electronics

MDPI

Article

Exploring Zero-Shot Semantic Segmentation with No Supervision Leakage

Yiqi Wang [1] and Yingjie Tian [2,3,4,5,*]

[1] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190, China; wangyiqi18@mails.ucas.edu.cn
[2] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[3] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
[4] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[5] MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation, University of Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: tyj@ucas.ac.cn

Abstract: Zero-shot semantic segmentation (ZS3), the process of classifying unseen classes without explicit training samples, poses a significant challenge. Despite notable progress made by pre-trained vision-language models, they have a problem of "supervision leakage" in the unseen classes due to their large-scale pre-trained data. For example, CLIP is trained on 400M image–text pairs that contain large label space categories. So, it is not convincing for real "zero-shot" learning in machine learning. This paper introduces SwinZS3, an innovative framework that explores the "no-supervision-leakage" zero-shot semantic segmentation with an image encoder that is not pre-trained on the seen classes. SwinZS3 integrates the strengths of both visual and semantic embeddings within a unified joint embedding space. This approach unifies a transformer-based image encoder with a language encoder. A distinguishing feature of SwinZS3 is the implementation of four specialized loss functions in the training progress: cross-entropy loss, semantic-consistency loss, regression loss, and pixel-text score loss. These functions guide the optimization process based on dense semantic prototypes derived from the language encoder, making the encoder adept at recognizing unseen classes during inference without retraining. We evaluated SwinZS3 with standard ZS3 benchmarks, including PASCAL VOC and PASCAL Context. The outcomes affirm the effectiveness of our method, marking a new milestone in "no-supervison-leakage" ZS3 task performance.

Keywords: zero-shot learning; semantic segmentation; transformer; supervision leakage

Citation: Wang, Y.; Tian, Y. Exploring Zero-Shot Semantic Segmentation with No Supervision Leakage. *Electronics* **2023**, *12*, 3452. https://doi.org/10.3390/electronics12163452

Academic Editor: Byung Cheol Song

Received: 9 July 2023
Revised: 2 August 2023
Accepted: 3 August 2023
Published: 15 August 2023

1. Introduction

Semantic segmentation is at the foundation of several high-level computer vision applications such as autonomous driving, medical imaging, and other areas involving identification and classification of objects within an image. Deep supervised learning has been instrumental in driving advancements in semantic segmentation [1–4]. However, fully supervised methods often require extensive labeled image databases with pixel-level annotations. They are typically designed to handle a pre-defined set of classes, restricting their application in diverse, real-world scenarios.

Some weakly supervised semantic segmentation (WSSS) approaches have been proposed for the above situation. These methods capitalize on easily accessible annotations like scribbles [5], bounding boxes [6], and image-level labels [7] and generate pseudo-ground-truths through visualization techniques [8,9]. However, this approach still relies on a certain degree of labeled data and needs to retrain the entire model if there are some new classes.

Humans possess an intuitive ability to recognize and classify new classes based solely on descriptive details, a powerful skill that current machine learning systems have yet to emulate fully. This observation has catalyzed the exploration of zero-shot semantic segmentation (ZS3) [10–13].

ZS3 aims to exploit the semantic relationships between image pixels and their corresponding text descriptions, predicting unseen classes through language-guided semantic information of the respective classes rather than the dense annotations. ZS3 techniques are broadly divided into generative and discriminative methods [14]. Generative ZS3 methods [15,16] usually train a semantic generator network which maps unseen class language embeddings into the visual feature space and fine-tunes the pre-trained classifier on these generated features. While these generative methods have demonstrated impressive performance, their effectiveness could be improved by a multi-stage training strategy. Discriminative methods directly learn the join embedding spaces for visual and language, like SPNet [17] and map the visual feature to the fixed semantic representations, bridging the gap between visual information and its corresponding semantic understanding. Similarly, JoEm [14] was proposed to optimize both the visual and semantic features within a joint embedding space.

However, both techniques employ local convolutional neural network (CNN) methods, which somewhat limit the global visual information utilized. They also implement an inconsistency loss for visual-language regression loss, as well as the cross-entropy segmentation ground-truth loss. These factors can potentially reduce the robustness of the models. To address this limitation, we propose an innovative strategy for ZS3. This approach eliminates the need for multi-stage training, thus directly tackling the inconsistency loss problem. We achieve this by minimizing Euclidean distances and implementing pixel-text score maps between the semantic prototypes generated by the language encoder and the visual features of the corresponding classes. This strategy obviates retraining during testing, enhancing the model's overall efficiency and flexibility. The network backbone of zero-shot networks offers another fruitful area of exploration for mitigating the bias problem. Traditional ZS3 models, as depicted in Figure 1, often suffer from a limited receptive field of CNNs and a lack of comprehensive attention mechanisms to extract global relations of visual features conditioned with language semantic information. So, we used the Swin Transformer [18] to extract the visual features on joint embedding which could offer a promising solution due to their ability to capture global feature relations and semantic information in visual features via the Multi-Head Self-Attention (MHSA) mechanism.

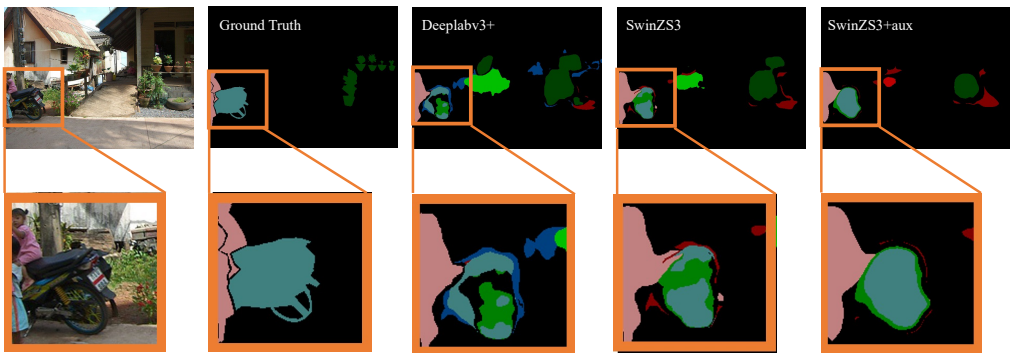

Figure 1. The impact of a transformer's global reasoning capability and the score map's decision boundary in the context of zero-shot semantic segmentation. We consider "motorbike" (represented in blue) as the unseen class. Existing solutions, such as Deeplabv3+, often produce imprecise segmentation results due to a limited receptive field and insufficient attention capabilities, resulting in a loss of fine-grained details. Implementing a transformer extractor considerably enhances the prediction accuracy of unseen classes. However, the bias towards seen classes remains, where unseen class pixels are misclassified as seen classes. To address this, our proposed SwinZS3 introduces a language-guided score map to mitigate such biases.

In this research, we combine convolutional layers with transformer blocks, enabling effective modeling of global information guided by pixel-text distances and score maps. We further refine the decision boundary by adapting the nearest neighbor (NN) classifier and introducing a score map-based weighted Euclidean distance to augment the precision of our model. Our method is validated using standard benchmarks for zero-shot semantic segmentation and shows remarkable success, surpassing the state-of-the-art performance on both PASCAL-VOC [19] and PASCAL-Context [20]. For avoiding the supervision leakage problem, we deleted all the corresponding unseen classes from PASCAL-Context and PASCAL-VOC on the pre-trained dataset ImageNet.

2. Related Work

Semantic Segmentation: Semantic segmentation has made significant progress with the advent of deep learning technologies. Chen et al. [1], Long et al. [2], Ronneberger et al. [3], and Zhao et al. [4] have leveraged deep learning architectures to enhance the performance of semantic segmentation, making it more accurate and efficient. and fully supervised semantic segmentation, operate under the assumption of pixel-level annotations throughout all training data. The DeepLab model has notably augmented segmentation performance on renowned datasets like PASCAL VOC2012 [19] and MS-COCO [21], employing sophisticated techniques such as multiple scales [13,22] and dilated convolution [23,24]. Other algorithms, such as UNet [3] and SegNet [25], have also demonstrated commendable performance using a diverse set of strategies.

Furthermore, the transformative potential of the vision transformer (ViT) [26], as the pioneer in deploying transformer architecture for recognition tasks, cannot be overstated. Concurrently, the Swin Transformer took a leap forward, extrapolating the transformer's capabilities for dense prediction tasks and achieving top-tier performance in the process. However, it must be acknowledged that these cutting-edge methods are heavily reliant on costly pixel-level segmentation labels and presuppose the presence of training data for all categories beforehand.

In the quest to circumvent these obstacles, weakly supervised semantic segmentation (WSSS) methods have emerged, leveraging more readily accessible annotations such as bounding boxes [6], scribbles [5], and image-level labels [7]. A cornerstone in prevailing WSSS pipelines is the generation of pseudo-labels, chiefly facilitated by network visualization techniques such as class activation maps (CAMs). Some works employ expanding strategies to stretch the CAM ground-truth regions to encapsulate entire objects. Still, obtaining pseudo-labels that accurately delineate entire object regions with fine-grained boundaries continues to pose a significant challenge [27,28].

Zero-shot semantic segmentation: Zero-shot semantic segmentation (ZS3) models are primarily categorized into two main types: discriminative and generative. Discerning the nuances within these two categories provides a comprehensive understanding of the current strategies utilized in the field.

Discriminative methods encompass several noteworthy studies. For instance, Zhao et al. [10] pioneered a groundbreaking study that proposed a novel strategy for predicting unseen classes using a hierarchical approach. This strategy represents an effort to build upon the data's inherent structure, using hierarchies to draw insights into unseen classes. Another study, SPNet [17], adopted a different approach by leveraging a semantic embedding space. Here, visual features are mapped onto fixed semantic representations, bridging the gap between visual information and its corresponding semantic understanding. Similarly, JoEm [14] was proposed as a method that aligns visual and semantic features within a shared embedding space, thereby fostering a direct correlation between these two aspects.

On the other hand, some studies explore the generative landscape of ZS3. ZS3Net [11], for example, employed a Generative Moment Matching Network (GMMN) to synthesize visual features. However, this model's intricate three-stage training pipeline can potentially introduce bias into the system. To mitigate this issue, CSRL [13] employed a unique strategy

that leverages relations of both seen and unseen classes to preserve these features during synthesis. Likewise, CaGNet [12] introduced a channel-wise attention mechanism in dilated convolutional layers, facilitating the extraction of visual features.

Recently, some works have explored the large-scale pre-trained model in zero-shot semantic segmentation [29–31]. Furthermore, the pre-trained data usually contain both seen and unseen labels (e.g., CLIP, WebImageText 400M) and have a supervision leakage problem. Supervision leakage is a crucial concern in machine learning, referring to the unintended incorporation of information about unseen classes during the training phase. Given that CLIP models are trained on a massive scale of approximately 400 million image–text pairs, it is conceivable that the text labels could encapsulate a diverse range of seen and unseen classes. Consequently, these models might unintentionally learn unseen classes during training, thus creating a form of supervision leakage. This situation could compromise the integrity of the zero-shot learning task, as the models are no longer genuinely learning from a "zero-shot" perspective.

In response to this significant challenge, our research introduces a distinct solution that effectively navigates the complexities of zero-shot learning while eliminating the risk of supervision leakage. By doing so, we enhance semantic segmentation models' reliability and adaptability. Our approach offers a robust framework to accurately process and categorize visual data in a genuinely zero-shot learning context. We envisage this method, free of supervision leakage, becoming a cornerstone for future research and advancements in semantic segmentation. Our work paves the way for more authentic and reliable zero-shot learning models, fostering a more resilient future for semantic segmentation in computer vision.

Visual-language learning: The domain of image–language pair learning has undergone a significant transformation marked by exponential growth. Several contributions have shaped the field, such as CLIP [32] and ALIGN [33]. Both models, pre-trained on hundreds of millions of image–language pairs, have marked substantial advancements in the field, pushing the boundaries of what is possible in image–language learning. Considering this, Yang et al. [34] put forth a unified contrastive learning method, successfully integrating both image–language techniques and image-label data. This method stands as an emblem of how these techniques can be effectively harnessed to push the frontier of the field further. Within the scope of ZS3, our study aims to build upon this method, adopting its fundamental principles and incorporating them at the pixel-level, ensuring enhanced segmentation and improved generalization capabilities. In the ever-evolving domain of zero-shot learning, CLIP-based methods [29,32,35–37] have been recognized for their substantial contributions and potential to provide effective solutions. These models capitalize on the strength of large-scale image–text pair datasets to deliver remarkable performance. However, a critical challenge that potentially undermines the legitimacy of their zero-shot learning capabilities is the risk of supervision leakage.

3. Method

3.1. Motivations

In the pursuit of improved semantic segmentation, this study seeks to enhance the performance of discriminative zero-shot semantic segmentation (ZS3) models, which fundamentally depend on the combined optimization of visual and language encoders to generate prototypes for unseen classes. This necessitates the network having a comprehensive understanding of the language context's structure. Existing networks such as JoEm [14] employ traditional convolutional layers to extract language information. However, these layers often fall short due to convolution operations' inherent locality and weak attention capabilities, failing to model the long-range and precise visual-language joint features effectively.

In response to this limitation, our study proposes using transformer-based blocks for better feature extraction. Furthermore, the study addresses another prevalent issue in ZS3: the seen bias problem. The visual encoder needs labeled data for unseen classes to extract

distinguishable features. Modulating the decision boundary could help alleviate this bias. traditional nearest neighbor (NN) classifiers' shortcomings are examined in this work, leading to the introduction of a novel decision boundary to enhance performance. With these motivations in mind, we designed our SwinZS3 framework, detailed in Figure 2.

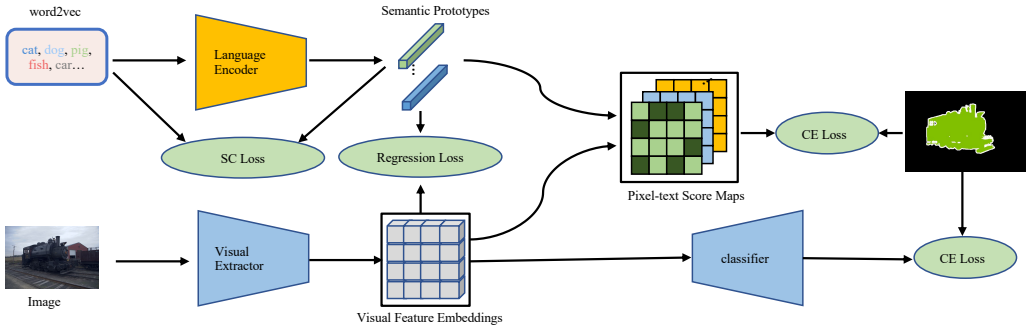

Figure 2. The overall framework of our approach SwinZS3. SSwinZS3 operates through a multi-stage process. Initially, it utilizes a transformer-based feature extractor to derive image visual embeddings and a language encoder to generate K-class semantic prototypes. These prototypes then undergo a regression loss with the visual features, and their inter-relationships are transferred from the language embeddings (word2vec) via semantic-consistency loss. Subsequently, SwinZS3 computes pixel-text score maps in a hyper-sphere space for the projected visual features and semantic prototypes. These score maps are then supervised by ground-truth labels. Simultaneously, the visual features are input into a classifier that is supervised by ground-truth labels, thereby introducing a cross-entropy loss.

3.2. Overview

In the SwinZS3 framework, we classify the dataset into seen classes (S) and unseen classes (U). During the training phase, our model, consisting of a visual feature extractor and a semantic prototype encoder, is exclusively trained on the set of seen classes (S). The primary objective of zero-shot semantic segmentation is to empower the model to identify both seen and unseen classes during testing.

Our strategy employs a visual extractor to derive visual features and a language encoder to obtain corresponding class semantic prototypes using language embeddings (specifically, word2vec). The visual features are input into a classifier then supervised by the ground-truth labels. The semantic prototypes undergo a regression loss with the visual features, and their interrelationships, established from language embeddings such as word2vec, are adjusted via semantic-consistency loss.

Following this, SwinZS3 calculates pixel-text score maps in a hyper-sphere space for the projected visual features and semantic prototypes. The ground-truth labels then supervise these score maps. Finally, the visual features are input into a classifier and supervised by ground-truth labels. The specifics of our framework are discussed in the subsequent sections.

3.3. Transformer Backbone

At the core of the SwinZS3 framework, we incorporate a transformer architecture [18], serving as a robust backbone for our model. As an initial step, the transformer takes an input image and partitions it into non-overlapping patches, each having dimensions $h \times w$. These individual patches are converted into tokens of equivalent dimensions, creating a grid-like representation of the original image.

This process is followed by introducing the Multi-Head Self-Attention (MHSA) layer into the transformer block. The MHSA layer captures global feature information from the transformed image tokens. The patch tokens are then projected into the query space $Q \in \mathbb{R}^{hw \times d_k}$, the critical space $K \in \mathbb{R}^{hw \times d_k}$, and the value space $V \in \mathbb{R}^{hw \times d_v}$, where h and w symbolize the dimensions of the feature map, and d_k and d_v represent the feature dimensionality.

The MHSA layer then calculates the outputs, X, according to the following equation:

$$X = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

Being the core operation of the transformer block, the MHSA layer fundamentally shapes the final output. Multiple such transformer blocks are stacked to generate the ultimate output of the transformer backbone.

3.4. Network Training

The SwinZS3 framework integrates four loss terms in its training process: cross-entropy loss L_{ce} [38], pixel-wise regression loss L_r, pixel-text score map loss L_{aux}, and semantic-consistency loss L_{sc}. The aggregate loss is formulated as follows:

$$L = L_{ce} + L_r + \lambda_1 L_{sc} + \lambda_2 L_{aux} \tag{2}$$

Here, λ_1 and λ_2 are employed to balance the contributions of the other losses.

Cross-entropy loss: Given the final output of feature maps $v \in \mathbb{R}^{h \times w \times c}$, where h, w, and c denote the height, width, and the number of channels, respectively, v is input into a classifier head f_c. In zero-shot settings, the classifier can learn seen classes. As such, we used cross-entropy loss [39], a popular choice in supervised semantic segmentation, on the seen classes set S as follows:

$$L_{ce} = -\frac{1}{\sum_{c \in S} \mid N_c \mid} \sum_{c \in S} \sum_{p \in N_c} log(\frac{e^{w_c v(p)}}{\sum_{j \in S} e^{w_j (v(p))}}) \tag{3}$$

where N_c denotes the label as class c in the ground truth.

Regression loss: While l_{ce} can guide the model to generate a discriminative embedding space for seen classes S, it is not adaptable for classifying unseen classes U since the classifier head needs to learn the prototypes of unseen classes. During inference, we intend to utilize the language prototypes of both seen and unseen classes as classifiers to identify the dense visual features extracted by the transformer backbone. This requires minimizing the distances between visual features and semantic prototypes. To this end, we introduced a regression loss, l_r.

Firstly, we obtain the final output visual feature maps $v \in \mathbb{R}^{h \times w \times c}$. Next, we derive the semantic feature maps $s \in \mathbb{R}^{h \times w \times d}$, where each pixel s_c of s corresponds to a language or word embedding of the same class as the corresponding visual feature pixel. Given these language embedding maps, we input them into a semantic encoder f_s as follows:

$$\mu = f_s(s) \tag{4}$$

Here, $\mu \in \mathbb{R}^{h \times w \times c}$, each pixel μ_c represents a semantic prototype for class c. The regression loss is then:

$$L_r = \frac{1}{\sum_{c \in S} \mid R_c \mid} \sum_{c \in S} \sum_{s \in R_c} d(v(s), \mu(s)) \tag{5}$$

$d()$ is the Euclidean distance metric, and R_c denotes the regions labeled with the same class in the ground truth. The l_r ensures that

The dense visual features and semantic prototypes are projected into a joint embedding space, where pixels of corresponding classes are closely aligned. However, like l_{ce}, l_r has a similar limitation in ZS3: it deals with pixel-wise visual features and semantic prototypes independently but does not explicitly consider other pixels' relationships. To address this issue, we propose using a contrastive loss.

Pixel-text score map: An integral component of our framework is the use of a pixel-text score map. This innovative inclusion helps alleviate the "seen bias" problem, a significant issue in ZS3, and reduce the discrepancy between the regression loss from semantic proto-

types and the cross-entropy loss from the ground truth. As depicted in Figure 1, the score map loss significantly enhances the smoothness of the results while concurrently reducing the noise in the semantic map. The score map functions to foster a more discriminative joint embedding space. Precisely, it is calculated using the language prototypes, denoted as $\mu_c \in \mathbb{R}^{k \times c}$, and the final output of the feature maps, represented as $v \in \mathbb{R}^{h \times w \times c}$. This calculation can be represented as follows:

$$s = \hat{v}\hat{\mu}_c^T, s \in \mathbb{R}^{h \times w \times k} \tag{6}$$

In this formulation, $\hat{\mu}_c$ and $\hat{v}$ represent the l_2 normalized versions of v and μ_c, respectively, performed along the channel dimension. Notably, the computation of the score map must strictly involve the seen class prototypes for μ_c. This restriction is essential to avoid exacerbating the unseen bias problem.

The generated score maps depict the alignment or matching outcomes between each visual feature pixel and its corresponding language-guided semantic prototype. They serve as a critical piece of the puzzle in our SwinZS3 framework. By leveraging the score maps, we calculate an auxiliary segmentation loss:

$$l_{aux} = CrossEntropy(Softmax(s/\tau), y) \tag{7}$$

In this equation, τ denotes a temperature coefficient, which we preset to 0.07, and y signifies the ground-truth label. The computation of this auxiliary segmentation loss serves a crucial purpose. It works to enhance the discriminative capacity of the joint embedding space, a characteristic that is fundamentally beneficial for zero-shot semantic segmentation. By introducing the auxiliary loss, we guide the embedding space to become more adept at differentiating between different classes. This enhancement makes the model better equipped to deal with both seen and unseen classes, thus enabling more effective and efficient semantic segmentation.

Semantic-consistency loss: In our model, we introduce the semantic-consistency loss (l_{sc}), designed to bridge the gap between the *word*2*vec* space and the embedding space of the semantic prototypes. This mechanism is essential as it capitalizes on the power of pre-trained word-embedding features, preserving vital class-contextual information within the system. Such information becomes crucial when attempting to maintain the relational significance of class prototypes.

The mathematical expression that defines the relationship between the prototypes is expressed as follows:

$$r^{\mu}ij = \frac{e^{-\tau\mu d(\mu_i,\mu_j)}}{\sum_{j \in S} e^{-\tau_{\mu} d(\mu_i,\mu_j)}} \tag{8}$$

Here, the function $d()$ represents the metric for the distance between two prototypes, and τ_{μ} acts as a temperature parameter. Similarly, we define the relationship within the word-embedding space as:

$$r_{ij} = \frac{e^{-\tau_s d(s_i,s_j)}}{\sum_{j \in S} e^{-\tau_s d(s_i,s_j)}} \tag{9}$$

The following equation then calculates the semantic-consistency loss:

$$Lsc = -\sum_{i \in S} \sum_{j \in S} r_{ij} log \frac{r^{\mu}ij}{rij} \tag{10}$$

By integrating l_{sc}, we distill contextual information from the word-embedding space and infuse it into the prototypes. This process makes them more representative, insightful, and attuned to the nuanced relationships inherent in the original data. In turn, this fosters improved segmentation performance when dealing with unseen classes.

3.5. Network Inference

In the network inference stage, we utilize semantic prototypes derived from the semantic encoder as a nearest neighbor (NN) classifier drawing from the research by [40]. We calculate the Euclidean distances and score maps between individual visual features and the language prototypes. Each visual feature is then classified according to the nearest language prototype as follows:

$$\hat{y}(p) = \underset{c \in S \cup U}{argmin}\, d(v(p), \mu_c)(1 - softmax(s)) \tag{11}$$

Here, d denotes the Euclidean distance metric, and s represents the score map. To reduce the inherent bias toward seen classes (S) in the context of unseen classes (U), we employ the strategy presented by [14] that proposed the use of the Apollonius circle. The $top2$ nearest language prototypes are defined as d_1 and d_2 for individual visual features. Here, d_1 represents the combined Euclidean and score distance to the language prototype μ_1, while d_2 stands for the distance to the language prototype μ_2. The classes of μ_1 and μ_2 are denoted by c_1 and c_2, respectively.

The decision rule is formalized with the use of the Apollonius circle as follows:

$$\hat{y}(p) = \begin{cases} c(p) & c_1 \in S \quad and \quad c_2 \in U \\ c_1 & otherwise \end{cases} \tag{12}$$

The classification is expressed as:

$$c(p) = c_1 \Pi[\frac{d_1}{d_2} \leq \gamma] + c_2 \Pi[\frac{d_1}{d_2} > \gamma] \tag{13}$$

The symbol Π represents a function that outputs a value of 1 if the argument is true and 0 otherwise. The variable γ is an adjustable parameter that can be modulated to reshape the decision boundary. This adaptability mitigates the classification bias, enhancing the semantic segmentation process's overall efficiency and precision.

4. Experiments

4.1. Implementation Details

Training: As our base model, we employ the Swin Transformer (specifically the swin-tiny variant) introduced by [27] as a proven foundation for transformer-based zero-shot semantic segmentation (ZS3) tasks. To effectively avert supervision leakage from unseen classes, an issue highlighted in the work of [41], we initialize the backbone parameters using the self-supervised model MoBY [42], which was pre-trained on the ImageNet dataset.

The optimization process is managed through an AdamW optimizer, which trains our SwinZS3 model. The initial learning rate for the backbone is set to 1×10^{-4}, with a polynomial scheduler being utilized to decay the rate at each iteration incrementally. We set the learning rate for the remaining parameters to be ten times that of the backbone parameters, thus ensuring a balance between learning efficiency and model stability. The weight decay factor is set to 0.01.

Data augmentation is a critical step in our training process; we adhere to the setting outlined in Baek et al. (2021) [14]. As for other key parameters, specifically (λ, γ), we set λ_1, λ_2 at 0.1 and γ at 0.6.

Dataset split: The experimentation was carried out on PASCAL VOC and PASCAL Context. The PASCAL-VOC2012 dataset comprises 1464 training images and 1449 validation images, spanning 21 categories (20 object categories plus one background). Conversely, the PASCAL Context dataset incorporates 4998 training and 5105 validation samples across 60 classes, including 59 distinct categories and a single background class.

In alignment with established practices, we adopted the expanded training set (10,582 samples) for PASCAL VOC. We split the Pascal-VOC2012 training samples into N-seen and

20-N unseen classes for the zero-shot semantic segmentation network. To illustrate, if we took "cow" and "motorbike" as unseen categories, removed any samples with these labels, and trained the segmentation network using the remaining samples.

During training, the segmentation model should ideally maintain a mean intersection over union ($mIOU$) of zero for unseen classes. This experimental design follows the parameters provided by ZS3Net, dividing the Pascal-VOC 2012 training samples into four different splits with an increasing number of unseen classes in each: 18-2, 16-4, 14-6, and 12-8 classes, respectively. The model is then evaluated on the full set of 1449 validation images.

The supervision leakage's methods, like Zegformer [29], zsseg [30], and ZegCLIP [31], are all based on the CLIP model, which is trained on WebImageText 400M. WebImageText 400M is collected from the website and contains a large-scale label space. Our method was pre-trained on the ImageNet datasets [43], from which we deleted all the PASCAL context and VOC consenting datasets.

Evaluation metrics: Our evaluation employs the mean intersection over union ($mIoU$) metric, in line with the methodology outlined by [2]. In particular, we computed the metrics separately for seen and unseen classes, represented as $mIoU_s$ and $mIoU_u$. Recognizing that the arithmetic mean can be heavily skewed by $mIoU_s$, we also calculated the harmonic mean ($hIoU$) of $mIoU_s$ and $mIoU_u$ to give a more balanced evaluation of the model's performance.

4.2. Ablation Study and Results

Ablation study: We aim to assess the impacts and effectiveness of our methodology through an ablation study, focusing on two particular aspects. These aspects include: (a) a comparison of convolutional neural networks (CNNs) and transformers, analyzing their relative impact on the task at hand and (b) the efficacy of utilizing the score map (l_{aux}) as a tool to adjust the decision boundary. We understand and acknowledge the critical role of cross-entropy loss and regression loss in recognizing unseen classes. Thus, we establish a baseline with Deeplabv3+, combined with l_{ce}, l_r, and l_{sc}.

In the first row of Table 1, we present the base intersection over union (IoU) scores achieved when the score map l_{aux} is not incorporated into the methodology. Following this, we compare the performance of the Swin Transformer (specifically the Swin-tiny variant) and the established Deeplabv3+ baseline. Our findings show that the transformer backbone provides a harmonic mean intersection over union ($hIoU$) gain of 1.0 when contrasted with the baseline.

Table 1. Ablation study on the unseen 6 split of PASCAL Context by comparing $mIoU$ scores using different loss terms. l_{ce}: cross-entropy loss; l_r: regression loss; l_{sc}: semantic-consistency loss; l_{aux}: pixel-text score map loss. Best numbers among the restricted sota are in bold.

Method	l_{ce}	l_r	l_{sc}	l_{aux}	$mIoU_s$	$mIoU_u$	$hIoU$
Deeplabv3+	✓		✓	✓	33.4	8.4	13.4
Deeplabv3+	✓	✓	✓		36.2	23.2	28.3
Deeplabv3+	✓	✓	✓	✓	37.7	25.0	30.2
SwinZS3	✓		✓	✓	25.8	12.0	16.4
SwinZS3	✓	✓	✓		37.1	24.3	29.3
SwinZS3	✓	✓	✓	✓	**39.3**	**26.2**	**31.4**

The benefit of introducing l_{aux} into the equation is displayed in the second and third rows of the table, as we report notable $mIoU_u$ gains of 3.0 and 3.1, respectively, over the baseline. Compared to the SwinZS3 baseline, these figures represent gains of 1.9 and 2.1, respectively. These observations underscore the significant improvements that can be realized in the zero-shot semantic segmentation (ZS3) domain, showcasing the effectiveness of our adopted methodologies.

When the transformer approach is combined with score maps, we observe the best $mIoU$ scores, solidifying the impact of this combination in performance enhancement.

The quantitative results of our research, as illustrated in Table 2, provide compelling evidence of the superior performance of our method compared to other leading approaches. These results were obtained through rigorous evaluation of well-established datasets such as PASCAL VOC and PASCAL Context. The highest performing scores across diverse split settings were reported for consistency, while the remainder of the $mIoU$ figures were referenced from [14].

Table 2. Quantitative results on the PASCAL VOC sets. The numbers in bold are the best performance. Datasets: the pre-training data for backbone; K: number of unseen classes. We highlight the improved mIoU compared to sota methods using the color green.

Datasets	K	Method	VOC		
			$mIoU_s$	$mIoU_u$	$hIoU$
Supervision leakage					
WebImageText 400M	5	Zegformer	86.4	63.6	73.3
		zsseg	83.5	72.5	77.5
		ZegCLIP	91.9	77.8	84.3
No supervision leakage					
ImageNet wo VOC	2	DeViSE	68.1	3.2	6.1
		SPNet	71.8	34.7	46.8
		ZS3Net	72.0	35.4	47.5
		CSRL	73.4	45.7	**56.3**
		JoEm	68.9	43.2	53.1
		Ours	69.2	**45.8** (+2.6)	55.3
ImageNet wo VOC	4	DeViSE	64.3	2.9	5.5
		SPNet	67.3	21.8	32.9
		ZS3Net	66.4	23.2	34.4
		CSRL	69.8	31.7	43.6
		JoEm	67.0	33.4	44.6
		Ours	68.9	**34.4** (+1.0)	**45.7** (+1.1)
ImageNet wo VOC	6	DeViSE	39.8	2.7	5.1
		SPNet	64.5	20.1	30.6
		ZS3Net	47.3	24.2	32.0
		CSRL	66.2	29.4	40.7
		JoEm	63.2	30.5	41.1
		Ours	62.6	**31.6** (+1.1)	**42.0** (+0.9)
ImageNet wo VOC	8	DeViSE	35.7	2.0	3.8 2
		SPNet	61.2	19.9	30.0
		ZS3Net	29.2	22.9	25.7
		CSRL	62.4	26.9	37.6
		JoEm	58.5	29.0	38.8
		Ours	60.2	**29.6** (+0.6)	**39.9** (+1.1)

When evaluated on the PASCAL Context dataset, as shown in the Table 3, our approach demonstrates a clear competitive edge, outstripping the second-best method, JoEm. An example is the 6 split setting, where our approach delivers a significant $mIoU_u$ improvement of 3.0 and a $hIoU$ enhancement of 3.1. This exceptional performance not only underscores the compelling capability of our method but also signals a considerable leap forward in the field of zero-shot semantic segmentation (ZS3).

The merits of our approach become even more pronounced when compared with the best generative ZS3 method, GSRL [13]. In this context, our discriminative method outshines GSRL by margins of 4.4 and 4.2 in $mIoU_u$ and $hIoU$, respectively. This comparative superiority underscores our method's efficacy, highlighting the inherent convenience of a discriminative approach. Moreover, our methodology provides a key advantage over approaches like CSRL, necessitating retraining whenever new unseen classes are introduced.

Contrarily, our framework offers a one-stage training strategy, improving computational efficiency and practical usability.

Our method consistently displays state-of-the-art performance across nearly all zero-shot settings, including unseen 2, 4, 6, and 8 splits. This consistency, notable for both $mIoU_u$ and $hIoU$ metrics, indicates our approach's robust capacity to learn discriminative representations, enhancing its generalization capability. Further validation of these results can be observed from the experiments conducted on the PASCAL VOC dataset. Our method demonstrates its competitive edge here, reinforcing our claim that this novel approach constitutes significantly to the advancement of semantic segmentation in computer vision.

Table 3. Quantitative results on the PASCAL Context validation sets. The numbers in bold are the best performance. Datasets: the pre-training data for backbone; K: number of unseen classes. We highlight the improved mIoU compared to sota methods using the color green.

Datasets	**K**	**Context**			
		Method	$mIoU_s$	$mIoU_u$	$hIoU$
Supervision leakage					
WebImageText 400M	5	Zegformer	-	-	-
		zsseg	-	-	-
		ZegCLIP	46.0	54.6	49.9
No supervision leakage					
ImageNet context	2	DeViSE	35.8	2.7	5.0
		SPNet	38.2	16.7	23.2
		ZS3Net	41.6	21.6	28.4
		CSRL	41.9	27.8	33.4
		JoEm	38.2	32.9	35.3
		Ours	39.8	**33.5** (+0.6)	**36.3** (+1.0)
ImageNet context	4	DeViSE	33.4	2.5	4.7
		SPNet	36.3	18.1	24.2
		ZS3Net	37.2	24.9	29.8
		CSRL	39.8	23.9	29.9
		JoEm	36.9	30.7	33.5
		Ours	38.7	**33.5** (+2.8)	**35.1** (+1.6)
ImageNet context	6	DeViSE	31.9	2.1	3.9
		SPNet	31.9	19.9	24.5
		ZS3Net	32.1	20.7	25.2
		CSRL	35.5	22.0	27.2
		JoEm	36.2	23.2	28.3
		Ours	**39.3** (+3.1)	**26.2** (+3.0)	**31.4** (+3.1)
ImageNet context	8	DeViSE	22.0	1.7	3.2
		SPNet	28.6	14.3	19.1
		ZS3Net	20.9	16.0	18.1
		CSRL	31.7	18.1	23.0
		JoEm	32.4	20.2	24.9
		Ours	**35.0** (+2.6)	**21.4** (+1.2)	**26.6** (+1.7)

4.3. Qualitative Results

To further evaluate the efficacy of our proposed methodology, we present several qualitative examples derived from the PASCAL VOC dataset, as shown in Figure 3. Here, we highlight the comparative superiority of the SwinZS3 model in accurately modeling unseen classes when juxtaposed with its competitor models. A key observation is that the SwinZS3 model can significantly reduce the incidence of false positive predictions. This demonstrates the model's strength in minimizing errors while discerning and categorizing unseen classes, enhancing its overall precision. By reducing false positives, the SwinZS3 model can optimize the segmentation output, leading to more accurate and reliable results.

However, as demonstrated in Figures 1 and 2, this approach may inadvertently lead to overly smoothed results at the edges. This is an area that could be further refined in future research.

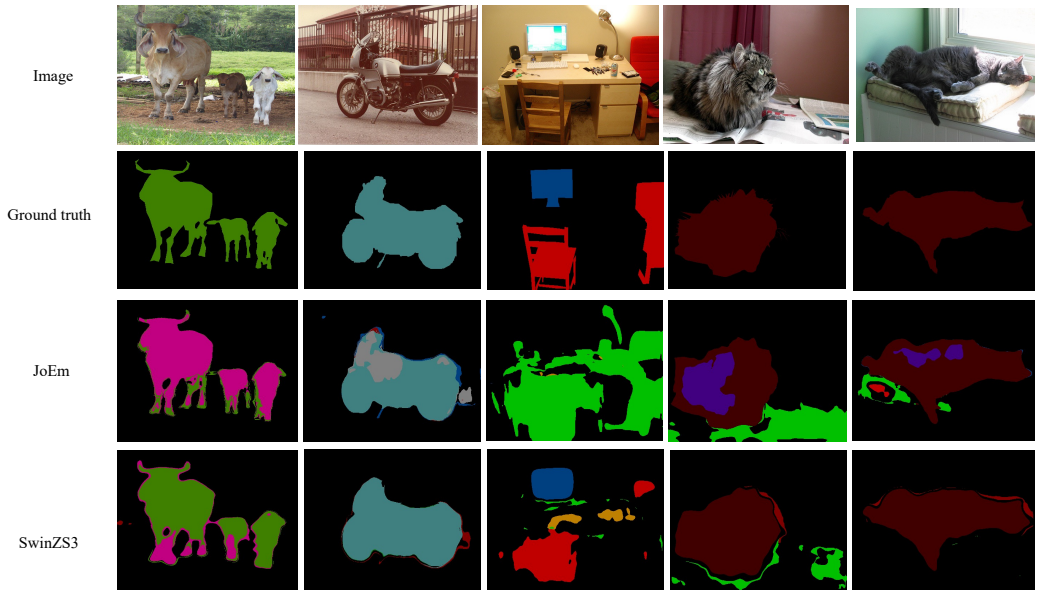

Figure 3. Qualitative results on PASCAL VOC. The unseen classes are "cow", "motorbike", and "cat". We compare the results of the other state-of-art method and our SwinZS3.

5. Conclusions

In this study, we have proposed a transformer-based framework that synergistically combines visual and language features within a unified embedding space to tackle the challenging problem of zero-shot semantic segmentation. Our approach offers a novel perspective on this task, deviating from conventional methods using a language-guided score map. This strategy enables the model to learn a more discriminative space, effectively differentiating between seen and unseen classes. Moreover, we innovatively altered the decision boundary to mitigate the prevalent seen bias issue that often undermines the performance of ZS3 models. To validate the effectiveness of our proposed methodology, we conducted extensive experiments on standard ZS3 benchmarks, and our method surpassed previous state-of-the-art performance levels. These experimental findings serve as robust evidence for the efficacy of our proposed method and underscore its promising potential for further development and application within the field of zero-shot semantic segmentation. Future research could potentially explore adapting this approach to other related tasks, such as object detection or image classification, thereby broadening its applicability and impact in computer vision. Ultimately, our study contributes to the ongoing efforts to advance the frontier of zero-shot learning, promising exciting new directions for future work.

Author Contributions: Methodology, Y.W.; Software, Y.W.; Writing—review & editing, Y.T.; Funding acquisition, Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China (No. 12071458, 71731009).

Data Availability Statement: The data presented in this study are openly available in ImageNet at https://doi.org/10.1109/CVPR.2009.5206848 (accessed on 8 July 2023) [43]; Pascal VOC at https://doi.org/10.1007/s11263-009-0275-4 (accessed on 8 July 2023) [19]; Pascal-Context dataset at https://ieeexplore.ieee.org/document/6909514 (accessed on 8 July 2023) [20].

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The authors declare no conflict of interest.

References

1. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
2. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
3. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
4. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
5. Sun, J.; Lin, D.; Dai, J.; Jia, J.; He, K.S. Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; Volume 26.
6. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
7. Hou, Q.; Jiang, P.; Wei, Y.; Cheng, M.M. Self-erasing network for integral object attention. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018; Volume 31.
8. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 818–833.
9. Zhang, D.; Zhang, H.; Tang, J.; Hua, X.S.; Sun, Q. Causal intervention for weakly-supervised semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 655–666.
10. Zhao, H.; Puig, X.; Zhou, B.; Fidler, S.; Torralba, A. Open vocabulary scene parsing. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2002–2010.
11. Bucher, M.; Vu, T.H.; Cord, M.; Pérez, P. Zero-shot semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2019**, *32* .
12. Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; Zhang, L. Context-aware feature generation for zero-shot semantic segmentation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1921–1929.
13. Li, P.; Wei, Y.; Yang, Y. Consistent structural relation learning for zero-shot segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 10317–10327.
14. Baek, D.; Oh, Y.; Ham, B. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9536–9545.
15. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
16. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
17. Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; Akata, Z. Semantic projection network for zero-and few-label semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8256–8265.
18. 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 IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canad, 11–17 October 2021; pp. 10012–10022.
19. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
20. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.G.; Lee, S.W.; Fidler, S.; Urtasun, R.; Yuille, A. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA , 23–28 June 2014; pp. 891–898.
21. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
22. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–284.
23. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
24. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
25. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
26. 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 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

27. Singh, K.K.; Lee, Y.J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Piscataway Township, NJ, USA, 2017; pp. 3544–3553.
28. Li, K.; Wu, Z.; Peng, K.C.; Ernst, J.; Fu, Y. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9215–9223.
29. Ding, J.; Xue, N.; Xia, G.S.; Dai, D. Decoupling Zero-Shot Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11583–11592.
30. Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; Bai, X. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 24–28 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 736–753.
31. Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; Liu, Y. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 11175–11185.
32. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
33. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 4904–4916.
34. Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; Gao, J. Unified Contrastive Learning in Image-Text-Label Space. *arXiv* **2022**, arXiv:2204.03610.
35. Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; Wang, X. GroupViT: Semantic Segmentation Emerges from Text Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18134–18144.
36. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6836–6846.
37. Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; Bai, X. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv* **2021**, arXiv:2112.14757.
38. Misra, I.; Maaten, L.v.d. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6707–6717.
39. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
40. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
41. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2251–2265. [CrossRef] [PubMed]
42. Xie, Z.; Lin, Y.; Yao, Z.; Zhang, Z.; Dai, Q.; Cao, Y.; Hu, H. Self-supervised learning with swin transformers. *arXiv* **2021**, arXiv:2105.04553.
43. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. 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. [CrossRef]