sensors

MDPI

Article

Blind Quality Assessment of Images Containing Objects of Interest

Wentong He [1,2] and Ze Luo [1,*]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; hwt0316@cnic.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: luoze@cnic.cn

Abstract: To monitor objects of interest, such as wildlife and people, image-capturing devices are used to collect a large number of images with and without objects of interest. As we are recording valuable information about the behavior and activity of objects, the quality of images containing objects of interest should be better than that of images without objects of interest, even if the former exhibits more severe distortion than the latter. However, according to current methods, quality assessments produce the opposite results. In this study, we propose an end-to-end model, named DETR-IQA (detection transformer image quality assessment), which extends the capability to perform object detection and blind image quality assessment (IQA) simultaneously by adding IQA heads comprising simple multi-layer perceptrons at the top of the DETRs (detection transformers) decoder. Using IQA heads, DETR-IQA carried out blind IQAs based on the weighted fusion of the distortion degree of the region of objects of interest and the other regions of the image; the predicted quality score of images containing objects of interest was generally greater than that of images without objects of interest. Currently, the subjective quality score of all public datasets is in accordance with the distortion of images and does not consider objects of interest. We manually extracted the images in which the five predefined classes of objects were the main contents of the largest authentic distortion dataset, KonIQ-10k, which was used as the experimental dataset. The experimental results show that with slight degradation in object detection performance and simple IQA heads, the values of PLCC and SRCC were 0.785 and 0.727, respectively, and exceeded those of some deep learning-based IQA models that are specially designed for only performing IQA. With the negligible increase in the computation and complexity of object detection and without a decrease in inference speeds, DETR-IQA can perform object detection and IQA via multi-tasking and substantially reduce the workload.

Keywords: blind image quality assessment; objects of interest; object detection; transformer; DETR; multi-task

Citation: He, W.; Luo, Z. Blind Quality Assessment of Images Containing Objects of Interest. *Sensors* **2023**, *23*, 8205. https://doi.org/10.3390/s23198205

Academic Editor: Sylvain Girard

Received: 24 July 2023
Revised: 28 August 2023
Accepted: 27 September 2023
Published: 30 September 2023

1. Introduction

In the real world, in order to monitor objects of interest, a large number of image-capturing devices, such as camera traps [1–3] for observing wild animals and video surveillance cameras for recording the activities of people, have been deployed. These automatic recording cameras produce a large number of images containing objects of interest at all times. These images can provide sufficient and diverse image data for many applications and research studies [4,5]. For example, the images from camera traps have recorded various animal behaviors without human interference, and people can use these images for animal protection and conduct various animal behavior analyses [6–9]. Thus, image quality assessment (IQA) plays an important role in image applications because the images can undergo various distortions during image acquisition, compression, transmission, etc. [9]. Since people are the final users of images, people's subjective assessment of an

image's quality is the most direct and reliable method [10]. However, the number of images substantially exceeds subjective human assessment capacities. Thus, objective image quality assessments, which enable computers to make assessments that are consistent with subjective assessments, have recently attracted increasing attention [11–13].

Generally, IQA methods can be categorized as full-reference IQA [14,15], reduced-reference IQA [16,17], and no-reference or blind IQA [18,19] via the availability of reference images. Full-reference IQA and reduced-reference IQA methods, respectively, utilize full or a part of the information from reference images to assess the distorted image's quality. Since they do not require reference images, which are usually difficult to obtain or completely unavailable in real-world applications, blind IQA methods are more challenging but more applicable [20,21].

In recent years, due to the achieved remarkable results of deep learning [22] in many computer vision tasks, many deep learning-based blind IQA models have been proposed [23–29]. Deep learning-based blind IQA methods have significantly outperformed other types of methods, such as natural scene statistic-based methods [12,30,31] and texture feature-based methods [32,33], which have currently become the dominant scheme [21]. Based on the assumption that the high-level semantic features extracted by pre-trained CNN (convolutional neural network) models for image classification tasks on large-scale datasets (such as ImageNet [34] and Places365 [35]) are quality-aware, many proposed CNN-based methods leverage these features to predict the subjective quality score, and this is carried out using mainly two methods: using a regression model such as support vector regression to regress the features onto the subjective score, or replacing the final classification layer with a regression layer and fine-tuning the regression layer with the IQA datasets to predict the image quality score [23,36–38]. The latter method results in end-to-end image quality assessments. Li et al. [9] utilized three well-known pretrained classification deep CNNs, including AlexNet [39] and ResNet [40], to extract deep semantic features and trained a linear regression model for regressing the aggregated feature onto the image quality score. Li et al. [41] utilized a network in network (NiN) model that was first pretrained on ImageNet to extract features and fine-tuned the concatenated new layers to map the learned features onto the subjective quality score. Finally, the model could directly take raw images as inputs and estimate the image quality. Very recently, end-to-end deep learning-based blind IQA methods have become the mainstream methods. Ma et al. [42] proposed a multi-task end-to-end CNN-based blind IQA method, namely MEON, which consists of two sub-networks for identifying the distortion type and predicting the quality score. Rehman et al. [43] proposed an end-to-end deep learning-based region and proposed the network (RPN) methodology for blind IQA, which used pretrained VGGNet and ResNet to extract feature maps, and RPN was used to extract region proposals from feature maps as the regions of interest (ROIs); moreover, fully connected and regression layers were used to compute the local quality score of ROI, and the average of all local ROI scores was used as the final total image quality score. For the first time, they leveraged the proposals to predict image quality and achieved standout accuracies with respect to synthetically distorted and real-world images.

It is well known that the receptive field of CNN is limited and only captures the local features of the image, thus losing non-local information and having a strong local bias. Moreover, due to the spatial invariance of shared convolution kernel weights, CNN has limited abilities in handling complex feature combinations. The attention mechanisms of the transformer [44] can aggregate global information from entire images [38]. In various computer vision tasks, such as classification [45] and object detection [46], transformer has achieved great success. In the real world, authentic images, especially the images captured from the wild, not only contain global distortions caused by low illumination, loss of focus, fog, rain, etc., but also suffer from distortions in local areas, which are caused by an object's fast movement, overexposure, etc. Thus, deep blind IQA models should accurately capture local and global features in order to fuse them and carry out image quality prediction [27]. Following the utilization of the transformer encoder in vision

transformer model [45], You et al. [47] proposed an architecture that leverages the feature maps extracted by CNN as input to the transformer encoder for image quality assessments. To address the shortcoming of image quality degradation caused by resizing and cropping images to fixed shapes in CNN-based models, Ke et al. [48] designed a multi-scale image quality transformer (MUSIQ) that encodes multi-aspect ratio multi-scale image features into a sequence of tokens as the input of the transformer encoder for image quality score prediction. Golestaneh et al. [38] sequenced the extracted features into the transformer encoder and applied relative ranking and self-consistency loss to reduce the uncertainty of the model. Yang et al. [49] first applied attention mechanisms across the channel and spatial dimensions on the extracted features via the vision transformer model to predict the quality score, and they achieved state-of-the-art performance.

In fact, the devices for monitoring objects of interest inevitably take on a large number of blank background images that do not contain objects of interest, such as animals. Currently, most deep learning-based blind IQA methods only evaluate the image's quality according to the distortion, and they do not consider the content of the objects of interest. When the distortion degree of the images containing objects of interest is greater than that of the images without objects of interest, the quality assessment predicted via the blind IQA of the former is worse than the latter, which is unreasonable or even incorrect in some real-world applications. For example, as shown in Figure 1, although the left image, which records the nocturnal behavior of animals, exhibits more severe distortion, such as dimming and overexposure, than the right image without animals, which was taken during the day, the left image is more valuable and exhibits better quality than the right image due to the presence of animals. Thus, we present a practical problem with respect to the blind IQA: how do we objectively and accurately evaluate the quality of images containing objects of interest?

Figure 1. A real-world example of an image containing animals and a blank image without animals. The left image should be more valuable and is of better quality than the right image because it captures Prionailurus bengalensis (marked in red rectangle) while exhibiting more severe distortion, such as dimming and overexposure, than the right image.

For obtaining accurate image quality assessments, humans should first focus on the image's content, especially the objects of interest. Moreover, the human visual system first pays attention to the objects of interest in an image and perceives its quality. Cao et al. [50] used an object detector to detect objects and aggregated the features of the detected objects and the image to assess image quality. The achieved state-of-the-art performance demonstrated that the objects in the images are highly relevant for IQA. Inspired by this work, the object detection network should have the potential to solve the above practical problem. In practical applications, the most direct way to blindly assess the quality of images containing objects of interest is to first filter out these images using a deep object detection model and then carry out a quality assessment on these images. Can a deep learning-based model perform both object detection and blind image quality assessment simultaneously in a multi-tasking manner (thus simplifying the process and reducing efforts, which renders it more suitable for practical application requirements)? Thus far, no research conducted has met the above requirements.

In this paper, we chose to use an advanced object detection model, DETR (detection transformer) [46], as the baseline model. Referring to the object detection heads of DETR, we added blind IQA heads on the top of the transformer decoder to extend its capability to perform object detection and blind IQA simultaneously; moreover, we named the final model DETR-IQA. The blind IQA heads consisting of simple multilayer perceptron (MLP) [51] translate the object queries from the decoder of DETR into image quality scores, and we set a hyperparameter to linearly combine the L2 loss of objects and no objects with a subjective score. We transformed the traditional blind IQA based on the distortion degree of the entire image into the blind IQA based on the weighted fusion of the distortion degree of the region of objects of interest and the other regions of the image. Finally, the proposed DETR-IQA can not only carry out accurate blind quality assessments of images containing objects of interestbut it can also carry out realistic and reasonable quality assessments relative to images without objects of interest. In addition, DETR-IQA performs object detection and blind IQA simultaneously with negligible increases in model computation and complexity, which can substantially reduce the practical workload.

The structure of this paper is as follows: In Section 2, we present the materials and methods used in this study. The details of images containing objects of interest and the results of several experiments are presented in Section 3. In Section 4, we discuss the results presented in Section 3 and the deficiencies and direction of our work. In Section 5, we present the conclusions.

2. Materials and Methods

2.1. Images Containing Objects of Interest

Several commonly used public datasets for image quality assessment can roughly be categorized as synthetic and authentic [21] datsets. Synthetic datasets, such as LIVE and TID2013 datasets, consist of reference images and distorted reference images of several types, such as JPEG2000 compression, white noise, Gaussian blur, and fast fading [52,53] distortion types in synthetic images could occur in real-world image applications, artificially synthesizing some complex distortions, such as blur, which is caused by the fast motion of objects, is difficult [54]. In this study, we worked on the blind IQA of real-world images containing objects of interest. Thus, we selected some images from the largest authentic dataset, KonIQ-10k [54], which contains 10,073 quality score images, with each image scoring a reliable 120 in terms of ratings from 1459 crowd workers on a crowdsourcing platform; these were used to construct the experimental dataset. The subjective score is in the form of a mean opinion score (MOS) ranging from 1 to 100, in which the larger MOS values present better image quality. The subjective quality score of images in KonIQ-10k is mainly used to consider 8 distortion types: noise, JPEG artifacts, aliasing, lens and motion blur, over-sharpening, wrong exposure, color fringing, and over-saturation. The crowd workers carried out image quality assessment based on the degree of the above types of distortion, and they were not instructed to consider the content of some classes of objects, such as people, birds, dogs, etc., in the images. Thus, before selecting the images, we firstly predefined some classes of objects of interest and then manually extracted the images relative to which the objects of interest were the main content component. The object of interest as the main content component can ensure that the MOS score of the image is mainly based on the object of interest. Some examples of images containing objects of interest are shown in Figure 2. The predefined class of objects of interest comprised people, birds, dogs, horses, and other animals, including sheep, animal sheer, monkey, etc. The total number of images is 3582, in which the number of images of people is 2069, bird images number 543, dog images number 456, horse images number 105, and other animal images number 541. Finally, we used a labeling tool to annotate images with a bounding box with respect to the objects of interest. We also randomly extracted 460 images without objects of interest to verify whether the proposed DETR-IQA can carry out reasonable and realistic image quality prediction: that is, it does not exhibit the problems discussed in the Introduction.

Figure 2. Example of images containing objects of interest, which are marked in rectangles and are the main content components of the image.

2.2. DETR-IQA

In real-world image applications, even if the distortion of the images containing objects of interest, such as rare animals, is more severe than that of the images without objects of interest, they are more valuable and exhibit better quality because they record the behavior and activity of the objects of interest. Therefore, people are more concerned about the distortion degree of the objects of interest in images. The human visual system first pays attention to the objects of interest in an image and perceives its quality. Since the behavior of the object of interest in the image is closely related to its surrounding environment, the image's background information can be used as auxiliary information for IQA. To mimic the above process of a human being perceiving the image's quality, the blind IQA model can be divided into two stages: object detection for extracting the features of the objects of interest and image quality assessment for predicting the quality score based on the extracted features.

Transformer, which solely relies on the attention mechanism, was first proposed in NLP tasks [44], and it achieved great performance with respect to various compute vision tasks. In the object detection task, DETR (detection transformer) [46] employs the pure transformer architecture to model object detection as a set prediction task without using hand-designed components, such as anchor design and non-maximum suppression. The multi-scale features of input images are first extracted by CNN and then fed into the encoder together with positional embedding. DETR uses a fixed number of learnable queries to probe the outputs of an encoder in a decoder and then adds a feedforward network on top of the decoder to predict either an object (object class with bounding box) or no-object (background class). DETR simplifies the detection pipeline without the need for hand-designed anchors [55] and non-maximum suppression [56], and it predicts all objects at once. The architecture of DETR is simple yet effective in the object detection task. The fixed number of decoder outputs contains the features of the objects of interest and can be used to assess the image's quality. Another advantage of DETR is that it is easy to expand because of its simple and advanced architecture. Thus, we extended the capabilities of DETR by adding IQA heads to carry out image quality assessment without degrading the object detection performance, and we named this model DETR-IQA. To the best of our knowledge, this is the first attempt to blindly assess image quality based on objects of interest. As a DETR-like model, DETR-IQA consists of four main components: a multi-scale CNN backbone, multi-layer transformer encoder–decoder, simple feedforward network (FFN) with a linear projection as detection heads, and simple multilayer perceptron (MLP) as IQA heads. In this study, the feedforward network and multilayer perceptron have the same structure, and only the dimensions of the outputs differ. The overall network is illustrated in Figure 3.

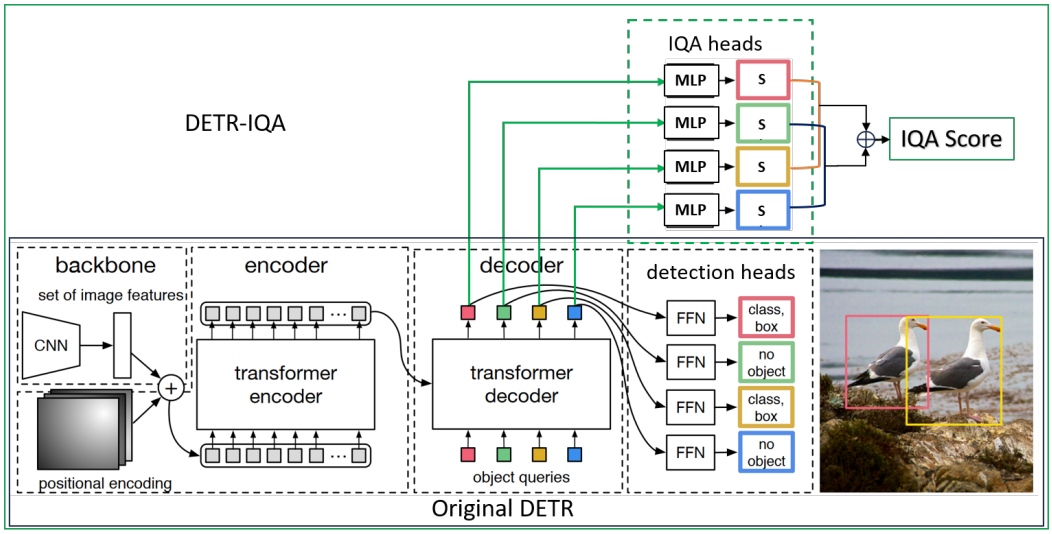

Figure 3. An overview of DETR-IQA, in which the bottom is the original DETR model and DETR-IQA is the DETR with IQA heads.

Given an input image, a conventional CNN backbone, such as ResNet-50, extracts multi-scale features. Flattened features supplemented with fixed positional encodings as a sequence are fed into the transformer encoder. The features from the encoder and N (e.g., N = 100) object queries, which are learnable positional embeddings, are fed into the transformer decoder, which contains cross-attention and self-attention modules. Then, the feedforward network regresses the object queries produced by the decoder relative to the bounding box coordinates, and a linear projection generates the classification results. Because the number of object queries is greater than the actual number of objects, the object queries not match the ground truth are denoted by the "no-object" class label. We intuitively think that the features of no objects contain some information in the background: more precisely, the surrounding information of objects that can assist the IQA. The multi-layer perceptron of IQA heads translates all object queries from the decoder into quality scores; then, all scores are averaged in some way as the image quality score. DETR-IQA is completely inherited from the original DETR and is then leveraged on the object queries of the DETR decoder to convert the IQA; this is carried out based on the distortion of the entire image relative to the IQA's object and no-object distortion. In addition, via multi-tasking, DETR-IQA can perform object detection and blind IQA simultaneously, which simplifies the practical process and reduces the practical workload.

The loss of DETR-IQ contains two parts—original loss from DETR and designed loss relative to the blind IQA head—and is defined as follows:

$$L_{DETR-IQA} = L_{DETR} + L_{IQA} \tag{1}$$

where L_{DETR} defines a linear combination of standard cross-entropy functions for classification and a combination of absolute error (L1 loss) and generalized IoU (intersection over union) for box coordinate prediction, and L_{IQA} defines a linear combination of object score loss and no-object score loss. For a more detailed description and details of the loss from DETR, we refer the reader to the DETR literature. Partly because of the blind IQA loss, we first caculated the L2 loss for all matched object regression scores with an averaged and partial MOS and all unmatched no-object regression scores with an averaged and partial MOS; finally, we added the two L2 losses as the image quality assessment's loss. Herein, the function is as follows:

$$L_{IQA} = L2(\frac{\sum_{i=1}^{O} S_i}{O}, \lambda * MOS) + L2(\frac{\sum_{j=1}^{NO} S_j}{NO}, (1-\lambda) * MOS) \tag{2}$$

where S_i and S_j, respectively, represent the score of one object and one no object, O represents the number of matched objects and NO represents the number of unmatched no objects. The sum of O and NO is equal to the fixed number of learnable quries. We intuitively considered that the MOS of an image consists of the MOS of objects and no objects, and we used λ as the hyperparameter to linearly split the MOS into two parts. We determined the value of λ based on the experimental results of object detection and image quality assessments.

2.3. Evaluation Metrics

Following prior studies, we selected Spearman's rank-order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) to evaluate the performance of DETR-IQA. The metric used to evaluate the performance of DETR-IQA relative to object detection comprises the traditional and commonly used average precision at different IoUs.

The SRCC calculates the monotonic relationship between the subjective and predicted scores and is defined as follows:

$$SRCC = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)} \tag{3}$$

where d_i represents the rank difference between the MOS and the predicted score of the i-th image, and N represents the number of test images.

The PLCC computes the linear correlation between the MOS values and predicted scores, and it is defined as follows:

$$PLCC = \frac{\sum_i (s_i - m_{s_i})(\hat{s}_i - m_{\hat{s}_i})}{\sqrt{\sum_i (s_i - m_{s_i})^2}\sqrt{\sum_i (\hat{s}_i - m_{\hat{s}_i})^2}} \tag{4}$$

where s_i and $\hat{s}_i$, respectively, represent the MOS and predicted scores of the i-th test image, and m_i and $\hat{m}_i$, respectively, represent the mean of the MOS and the predicted scores of all test images.

When values of SRCC and PLCC are closer to 1, this indicates the better performance of the blind IQA.

2.4. Implementation Details

In this study, we implemented the DETR-IQA model via PyTorch, and experiments were run on 2 NVIDIA TITAN RTX GPUs, each with 24G VRAM size. We trained 300 epochs with a batch size of 2. Moreover, we leveraged the pre-trained DETR model with the ResNet-50 backbone on COCO 2017 val5k and fine-tuned it using our constructed dataset. We did not change the structure of the baseline DETR model and used the default parameters. DETR-IQA is composed of ResNet-50, a 6-layer transformer encoder, and a 6-layer transformer decoder. We used an AdamW optimizer with a weight decay of 1×10^{-4} and at most 200 epochs. We set the initial transformer's learning rate to 1×10^{-4} and the backbone's learning rate to 1×10^{-5}. The weights of DETR-IQA were initialized with a COCO-pretrained DETR model. The hyperparameters λ were empirically set to 0.99, 0.9, 0.8, 0.7, and 0.6. According to the ratio of 8:2, the constructed images containing objects of interest were randomly divided into the training set and testing set.

3. Results

3.1. Images Containing Objects of Interest

In some real-world applications, especially monitoring objects of interest such as animals, it is common that the images containing objects of interest, such as nocturnal animals and rare animals, exhibit greater distortion than images without objects of interest. It is unreasonable or even incorrect to assess the images' quality only based on image distortion because people subjectively think that images containing objects of interest are more valuable and of better quality than images without objects of interest and, therefore,

should have greater quality scores. In order to simulate the images' data in real-world applications, we predefined five classes of objects of interest and extracted images in which the objects of interest are the main content components in KonIQ-10k. The object of interest occupying the major region of an image can ensure that the subjective quality score is mainly based on the perception of objects of interest. Then, we manually annotated the bounding box of the objects in each image. At the same time, we also randomly selected 460 images without objects of interest to evaluate the performance of the proposed model relative to images without objects of interest. The details of images containing objects of interest are shown in Table 1. The distribution histogram of the MOS of these extracted images is shown in Figure 4.

Table 1. The details of images containing five classes of objects of interest.

Class	No. of Images
Person	2069
Bird	543
Dog	456
Horse	105
Other animal	541

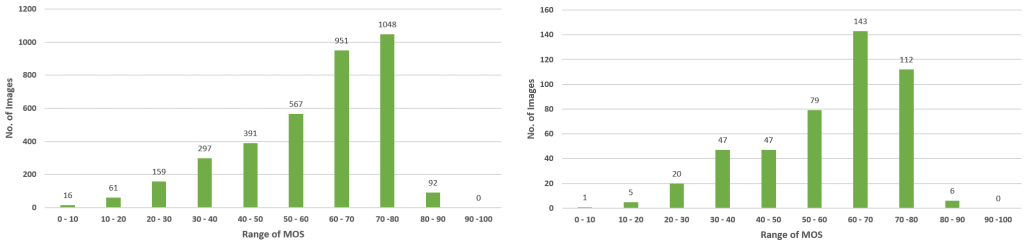

Figure 4. The distribution histogram of the MOS of images containing objects of interest (**left**) and images without objects of interest (**right**).

According to the ratio of 8:2, the images containing objects of interest were randomly divided into the training set and testing set, as shown in Table 2.

Table 2. The extracted images' training set and testing set assignments.

Training Set	Test Set
2866	716

3.2. Performance Evaluation

In this study, we added the MLP on top of the transformer decoder to extend the capability of DETR to perform object detection and blind IQA via multi-tasking. The IQA heads and detection heads shared the same backbone: the transformer encoder and transformer decoder. Because the ground-truth image quality score's MOS did not contain any information about the location of objects in the image, the IQA heads could degrade the performance of object detection. However, the IQA heads depend on the outputs of the decoder; thus, the object detection performance of DETR-IQA should not degrade too much and should preferably be close to the performance of the baseline DETR model.

We firstly conducted a DETR experiment without using IQA heads to provide the baseline object detection performance as shown in Table 3.

Table 3. The results of DETR based on the constructed dataset.

Model Name	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR	0.74	0.663	0.112	0.269	0.678

Note: AP_{50} is the average precision calculated when IoU is 0.5; AP_{75} is the average precision calculated when Iou is 0.75; AP_S is the average precision calculated when the object is small; AP_M is the average precision calculated when the object is sized in the middle; AP_L is the average precision calculated when object is large.

Then, we conducted several DETR-IQA experiments using different hyperparameters λ to find the suitable value. The results of the experiments are shown in Table 4.

Table 4. The results of several DETR-IQA experiments with different hyperparameters, λ.

λ	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	$PLCC_T$	$SRCC_T$	$PLCC_{NO}$	$SRCC_{NO}$
0.99	0.728	0.646	0.118	0.21	0.657	0.746	0.707	0.282	0.299
0.90	0.726	**0.647**	**0.224**	0.201	0.665	0.732	0.686	0.549	0.514
0.80	**0.741**	**0.647**	0.114	0.267	**0.668**	**0.785**	**0.727**	**0.659**	**0.609**
0.70	0.73	0.642	0.083	0.256	0.663	0.763	0.721	0.643	0.591
0.60	0.672	0.589	0.107	**0.299**	0.61	0.735	0.709	0.645	0.565

Note: $PLCC_T$ and $SRCC_T$ are, respectively, the Pearson and Spearman correlation coefficients relative to the test set; $PLCC_{NO}$ and $SRCC_{NO}$ are, respectively, the Pearson and Spearman correlation coefficients of the images without objects of interest.

We used AP_{50}, AP_{75}, AP_S, AP_M, and AP_L to comprehensively measure object detection performance, and PLCC and SRCC were used to measure the image quality prediction accuracy of the model relative to the test dataset (represented by T) and images without objects of interest (represented by NO). Some examples of images with ground-truth MOS and the predicted score of DETR-IQA are illustrated in Figure 5, and it can be observed that DETR-IQA can carry out reasonable and accurate image quality assessment.

Figure 5. Some examples of images with MOS and a predicted score of DETR-IQA. Upper four images contain objects of interest and lower four images do not contain objects of interest.

Finally, we calculated the FLOPs and FPS of DETR and DETR-IQA to compare their computation and complexity, and the results are shown in Table 5.

Table 5. The results of FLOPs and the FPS of DETR and DETR-IQA.

Model Name	FLOPs	Inference FPS	Params
DETR-IQA	70.09 G	41.41 M	20
DETR	70.01 G	41.27 M	20

4. Discussion

In this study, we attempted to solve the practical problem of blindly assessing the quality of images containing objects of interest in some real-world applications. Thus, we selected the largest-scale authentic dataset KonIQ-10k [54] as the base dataset. Since the crowd workers were only instructed to consider the distortion of the entire image, the subjective quality score they provided was not closely related to the objects in the images. In order to solve this problem, we first predefined five classes—person, bird, dog, horse, and other animals (such as sheep and monkey)—and manually extracted the images containing objects belonging to these classes. To ensure that the subjective score is closely related to the objects of interest, we extracted images in which the objects of interest were the main content component in order to construct the experimental dataset. As shown in Table 1, the extracted images exhibited a long-tailed distribution, in which the number of images containing a person was close to 60% and the proportion of images containing other classes of objects comprised a smaller portion; real-world datasets typically exhibit this imbalanced distribution [57]. The consistency with the distribution of the real-world dataset indicated that the constructed dataset could be used to simulate the data generated by real-world scenarios, such as camera traps used to monitor wild animals. To test whether DETR-IQA has the ability to carry out reasonable and accurate quality assessments on images without objects of interest, we also manually extracted images without the above classes of objects from the KonIQ-10k dataset. We used 10 intervals as the MOS segment, and we counted the number of extracted images in each MOS segment. As shown in Figure 4, most images in the two extracted images datasets have an MOS between 50 and 80, and the number of images accounts for more than 70%. There are very few images in the two extracted image datasets with an MOS that is lower than 20, and the number of images does not exceed 2%. The MOS distribution of both datasets is consistent with the MOS distribution of the entire KonIQ-10k dataset.

In this study, we focus on a practical problem in the field of blind image quality assessment: In real-world applications, for the purpose of monitoring certain objects, the quality of images should be better than that of images without objects of interest even if the former images exhibit more severe distortion than the latter images because the images containing objects record important information—such as the activity and behavior of objects. However, according to the current methods, the quality assessment produced opposite results. Ignoring the distortion of the image region of objects of interest may lead to inconsistency between the predicted quality and human visual perception because the human visual system is sensitive to the visual quality of objects of interest in an image. Based on the assumption that people are more concerned about the visual quality of objects of interest, we added a simple multi-layer perceptron on the top of the transformer decoder of DETR to regress the feature of objects onto the quality score. Because of the surrounding auxiliary information relative to objects, we also simply took advantage of the feature of no-objects to assist in quality prediction. The proposed end-to-end model, DETR-IQA, possesses the capability to simultaneously carry out object detection and blind image quality assessment. Thus, the performance evaluation should comprehensively consider object detection performance and image quality assessment accuracy. In DETR-IQA, the image quality assessment and object detection shared the same original DETR architecture. The quality evaluation score not only did not contain any information, such as location, that could help improve the performance of object detection, but the quality evaluation head could also increase the loss of object detection and degrade the performance results. However, DETR-IQA carried out quality assessment based on

the results of DETR. Therefore, the object detection performance was vital. We conducted baseline object detection experiments using DETR without IQA heads, and the baseline object detection results are shown in Table 3. We designed the loss of IQA heads with hyperparameter lambda to simply divide the MOS linearly in order to map the features of objects with the auxiliary features of noobjects to obtain the subjective quality score. We conducted several experiments using DETR-IQA with different λ—0.99, 0.9, 0.8, 0.7, and 0.6—and the results, including the object detection performance and image quality assessment accuracy, are shown Table 4. Comparing the five types of AP in Tables 3 and 4, the various declining trends in object detection can be observed when hyperparameter lambda is set to different values. When λ was 0.8, the performance of object detection degraded the least. At the same time, the IQA accuracy was the best relative to the test set and image set without objects of interest. This result shows that, to some extent, the performance of object detection determines the accuracy of the image quality assessment. When λ was set to 0.99, 0.9, 0.7, and 0.6, the results indicated that the feature of no objects can provide auxiliary but limited information for image quality assessments. A λ value of 0.8 also indicated that the predicted quality score of an image without objects of interest would not exceed 20 due to the absence of objects of interest. As shown in Figure 4, the number of images with an MOS below 20 did not exceed 2% in the two extracted images datasets; this is, again, relative to the entire KonIQ-10k dataset. When λ was set to 0.8, DETR-IQA guaranteed that the predicted quality score of images containing objects of interest was generally greater than that of images without objects of interest unless the former had particularly severe distortion. Moreover, according to the results of PLCC and SRCC in the last two columns of Table 3, the quality assessment scores of DETR-IQA for images without objects of interest are highly correlated with subjective scores. This indicates that DETR-IQA can carry out reasonable and accurate image quality score predictions. Some examples of images with MOS and the predicted score of DETR-IQA in Figure 5 show that DETR-IQA can produce accurate quality assessments of images containing objects of interest and reasonable quality assessments of images without objects of interest even if the MOS of images containing objects of interest is lower than that of images without objects of interest; moreover, the predicted quality scores are highly related to the distortion of images.

We investigated the performance of methods proposed in the other blind IQA research literature on the entire KonIQ-10k dataset. The existing methods adopted many advanced architectures and tricks to fit the predicted score and the ground-truth subjective score, MOS, on the KonIQ-10k dataset. Moreover, the MOS only functions in accordance with the distortion of an entire image without considering the objects of interest. The dataset constructed in this study only ensures that MOS is mainly based on the objects in the image. Even so, the performance of our proposed DETR-IQA method with respect to predicting image quality exceeded that of some deep models that are specially designed for only performing image quality assessments, such as MEON [42] and CNN [23]. Moreover, our proposed DETR-IQA has the advantage of avoiding false quality assessments in which the predicted quality score of images containing objects of interest is lower than that of images without objects of interest. The results in Table 5 show that the proposed method negligibly increased the calculation and complexity of the DETR model and did not decrease inference speeds. If applied in practical applications, the proposed DETR-IQA method can simultaneously perform object detection and IQA via multi-tasking, which can simplify the conventional process—in which object detection is first carried out, followed by image quality assessment—and greatly reduce the workload.

Finally, in this study, the accuracy of DETR-IQA in image quality assessments was constrained by the performance of object detection. Since the original DETR has drawbacks such as slow convergence and instability, we leveraged SOTA adaptive activation functions [58] to accelerate the convergence and stabilize the performance of DETR-IQA. In addition, other models with better detection performances, such as Deformable-DETR [59],DINO [60], Co-DETR [61], etc., can be used to extract the features of objects more accurately, thus

improving the accuracy of IQA. Moreover, the IQA heads in this study only used a very simple multi-layer perceptron to achieve better results than some deep learning-based models. Therefore, the design and use of more complex and advanced IQA heads, such as SwinTransformer [62], and the combination of CNN and transformer, could improve IQA accuracy.

5. Conclusions

In some practical image applications, images containing objects of interest are more valuable than images without objects of interest, even though the former exhibit greater distortion. Therefore, humans should be more concerned with the distortion of objects of interest. In this paper, the proposed model, DETR-IQA, extended DETR using IQA heads to simulate how humans perceive the quality of images containing objects of interest. The model can blindly predict the image quality score based on the features of objects detected using DETR, and the object detection performance only exhibits a slight decrease. The IQA heads consisting of very simple multi-layer perceptron architecture can not only accurately and blindly assess the quality of images containing objects of interest but also perform reasonable quality assessments on images without objects of interest. With the negligible increase in the model's computation and the complexity of object detection, DETR-IQA can perform object detection and IQA simultaneously. DETR-IQA is a simple yet meaningful effort toward solving the practical IQA problem. In the future, we will use more advanced object detection networks and design more powerful IQA heads to improve their potential in practical image applications.

Author Contributions: W.H. and Z.L. conceived the idea and analyzed the data; W.H. performed the experiments and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Special Project on Network Security and Informatization, CAS (CAS-WX2022GC-0106).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public KonIQ-10k dataset link: https://osf.io/hcsdy/.

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

References

1. Rowcliffe, J.M.; Carbone, C. Surveys using camera traps: Are we looking to a brighter future? *Anim. Conserv.* **2008**, *11*, 185–186. [CrossRef]
2. O'connell, A.F.; Nichols, J.D.; Karanth, K.U. *Camera Traps in Animal Ecology: Methods and Analyses*; Springer: New York, NY, USA, 2011.
3. McCallum, J. Changing use of camera traps in mammalian field research: Habitats, taxa and study types. *Mammal Rev.* **2013**, *43*, 196–206. [CrossRef]
4. Tang Z, Yang J, Liu, X.H.; Wang, P.Y.; Li, Z.Y.; Liu, C.S. Activity pattern of Lophophorus lhuysii by camera-trapping in Wolong National Nature Reserve, China. *Sichuan J. Zool.* **2017**, *36*, 582–587.
5. Royle, J.A.; Nichols, J.D.; Karanth, K.U.; Gopalaswamy, A.M. A hierarchical model for estimating density in camera-trap studies. *J. Appl. Ecol.* **2009**, *46*, 118–127. [CrossRef]
6. Karlin, M.; De La Paz, G. Using Camera-Trap Technology to Improve Undergraduate Education and Citizen-Science Contributions in Wildlife Research. *Southwest. Nat.* **2015**. *60*, 171–179. [CrossRef]
7. Yin, Y.F.; Drubgyal, A.; Lu, Z.; Sanderson, J. First photographs in nature of the Chinese mountain cat. *Cat News* **2007**, *47*, 6–7.
8. Huang, Z.; Qi, X.; Garber, P.A.; Jin, T.; Guo, S.; Li, S.; Li, B. The use of camera traps to identify the set of scavengers preying on the carcass of a golden snub-nosed monkey (Rhinopithecus roxellana). *Sci. Rep.* **2014**, *9*, e87318. [CrossRef]
9. Li, D.; Jiang, T.; Lin, W.; Jiang, M. Which Has Better Visual Quality: The Clear Blue Sky or a Blurry Animal? *IEEE Trans. Multimed.* **2018,** *21*, 1221–1234. [CrossRef]
10. Li, F.; Shuang, F.; Liu, Z.; Qian, X. A cost-constrained video quality satisfaction study on mobile devices. *IEEE Trans. Multimed.* **2017**, *20*, 1154–1168. [CrossRef]

11. Zhang, L.; Zhang, L.; Mou, X.Q.; Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef] [PubMed]
12. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef]
13. Cheon, M.; Yoon, S.J.; Kang, B.; Lee, J. Perceptual image quality assessment with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 433–442.
14. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
15. Chandler, D.M.; Hemami, S.S. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. Image Process.* **2007**, *16*, 2284–2298. [CrossRef] [PubMed]
16. Liu, Y.; Zhai, G.; Gu, K.; Liu, X.; Zhao, D.; Gao, W. Reduced-reference image quality assessment in free-energy principle and sparse representation. *IEEE Trans. Multimed.* **2017**, *20*, 379–391. [CrossRef]
17. Wu, J.; Liu, Y.; Li, L.; Shi, G. Attended Visual Content Degradation Based Reduced Reference Image Quality Assessment. *IEEE Access* **2018**, *6*, 2169–3536. [CrossRef]
18. Shi, Y.; Guo, W.; Niu, Y.; Zhan, J. No-reference stereoscopic image quality assessment using a multi-task cnn and registered distortion representation. *Pattern Recognit* **2020**. *100* 107168. [CrossRef]
19. Li, J.; Yan, J.; Deng, D.; Shi, W.; Deng, S. No-reference image quality assessment based on hybrid model. *Signal Image Video Process.* **2017**, *11*, 985–992. [CrossRef]
20. Cai, W.; Fan, C.; Zou, L.; Liu, Y.; Ma, Y.; Wu, M. Blind Image Quality Assessment Based on Classification Guidance and Feature Aggregation. *Electronics* **2020**, *9*, 1811. [CrossRef]
21. Yang, P.; Sturtz, J.; Qingge, L. Progress in Blind Image Quality Assessment: A Brief Review. *Mathematics* **2023**, *11*, 2766. [CrossRef]
22. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
23. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional neural networks for no-reference image quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1733–1740.
24. Lin, K.-Y.; Wang, G. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 732–741.
25. Zhang, W.; Ma, K.; Yan, J.; Deng, D.; Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *30*, 36–47. [CrossRef]
26. Bianco, S.; Celona, L.; Napoletano, P.; Schettini, R. On the use of deep learning for blind image quality assessment. *Signal Image Video Process.* **2018**, *12*, 2. [CrossRef]
27. Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; Zhang, Y. Blindly assess image quality in the wild guided by a selfadaptive hyper network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3667–3676.
28. Zhu, H.; Li, L.; Wu, J.; Dong, W.; Shi, G. Metaiqa: Deep meta-learning for no-reference image quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
29. Kim, J.; Lee, S. Fully deep blind image quality predictor. *IEEE J. Sel. Top. Signal Process.* **2016**, *11*, 206–220. [CrossRef]
30. Moorthy, A.K.; Bovik, A.C. A two-step framework for constructing blind image quality indices. *IEEE Signal Process. Lett.* **2010**, *17*, 513–516. [CrossRef]
31. Liu, L.; Dong, H.; Huang, H.; Bovik, A.C. No-reference image quality assessment in curvelet domain. *Signal Process. Image Commun.* **2014**, *29*, 494–505. [CrossRef]
32. Xue, W.; Mou, X.; Zhang, L.; Bovik, A.C.; Feng, X. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Trans. Image Process.* **2014**, *23*, 4850–4862. [CrossRef] [PubMed]
33. Freitas, P.G.; Akamine, W.Y.L.; Farias, M.C.Q. Blind image quality assessment using multiscale local binary patterns. *J. Imaging Sci. Technol.* **2017**, *29*, 7–14. [CrossRef]
34. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; FeiFei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
35. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1452–1464. [CrossRef]
36. Tang, H.X.; Joshi, N.; Kapoor, A. Blind image quality assessment using semi-supervised rectifier networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 287–2884.
37. Sun, C.R.; Li, H.Q.; Li, W.P. No-reference image quality assessment based on global and local content perception. In Proceedings of the Visual Communications and Image Processing, Chengdu, China, 27–30 November 2016; pp. 1–4.
38. Golestaneh, S.A.; Dadsetan, S.; Kitani, K.M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 1220–1230.
39. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

41. Li, Y.; Po, L.-M.; Feng, L.; Yuan, F. No-reference image quality assessment with deep convolutional neural networks. In Proceedings of the 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, China, 16–18 October 2016; pp. 685–689.
42. Ma, K.; Liu, W.; Zhang, K.; Duanmu, Z.; Wang, Z.; Zuo, W. End-to-End Blind Image Quality Assessment Using Deep Neural Networks. *IEEE Trans. Image Process.* **2018**, *27*, 1202–1213. [CrossRef] [PubMed]
43. Rehman, M.U.; Nizami, I.F.; Majid, M. DeepRPN-BIQA: Deep architectures with region proposal network for natural-scene and screen-content blind image quality assessment. *Displays* **2021**, *71*, 102101. [CrossRef]
44. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All Your Need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
45. 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* **2010**, arXiv:2010.11929.
46. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switerland, 2020.
47. You, J.; Korhonen, J. Transformer For Image Quality Assessment. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1389–1393.
48. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. MUSIQ: Multi-scale Image Quality Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 5128–5137.
49. Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; Yang, Y. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 1190–1199.
50. Cao, J.; Wu, W.; Wang, R.; Kwong, S. No-reference image quality assessment by using convolutional neural networks via object detection. *Int. J. Mach. Learn. Cyber.* **2022** *13*, 3543–3554. [CrossRef]
51. Popescu, M.; Balas, V.E.; Perescu-Popescu, L.; Mastorakis, N.E. Multilayer perceptron and neural networks. *WSEAS Trans. Circuits Syst. Arch.* **2009**. *8*, 579–588.
52. Sheikh, H.R.; Sabir, M.F.; Bovik, A.C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* **2006**, *15*, 3440–3451. [CrossRef]
53. Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Process. Image Commun.* **2015**, *30*, 57–77. [CrossRef]
54. Hosu, V.; Lin, H.; Sziranyi, T.; Saupe, D. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.* **2020**, *29*, 4041–4056. [CrossRef]
55. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007
56. Hosang, J.; Benenson, R.; Schiele, B. Learning Non-maximum Suppression. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6469–6477.
57. 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. [CrossRef]
58. Jagtap, A.D.; Karniadakis, G.E. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *J. Comput. Phys.* **2019**, *404*, 109136. [CrossRef]
59. 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.
60. 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.
61. Zong, Z.; Song, G.; Liu, Y. DETRs with Collaborative Hybrid Assignments Training. *arXiv* **2022**, arXiv:2211.12860.
62. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002.