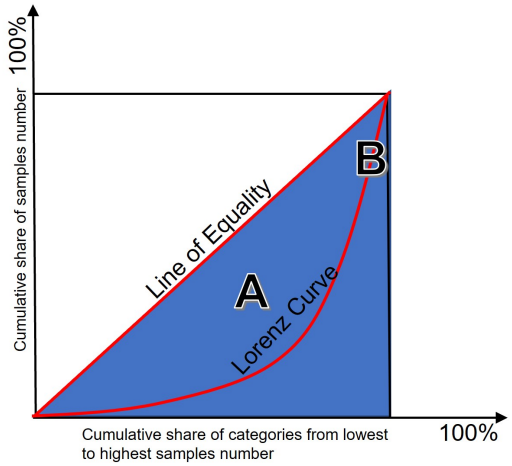

Figure 2. Calculation of the Gini coefficient (adapted from [20]).

2.2.2. Imbalance Factor

The imbalance factor refers to the number of samples in the largest class divided by the smallest:

$$IF = \frac{Max(C_1, C2, \cdots, C_k)}{Min(C_1, C2, \cdots, C_k).} \tag{4}$$

2.2.3. Standard Deviation

The standard deviation can be expressed as

$$SD = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(C_i - \mu)^2} \tag{5}$$

where μ represents the mean number of samples.

2.2.4. Mean/Median

The ratio of mean to median can be expressed as

$$MM = \frac{Mean(C_1, C2, \cdots, C_k)}{Median(C_1, C2, \cdots, C_k).} \tag{6}$$

2.3. Object Detection Network

The goal of object detection in camera trap datasets is to determine what and where animals are in images. Transformer architecture, which relies solely on the attention mechanism, and eliminates the convolution operator, has significantly impacted deep learning, particularly computer vision [45,46]. Transformer can provide the overall perception of one image, while the convolution network has limited receptive fields, and lacks a global understanding of the image. Researchers have introduced Transformer into the field of computer vision, and some pioneering work, such as ViT (Vision Transformer) and its follow-ups, has achieved better performance than CNNs in classification tasks [47–50]. The DETR (DEtection TRansformer) model introduced Transformer into the object detection task without using hand-designed components, such as anchor design and Non-Maximum Suppression (NMS), and many follow-up methods, such as Deformable DETR, DN-DETR (DeNoising-DETR), DAB-DETR (Dynamic Anchor Box DETR), etc., have attempted to

address slow convergence and limited performance in the detection of small objects [51–54]. Based on the DAB-DTER and DN-DETR, Zhang et al. [55] proposed a DETR-like, end-to-end object detector, DINO (**D**ETR with **I**mproved de**N**oising anch**O**r boxes). DINO improved over previous DETR-like models in performance and efficiency, by using a contrastive learning method for denosing training, to stabilize bipartite matching during training, the mixed query selection method for anchor initialization, and the look forward twice scheme for box prediction, as shown in Figure 3. For the first time, DINO, as an end-to-end Transformer detector, achieved the best results against both COCO val2017 (63.2AP) and test-dev (63.3AP) benchmarks; therefore, in this study, we chose DINO as our baseline applied to the camera trap datasets object detection.

In long-tailed camera trap datasets, the scarcity of images of rare species can result in insufficient features learning by the object detection model: this often leads to errors or omissions in the detection of rare species, resulting in poor overall detection performance. In nature, some species—especially those from the same Class, Family or Genus—tend to have similar characteristics. Transferring shared knowledge from frequent or common species to rare species can enhance the features of rare species, helping to overcome the disadvantages of the scarcity of samples of rare species. In addition, because images obtained from the different static cameras have different backgrounds, invariant features between images of the same species with different backgrounds are crucial for robust representation learning with limited samples. Thus, the diverse, firm, and complex sample relationships in camera trap datasets can be used to address the imbalance problem. Hou et al. [27] proposed one simple yet effective module, referred to as BatchFormerV2 (BF), to enable deep neural networks with the ability to learn the sample relationships from each mini-batch. Traditionally, Transformer block is applied to pixel/patch-level feature maps, while the BF transformer block is applied to feature maps where the length of sequence is batch-sized. An overview of the deep learning network with BF is shown in Figure 4. By integrating the BF module, the vision transformers network forms a two-stream pipeline that shares the same training weights. The outputs of the two streams are fed into the same transformer decoder. With the two-stream design, all shared blocks are trained with shared weights during training, and the original blocks can work well without BF, to avoid any extra inference load during testing. Hou et al. integrated the BF into different vision transformer models, such as DETR and Deformable-DETR, and consistently and significantly improved them by over 1.3% on the MSCOCO benchmarks, while this was not achieved for any of the long-tailed camera trap datasets.

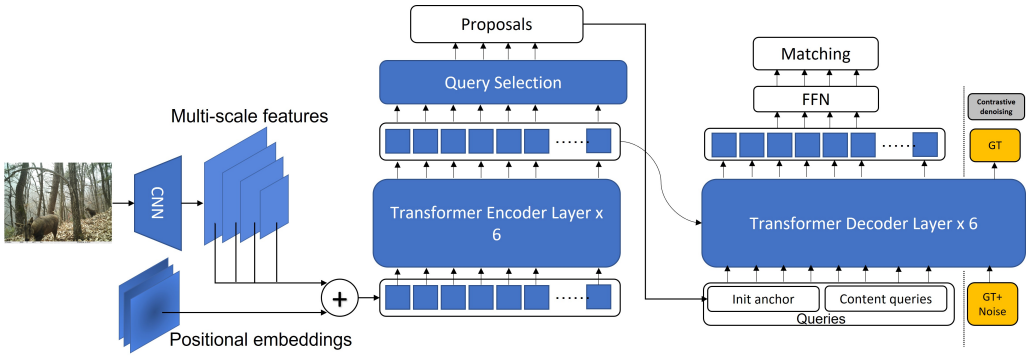

Figure 3. Overview of DINO network (adapted from [55]).

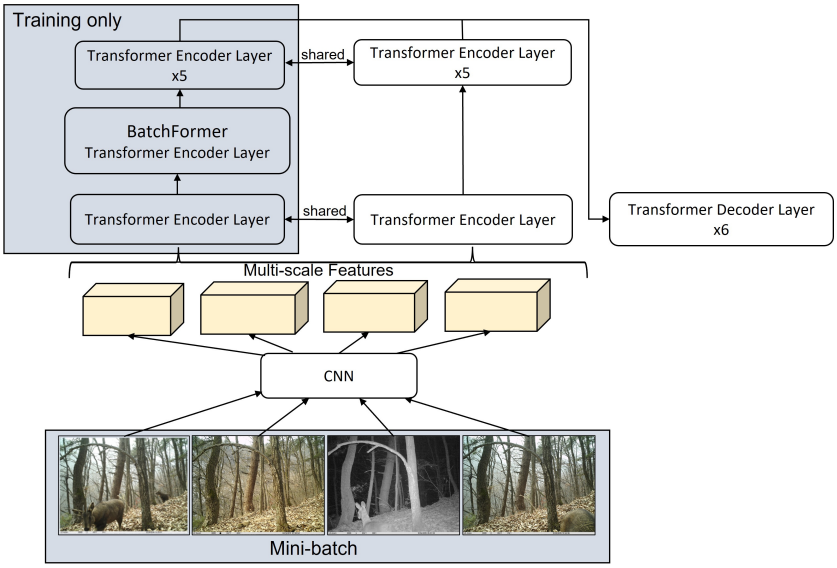

Figure 4. Overview of deep learning network with BatchFormer.

In this study, we evaluated the performance of the original DINO model and a modified version that integrated the BatchFormer module with multiple real-world camera trap datasets (SWG, WCS, and CBL), to demonstrate the efficacy of learning sample relationships for long-tailed object detection. We seamlessly inserted the BF module after the first encoder layer of DINO, and did not change the structure and parameters of the DINO framework. The original DINO was composed of a ResNet-50 backbone, a 6-layer Transformer encoder, a 6-layer Transformer decoder, and multiple prediction heads. The training of the models was based on PyTorch. Experiments were run on NVIDIA TITAN RTX GPU with 24G VRAM size. We trained 12 epochs with a batch size of 2. We set the initial learning rate to 1×10^{-4}, and adopted a simple learning rate scheduler; we also used the AdamW optimizer with a weight decay of 1×10^{-4}.

3. Results

3.1. Camera Trap Datasets

Before analyzing the data, we directly filtered and deleted images with labels such as 'end', 'start', 'blurred', 'car', etc., that were not related to animals. Due to poor image quality caused by lighting, motion blur, distance, etc., experts could not accurately judge the species or even distinguish between animals and background, so the images were labeled with non-deterministic tags, such as 'unknown small mammal', 'unidentified bird', 'unknown', etc. We also directly filtered and deleted. Some low-quality images are shown in Figure 5. To protect people's privacy, we also removed images labeled 'human'.

Figure 5. Examples of low-quality images.

Details of 12 datasets are shown in Table 1. As shown in this table, the number of images labeled 'unrelated' and 'uncertain' was only a small part of the whole, and filtering out this part did not change the overall distribution of the data. The samples labeled as 'empty' in the 12 datasets accounted for the largest proportion, reaching up to 92.17%. Accurately and effectively distinguishing between blank images and images containing objects is a very important task in the field of deep learning of the camera trap dataset. Some images from the SWG and WCS datasets contained bounding box annotations. Finally, we selected SWG, WCS, and self-constructed CBL datasets as the experimental datasets for object detection.

Table 1. The details of 12 camera trap datasets.

Dataset	No. of Species	No. of Total	No. of Filtered Species	No. of Filtered Samples	No. of Blank	Percent of Blank
Orinoquía	57	112,267	18	15,157	20,334	21%
SWG	120	2,039,657	28	885,445	264,755	23%
Island	48	142,341	6	7302	77,670	23%
Karoo	37	142,341	6	363	31,792	58%
Kgalagadi	30	10,402	2	459	7886	79%
Enonkishu	38	30,542	2	1345	19,048	65%
Camdeboo	42	30,717	3	337	13,363	44%
Zebra	53	73,606	4	797	67,115	92.17%
Kruger	45	10,637	2	610	6532	65.14%
Serengeti	60	7,261,545	4	82,384	5,445,842	75.86%
WCS	675	1,369,991	37	203,420	591,874	51%
CBL	70	48,078	0	0	0	0

According to the ratio of 8:2, the three object detection datasets were randomly divided into the training set and test set, as shown in Table 2.

Table 2. The three object detection datasets: training set and test set assignments.

Dataset	Training Set	Test Set
SWG	63,722	15,791
WCS	275,514	68,804
CBL	38,391	9687

3.2. The Long-Tailedness Metrics Results

Based on four commonly used metrics, we quantified the long-tailedness of 12 camera trap datasets and other public benchmark datasets, as shown in Table 3. In the table, we listed the values of the largest and smallest number of samples of species, named 'Max size' and 'Min size'.

To better measure the long-tailedness of camera trap datasets, we calculated the four metrics of the public balanced CIFAR, long-tailed ImageNet-LT, and LVIS 1.0. The sample size of each category of the CIFAR was equal, the Gini Coefficient and Standard Devian were 0.0, and the Imbalance factor and Mean/Median were 1.0. According to the calculation results, we can conclude that the two metrics, imbalance factor and standard deviation, are easily affected by the extreme species, and cannot reflect the overall long-tailed data distribution. The ratio of mean to median can reflect the imbalance distribution of data, including the difference in the number of frequent and rare classes; however, the size of this indicator has no upper limit, and is susceptible to extreme absolute numbers of samples of classes. Unlike the other indicators, the Gini Coefficient has a bounded distribution, is not affected by the extreme differences in the number of samples, and can represent the overall class imbalance of data and the differing long-tailedness of different datasets. According to the result of CIFAR, ImageNet-LT, and LVIS 1.0, and similar to the work of Lu et al. [20], the

Gini Coefficient is the recommended, reasonable, effective, and quantitative indicator by which to reflect the long-tailedness of the camera trap dataset. Compared to ImageNet-LT and LVIS 1.0, the imbalance in the public benchmark dataset varies widely, while the Gini Coefficient of the camera trap datasets is generally very large, reaching over 0.9: this indicates that the sample size of various species in camera trap datasets is generally and extremely imbalanced.

Table 3. The metrics results of 12 camera traps datasets, CIFAR, ImageNet-LT, and LVIS 1.0.

Dataset	GC	IF	SD	MM	Max Size	Min Size
Orinoquía	0.824	24,784	4496.7	10.64	24,784	1
SWG	0.875	234,736	30,178.71	42.68	234,736	1
Island	0.817	16,338	3076.35	9.2	16,338	1
Karoo	0.8	1698	363.33	14.68	1698	1
Kgalagadi	0.847	1378	253.77	7.35	1378	1
Enonkishu	0.761	1857	498.75	9.24	1857	1
Camdeboo	0.741	1167.67	744.8	9.49	3503	3
Zebra	0.773	1895	284.48	5.47	1895	1
Kruger	0.792	1379	220.93	8.94	1379	1
Serengeti	0.864	533,478	92,640.68	21.24	533,478	1
WCS	0.905	95,788	4598.39	33.36	95,788	1
CBL	0.872	19,728	2481.93	19.08	19,728	1
CIFAR	0.0	1.0	0.0	1.0	6000	6000
ImageNet-LT	0.524	256	139	1.58	1280	5
LVIS 1.0	0.82	50,552	2789	11.1	50,552	1

Note: GC is the Gini Coefficient; IF is the Imbalance Factor; SD is the Standard Deviation; MM is the Mean/Median. The values of the largest and smallest number of samples of species were named 'Max size' and 'Min size'.

The Gini coefficients of several object detection datasets with bounding box annotations are shown in Table 4. The Gini coefficients of three of the object detection datasets exceeded the Gini Coefficient of the LVIS 1.0. The bounding box annotations of the camera trap datasets also demonstrated an extreme class imbalance.

Table 4. The Gini coefficient of three object detection datasets, COCO, and LVIS 1.0.

Dataset	Gini Coef.	Gini Coef. of Area
SWG	0.857	0.614
WCS	0.92	0.521
CBL	0.872	0.621
COCO	0.564	0.361
LVIS 1.0	0.82	0.475

Note: GC of Area is the Gini Coefficient of object/box-level scale imbalance.

According to the definition of small objects in the current benchmark datasets, MS COCO defines objects less than 32 × 32 pixels as small; LVIS defines less than 32 × 32 as small; between 32 × 32–96 × 96 are defined as medium; greater than 96 × 96 are defined as large; TinyPerson defines objects between 20 × 20–32 × 32 as small objects, and 2 × 2–20 × 20 are defined as tiny. We divided the size of the objects of the camera trap datasets into eight ranges, which were 0–16 × 16 (Very Small, VS), 16 × 16–32 × 32 (Small, S), 32 × 32–64 × 64 (Small Medium, SM), 64 × 64–128 × 128 (Medium, M), 128 × 128–256 × 256 (Medium Large), 256 × 256–512 × 512 (Large), 512 × 512–1024 × 1024 (Very Large, VL), and >1024 × 1024 (Super Large, SL), to study the object/box-level scale imbalance of the camera trap datasets. The Gini Coefficients of the object/box-level scale imbalance of the three camera trap datasets were all greater than COCO and LVIS 1.0, so the object/box-level scale imbalance was also extremely long-tailed. Moreover, according to the areas of the objects shown in Table 5, we can conclude that there is a huge difference between camera trap and benchmark datasets.

Table 5. The number of various sizes of objects of three object detection datasets, COCO, and LVIS 1.0.

Dataset	VS	S	SM	M	ML	L	VL	SL
SWG	24	6	62	1125	8180	23,710	23,305	7310
WCS	490	2175	9594	30,657	72,885	91,432	56,302	11,979
CBL	0	0	6	331	3991	12,897	14,741	6425
COCO	107,520	160,177	192,171	178,179	137,132	77,814	6468	0
LVIS 1.0	364,697	301,748	272,726	185,982	100,047	41,770	3171	0

3.3. Object Detection

We conducted several experiments on three datasets, using DINO and DINO with BF. As shown in Table 6, the BF module significantly improved the DINO detection performance by 2.3% for SWG, by 0.8% for WCS, and by 1.2% for CBL.

Table 6. The overall results for DINO and DINO with BatchFormer Models on SWG, WCS, and CBL.

Dataset	Model	AP	AP_{50}	AP_{75}
SWG	DINO	64.8	87.5	70.6
	+BF	**67.1**	**88.6**	**71.3**
WCS	DINO	70.3	86.7	76.2
	+BF	**71.1**	**87.4**	**77.8**
CBL	DINO	72.3	94.9	78.2
	+BF	**73.5**	**94.9**	**78.5**

Note: AP is the average precision; AP_{50} is the average precision calculated when IoU is 0.5; AP_{75} is the average precision calculated when Iou is 0.75. Bold numbers are used to highlight better results.

According to the study of Scineider et al. [14], we classified the species with sample sizes of less than 500 as Rare, those between 500 and 1000 as Common, and those greater than 1000 as Frequent. As illustrated in Table 7, the BF module can also improve the DINO performance in the following categories: Rare, Common, and Frequent.

As shown in Table 8, BF can also improve the performance of DINO in terms of object detection of different sizes, but it still cannot make up for the shortcomings of too-few and too-small objects.

Table 7. The result for DINO and DINO with BatchFormer on different frequency classes of three datasets.

Dataset	Model	AP_r	AP_c	AP_f
SWG	DINO	62.5	67.6	70.1
	DINO+BF	**65.4**	**69.2**	**70.9**
WCS	DINO	69	77	76.7
	DINO+BF	**69.6**	**79**	**78.4**
CBL	DINO	70.8	72.9	78.9
	DINO+BF	**72.1**	**75.3**	**79.5**

Note: Bold numbers are used to highlight better results.

Table 8. The results for DINO and DINO with BatchFormer on various-sized objects of three datasets.

Dataset	Model	AP_{VS}	AP_{S}	AP_{SM}	AP_{M}	AP_{ML}	AP_{L}	AP_{VL}	AP_{SL}
SWG	DINO	0	0	7.2	16.5	**43.5**	57.3	78.7	84.2
	DINO+BF	0	0	**9.2**	**17.4**	42.3	**60**	**80**	**86.7**
WCS	DINO	**1.6**	20.4	35.8	51.8	66.6	79.9	86.7	92.2
	DINO+BF	1.2	**22**	**38.8**	**55.1**	**67.8**	**81.1**	**88**	**93.6**
CBL	DINO	-	-	**22.3**	42.6	57.6	70.9	80.9	87.9
	DINO+BF	-	-	21	**43.6**	**57.7**	**73.2**	**82.7**	**90**

Note: Bold numbers are used to highlight better results.

4. Discussion

The application and development of deep learning technologies have revolutionized the analysis and utilization of camera trap data. However, the imbalanced distribution of data in these datasets can often lead to the poor performance of deep learning models. In this study, we analyzed 12 camera trap datasets obtained from various habitats worldwide. These datasets exhibited significant differences in the number of species and the size of camera trap images, but the proportion of images with empty labels was the largest among all classes, reaching up to 92.17% in Snapshot Mountain Zebra. Next, we utilized four quantitative metrics to objectively and accurately quantify long-tailedness in camera trap datasets. Based on our results, we recommended the Gini Coefficient as an effective and appropriate measure of imbalance in camera trap datasets. Compared to the benchmark balanced CIFAR and long-tailed ImageNet-LT, LVIS 1.0, the class imbalance in different camera trap datasets was prevalent and very severe, consistently surpassing 0.7 in 12 datasets. Moreover, the GC of three object detection datasets was greater than COCO and LVIS 1.0, indicating that for various deep learning tasks, such as animal recognition and detection in a camera trap dataset, long-tailed distribution is a very challenging problem. In addition, the rarity of samples of some tail species in the camera trap datasets was mainly due to the fact that tail species are rare or even endangered in the wild. The ongoing global trend of anthropogenic biodiversity loss, which involves extinction or a dramatic decline in both species and population size, will further exacerbate the class imbalance in camera trap datasets; therefore, compared with the head species, which may even be over-represented, determining whether deep learning can accurately extract the information of tail species is more difficult and urgent.

Object detection accuracy varies greatly for different-sized objects. Accurate detection of small objects remains particularly challenging. Because of the different body size of animals, and different distances from the camera, the size of animal objects varies greatly. Thus, we also need to pay attention to the object/box-level scale imbalance in camera trap datasets. In this study, we calculated the GC of area to measure the object/box-level scale imbalance in three object detection datasets: the results were all greater than 0.5, demonstrating that camera trap datasets exhibit object/box-level scale long-tailed distribution as well. As shown in Table 5, camera trap datasets exhibit a positive correlation between the number of samples and object size, which is completely different from a skewness in the distribution in favor of small objects in the COCO and LVIS 1.0. However, we also note that the number of very small (0–16 $\times$ 16), small (16 $\times$ 16–32 $\times$ 32), and small–medium (32 $\times$ 32–64 $\times$ 64) objects is too few. Even worse, the camera trap images resolution is generally much higher than in the benchmark dataset, the natural background of the images is very complex, light conditions are variable, and small animals move quickly, leading to detection-performance issues for small objects, and making object detection for the camera trap dataset more challenging.

To exploit the diverse and firm sample relationships, we introduced the simple yet effective module BatchFormer into the DINO model, to transfer shared knowledge from head to tail, so as to enhance the representation of tail species. In this experiment, the BatchFormer module improved the DINO overall detection performance by 2.3% on SWG, by 0.8% on WCS, and by 1.2% on CBL. On the class imbalance, the BatchFormer module

improved the performance of DINO by up to 2.9 % on Rare, 2.4% on Common, and 1.7% on Frequent. On the object/box-level scale imbalance, the BatchFormer module can also improve the performance of DINO by up to 3.3 % on eight types of object sizes, while the AP on very small, small, and small–medium objects is too low, demonstrating that it cannot make up for the shortcoming of too few and too small objects. Exploiting sample relationships is a simple yet effective way to promote long-tailed deep learning problems for camera trap dataset solving.

Finally, due to the limitations of the experimental environments, this study did not make generalization experiments to test the performance on new camera trap data. Practically, the images from camera traps situated at new locations not included in training sets have different backgrounds (grasslands, forest, etc.), different prominent objects (tree stumps, rocks, etc.), and different environmental conditions (day, night, season, etc.), and should be considered as different domains. The deep learning models generalization performance declines in new locations [14]. In future, we will leverage the approaches of few-shot and zero-shot learning, to improve the generalization. Except for the imbalance and generalization problem, the classification and detection of nocturnal animals, such as rodents, are considerably more challenging, due to issues such as low light, fast movement, small body size, etc. Data augmentation methods such as deblur, colorization, low-light enhancement, etc., can be implemented to increase the quality of night-time images, further improving classification and detection accuracy [56–58].

5. Conclusions

Camera traps have become a popular method for collecting vast numbers of animal images. However, manually analyzing the resulting data can be slow, labor intensive, and tedious. Deep learning has emerged as a solution to overcome these obstacles. However, the long-tailed distribution of camera trap datasets becomes a barrier to taking advantage of deep learning. Our paper used four metrics to quantify the long-tailedness of 12 camera trap datasets obtained from various habitats worldwide. The results showed that long-tailedness in camera trap datasets is prevalent and very severe. Then, we analyzed the object/box-level scale imbalance for the first time, and found that object/box-level scale imbalance is long-tailed and poorer than the benchmark long-tailed dataset. To make matters worse, the number of small objects is low, making deep learning more challenging in the camera trap dataset. We employed the BatchFormer module to leverage sample relationships and enhance the performance of a general object detection model, in terms of class imbalance and object/box-level scale imbalance. In summary, the severe issue of imbalance in camera trap datasets is widely prevalent. Nevertheless, the development of deep learning techniques provides promising solutions for addressing this challenge. With the aid of deep learning techniques, camera trap data can foster biodiversity conservation.

Author Contributions: W.H. and Z.L. collected the data and calculated the results of the long-tailed metrics; W.H. and Z.L. designed the deep learning model and performed the experiments; X.T., X.H. and Z.S. constructed the CBL dataset; W.H. wrote the paper; C.C. reviewed the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19020104), and in part by the Special Project on Network Security and Informatization, CAS (CAS-WX2022GC-0106).

Institutional Review Board Statement: Not applicable. No ethical approval was required, as camera trapping is a noninvasive method.

Informed Consent Statement: Not applicable.

Data Availability Statement: CBL dataset link: http://cbl.elab.cnic.cn.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carl, C.; Schönfeld, F.; Profft, I.; Klamm, A.; Landgraf, D. Automated detection of European wild mammal species in camera trap images with an existing and pre-trained computer vision model. *Eur. J. Wildl. Res.* **2020**, *66*, 62. [CrossRef]
2. Rowcliffe, J.M.; Carbone, C. Surveys using camera traps: Are we looking to a brighter future? *Anim. Conserv.* **2008**, *11*, 185–186. [CrossRef]
3. O'connell, A.F.; Nichols, J.D.; Karanth, K.U. *Camera Traps in Animal Ecology*; Springer: New York, NY, USA, 2011.
4. McCallum, J. Changing use of camera traps in mammalian field research: Habitats, taxa and study types. *Mammal Rev.* **2013**, *43*,196–206. [CrossRef]
5. Newey, S.; Davidson, P.; Nazir, S.; Fairhurst, G.; Verdicchio, F.; Irvine, R.J.; van der Wal, R. Limitations of recreational camera traps for wildlife management and conservation research: A practitioner's perspective. *Ambio* **2015**, *44*, 624–635. [CrossRef]
6. Rovero, F.; Zimmermann, F.; Berzi, D.; Meek, P.D. "Which camera trap type and how many do I need?" A review of camera features and study designs for a range of wildlife research applications. *Hystrix-Ital. J. Mammal.* **2013**, *24*, 148–156.
7. Steenweg, R.; Hebblewhite, M.; Kays, R.W.; Ahumada, J.A.; Fisher, J.T.; Burton, C.; Burton, C.; Townsend, S.; Carbone, C.; Rowcliffe, J.M.; et al. Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Front. Ecol. Environ.* **2017**, *15*, 26–34. [CrossRef]
8. Tuia, D.; Kellenberger, B.; Beery, S.; Costelloe, B.R.; Zuffi, S.; Risse, B.; Mathis, A.; Mathis, M.W.; Langevelde, F.V.; Burghardt, T.; et al. Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **2021**, *13*, 792. [CrossRef]
9. Norouzzadeh, M.S.; Nguyen, A.M.; 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* **2017**, *115*, E5716–E5725. [CrossRef]
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
11. Banupriya, N.; Saranya, S.; Swaminathan, R.; Harikumar, S.; Palanisamy, S. Animal detection using deep learning algorithm. *J. Crit. Rev.* **2020**, *7*, 434–439.
12. Miao, Z.; Gaynor, K.M.; Wang, J.; Liu, Z.; Muellerklein, O.C.; Norouzzadeh, M.S.; McInturff, A.; Bowie, R.C.; Nathan, R.; Yu, S.X.; et al. Insights and approaches using deep learning to classify wildlife. *Sci. Rep.* **2019**, *9*, 8137. [CrossRef] [PubMed]
13. Tabak, M.A.; Norouzzadeh, M.S.; Wolfson, D.W.; Sweeney, S.J.; Vercauteren, K.C.; Snow, N.P.; Halseth, J.M.; Di Salvo, P.A.; Lewis, J.S.; White, M.; et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods Ecol. Evol.* **2019**, *10*, 585–590. [CrossRef]
14. Schneider, S.; Greenberg, S.; Taylor, G.W.; Kremer, S.C. Three critical factors affecting automated image species recognition performance for camera traps. *Ecol. Evol.* **2020**, *10*, 3503–3517. [CrossRef] [PubMed]
15. Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; Feng, J. Deep Long-Tailed Learning: A Survey. *arXiv* **2021**, arXiv:2110.04596.
16. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Large-Scale Long-Tailed Recognition in an Open World. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
17. Cui, Y.; Jia, M.; Lin, T.; Song, Y.; Belongie, S.J. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9260–9269.
18. Horn, G.V.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S.J. 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; pp. 8769–8778.
19. Gupta, A.; Dollár, P.; Girshick, R.B. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5351–5359.
20. Yang, L.; Jiang, H.; Song, Q.; Guo, J. A Survey on Long-Tailed Visual Recognition. *Int. J. Comput. Vis.* **2022**, *130*, 1837–1872. [CrossRef]
21. Zhao, Z.; Zheng, P.; Xu, S.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 3212–3232. [CrossRef] [PubMed]
22. Murthy, C.B.; Hashmi, M.F.; Keskar, A.G. EfficientLiteDet: A real-time pedestrian and vehicle detection algorithm. *Mach. Vis. Appl.* **2022**, *33*, 47. [CrossRef]
23. Li, M.-L.; Sun, G.-B.; Yu, J.-X. A Pedestrian Detection Network Model Based on Improved YOLOv5. *Entropy* **2023**, *25*, 381. [CrossRef]
24. Wang, M.; Ma, H.; Liu, S.; Yang, Z. A novel small-scale pedestrian detection method base on residual block group of CenterNet. *Comput. Stand. Interfaces* **2022**, *84*, 103702. [CrossRef]
25. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance Problems in Object Detection: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 3388–3415. [CrossRef]
26. Hou, Z.; Yu, B.; Tao, D. BatchFormer: Learning to Explore Sample Relationships for Robust Representation Learning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7246–7256.
27. Hou, Z.; Yu, B.; Wang, C.; Zhan, Y.; Tao, D. BatchFormerV2: Exploring Sample Relationships for Dense Representation Learning. *arXiv* **2022**, arXiv:2204.01254.
28. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.

29. Liu, Y.; Sun, P.; Wergeles, N.M.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
30. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
31. LILA BC:Labeled Information Library of Alexandria: Biology and Conservation. Available online: https://lila.science (accessed on 28 November 2022)
32. Vélez, J.; Castiblanco-Camacho, P.J.; Tabak, M.A.; Chalmers, C.; Fergus, P.; Fieberg, J. Choosing an Appropriate Platform and Workflow for Processing Camera Trap Data using Artificial Intelligence. *arXiv* **2022**, arXiv:2202.02283.
33. SWG Camera Traps 2018-2020. Available online: https://lila.science/datasets/swg-camera-traps (accessed on 28 November 2022).
34. Island Conservation Camera Traps. Available online: https://lila.science/datasets/island-conservation-camera-traps (accessed on 28 November 2022).
35. Snapshot Karoo. Available online: https://lila.science/datasets/snapshot-karoo (accessed on 28 November 2022).
36. Snapshot Kgalagadi. Available online: https://lila.science/datasets/snapshot-kgalagadi (accessed on 28 November 2022).
37. Snapshot Enonkishu. Available online: https://lila.science/datasets/snapshot-enonkishu (accessed on 28 November 2022).
38. Snapshot Camdeboo. Available online: https://lila.science/datasets/snapshot-camdeboo (accessed on 28 November 2022).
39. Snapshot Mountain Zebra. Available online: https://lila.science/datasets/snapshot-mountain-zebra (accessed on 28 November 2022).
40. Snapshot Kruger. Available online: https://lila.science/datasets/snapshot-kruger (accessed on 28 November 2022).
41. Swanson, A.B.; Kosmala, M.; Lintott, C.J.; Simpson, R.J.; Smith, A.; Packer, C. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci. Data* **2015**, *2*, 150026. [CrossRef] [PubMed]
42. WCS Camera Traps. Available online: https://lila.science/datasets/wcscameratraps (accessed on 28 November 2022).
43. COCO Camera Trap Format. Available online: https://github.com/Microsoft/CameraTraps/blob/main/datamanagement/README.md (accessed on 30 November 2022).
44. Gini, C. Variabilita e Mutabilita. *J. R. Stat. Soc.* **1913**, *76*, 326.
45. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2019**, *111*, 257–276. [CrossRef]
46. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:1706.03762.
47. 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.
48. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 538–547.
49. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; J'egou, H. Training data-efficient image transformers distillation through attention. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 13–18 July 2020.
50. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; J'egou, H. Going deeper with Image Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 32–42.
51. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872.
52. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13609–13617.
53. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
54. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic anchor boxes are better queries for DETR. *arXiv* **2022**, arXiv:2201.12329.
55. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605.
56. Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; Matas, J. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8183–8192.
57. Huang, S.; Jin, X.; Jiang, Q.; Liu, L. Deep learning for image colorization: Current and future prospects. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105006. [CrossRef]
58. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.T.; Cong, R. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1777–1786.