Expediting Contrastive Language-Image Pretraining via Self-Distilled Encoders

Bumsoo Kim*, **Jinhyung Kim, Yeonsik Jo, Seung Hwan Kim**

LG AI Research

Abstract

Recent advances in vision language pretraining (VLP) have been largely attributed to the large-scale data collected from the web. However, uncurated dataset contains weakly correlated image-text pairs, causing data inefficiency. To address the issue, knowledge distillation have been explored at the expense of extra image and text momentum encoders to generate teaching signals for misaligned image-text pairs. In this paper, our goal is to resolve the misalignment problem with an efficient distillation framework. To this end, we propose ECLIPSE: Expediting Contrastive Language-Image Pretraining with Self-distilled Encoders. ECLIPSE features a distinctive distillation architecture wherein a shared text encoder is utilized between an online image encoder and a momentum image encoder. This strategic design choice enables the distillation to operate within a unified projected space of text embedding, resulting in better performance. Based on the unified text embedding space, ECLIPSE compensates for the additional computational cost of the momentum image encoder by expediting the online image encoder. Through our extensive experiments, we validate that there is a sweet spot between expedition and distillation where the partial view from the expedited online image encoder interacts complementarily with the momentum teacher. As a result, ECLIPSE outperforms its counterparts while achieving substantial acceleration in inference speed.

Introduction

Transformers (Vaswani et al. 2017) have achieved significant progress across various challenging vision tasks such as image classification (Dosovitskiy et al. 2021; Touvron et al. 2021; Jiang et al. 2021; Graham et al. 2021), object detection (Carion et al. 2020), semantic segmentation (Xie et al. 2021; Liu et al. 2021b; Wang et al. 2021) and visual relationship detection (Kim et al. 2021, 2022). Following this success in vision tasks, recent studies demonstrated that large-scale vision-language pretraining (VLP) (Li et al. 2019; Chen et al. 2020c; Huang et al. 2019; Li et al. 2020, 2021b; Lu et al. 2019; Tan and Bansal 2019; Jia et al. 2021; Radford et al. 2021) with ViTs is scalable to large uncurated datasets and transferable to various downstream tasks.

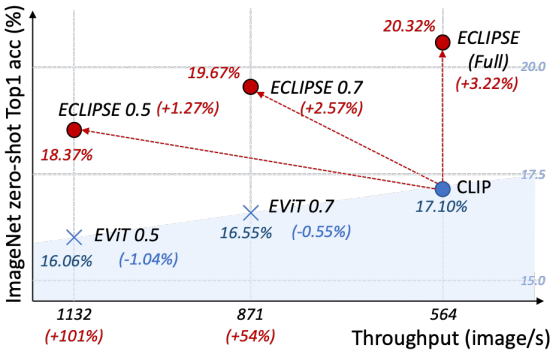

Figure 1: Time vs. ImageNet zero-shot performance analysis for Contrastive Language-Image Pretraining with existing ViT accleration framework (EViT). We compare the results between EViT directly applied on CLIP and EViT trained with our proposed meta-architecture, ECLIPSE. Our proposed framework enables even streamlined ViTs with 101% faster throughputs to outperform the full ViT of CLIP. Model performance and inference time are measured with ViT-B/16 backbone.

However, the large scale image-text pairs for VLP are usually collected from the web; thus they are often noisy, i.e., having weak correlation between the image and its corresponding text description. To alleviate the image-text misalignment problem, previous works (Li et al. 2021a; Lu et al. 2022) have proposed knowledge distillation framework (Hinton et al. 2015) with a momentum encoder for both image and text. However, adopting two additional momentum encoders and calculating soft alignments for distillation loss inevitably increase the computational cost for training, which hinders training for large-scale VLP.

In this work, we propose a efficient formulation for distilling soft image-text alignment matrix without text momentum encoder for contrastive language-image pretraining (Jia et al. 2021; Radford et al. 2021). Inspired from SimSiam (Chen and He 2021), we simply replace the text momentum encoder with stop-gradient operation. This design not only eliminates the computational cost for an additional text momentum encoder, but also enables the distillation

*correspondence to: bumsoo.kim@lgresearch.ai

to operate within a unified projected space of text embedding, resulting in better performance. Based on this shared projected space, we adopt token sparsification (Liang et al. 2022b) for the online image encoder to i) provide a partial view that complementarily interacts with the full-view of the momentum image encoder, ii) compensate for the computational overhead of training the momentum image encoder, and iii) accelerate inference speed. While our distillation architecture effectively improves data efficiency by alleviating the natural misalignment between images and text, the expedited online image encoder and the momentum teacher positively interacts with a sweet spot that achieves speed improvement without degrading performance. We name this meta-architecture as ECLIPSE: **E**xpediting **C**ontrastive **L**anguage-**I**mage **P**retraining with **S**elf-distilled **E**ncoders. ECLIPSE is trained with a loss jointly obtained from two image-text alignment matrices (i.e., $\bar{A}$ and A in Fig. 2):

- The batch-wise image-text alignment matrix $\bar{A}$ between the text encoder and the momentum teacher is trained with an InfoNCE loss (Oord, Li, and Vinyals 2018) with hard alignment labels for matching image-text pairs.
- Student-text alignment matrix A is obtained likewise with the online network and the text encoder with stop gradient. We train the online network to match A with the soft alignment matrix $\bar{A}$ obtained above.

The momentum parameters are updated with an exponential moving average (EMA) of the parameters of online encoder.

Extensive experiments demonstrate the effectiveness of ECLIPSE, showing that our distillation architecture significantly improves data efficiency while achieving substantial model acceleration. For example, when applied to CLIP (Radford et al. 2021), our proposed architecture improves 1.27% zero-shot accuracy in ImageNet classification while achieving 101% acceleration in inference speed. Moreover, ECLIPSE can be also trained without expedition, which then shows a large 3.22% gain compared to ViT, thus offers a model choice between an accelerated model with competitive performance and a full-capacity model with enhanced performance (see Fig. 1 and Tab. 3). Furthermore, scaling to large-scale datasets, ECLIPSE achieves state-of-the-art on several downstream tasks, outperforming CLIP variants with a model accelerated by more than 54%.

Related Work

Vision-Language Pretraining (VLP) learns a joint representation between two modalities on large-scale image-text pairs. VLP covers both *single*-stream models (Li et al. 2019; Chen et al. 2020c; Huang et al. 2019; Li et al. 2020, 2021b; Lu et al. 2019; Tan and Bansal 2019) and *dual*-stream models (Jia et al. 2021; Radford et al. 2021; Li et al. 2022). Single-stream models jointly encode both image and text input with a single multi-modal encoder. Though they have shown impressive performance in several image-text downstream tasks, single-stream models suffer from their large inference cost for the cross-modal retrieval. Also, how to transfer the pretrained joint encoder to the unimodal downstream tasks, e.g., image recognition, is not trivial. On the contrary, dual-stream models encode the images and texts separately with independent encoders, and thus have several advantages: simplicity, versatility, and relatively cheaper computational cost. In this work, we focuses on a dual-stream encoder trained with a contrastive objective.

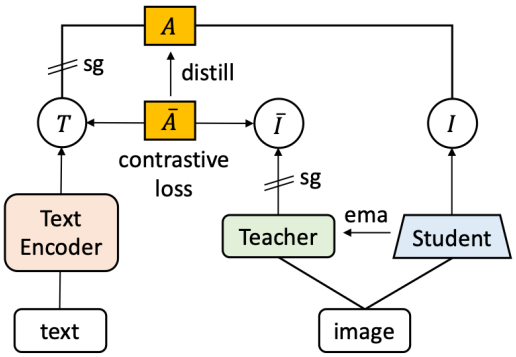

Figure 2: Overview of ECLIPSE. Student encoder is trained to estimate the soft alignment matrix $\bar{A}$ predicted by Text Encoder and the Teacher network. sg stands for stop-gradient, I and $\bar{I}$ are encoded image with student and teacher network, respectively.

Contrastive Language-Image Pretraining

In Contrastive Language-Image Pretraining (Jia et al. 2021; Li et al. 2021b; Radford et al. 2021), the model is trained via a contrastive loss with large-scale image-text pairs, where the matching image-text pairs comprise a positive pair while other arbitrary pairs are treated as negative pairs. Several works (Mu et al. 2021; Li et al. 2022) introduce additional form of supervision such as self-supervision between augmented views of the image (e.g., SimCLR (Chen et al. 2020a), SimSiam (Chen and He 2021)), language self-supervision (e.g., supervision with text augmentation (Wei and Zou 2019), masked language modeling (Devlin et al. 2019)) and momentum contrast with nearest neighbor (He et al. 2020) to further improve downstream performance. Recently, FLIP (Li et al. 2023) borrowed random masking strategy (He et al. 2022) for the input token which substantially improves the training efficiency. However, FLIP employs unmasked images during inference, which means that there is no speed improvement at inference time. Furthermore, due to the discrepancy between the training and test distributions, an additional unmasked tuning process is necessary. On the other hand, ECLIPSE improves both training and inference speed without extra tuning strategy.

Distillation and Momentum Contrast for VLP

Knowledge distillation (Hinton et al. 2015) has been initially proposed to transfer the knowledge of a large model (the teacher) to a smaller model (the student). Consecutive works (Romero et al. 2015; Park et al. 2019) have been explored different distilling targets other than direct output. Extending the concept, distillation from an identically structured model (Furlanello et al. 2018; Hessam Bagherinezhad and Farhadi 2018) or a momentum network (teacher) (Tar-

vainen and Valpola 2017; He et al. 2020; Grill et al. 2020; Caron et al. 2021), whose parameters are updated with the exponential moving average (EMA) of a online network (student), have been proposed. Momentum contrast (MoCo (He et al. 2020)) is the pioneering contrastive learning method for images without labels that uses momentum encoder and memory queue to increase the number of negative samples. Inspired from MoCo, HIT (Liu et al. 2021a) adopted momentum encoders and memory bank for video-text contrastive matching without distillation. ALBEF (Li et al. 2021a) and COTS (Lu et al. 2022) distill soft-alignment matrix obtained from both image and text momentum encoders to the online encoders. Andonian et al. (Andonian, Chen, and Hamid 2022) proposed self-distillation via swapping image-text alignment matrix without momentum encoder. MCD (Kim et al. 2023) proposes a distillation where the misalignments caused by image augmentation serves as a training signal. We introduce a novel effective distillation method called ECLIPSE whose online and momentum image encoder share text encoder. Through a systematic analysis, we validate diverse distillation designs (Table 2) and demonstrate effectiveness of ECLIPSE.

Method

In this section, we propose ECLIPSE: Expediting Contrastive Language-Image Pretraining with Self-distilled Encoders. Our goal is to resolve image-text misalignment problem of Contrastive Language-Image Pretraining (i.e., CLIP) for uncurated image-text pairs via efficient distillation formulation without extra text momentum encoder. We further compensate the heavy computational cost of distillation by adopting model expediting (i.e., EViT) to the online encoder that requires gradient computation. We start from revisiting basic concepts of CLIP and EViT (Liang et al. 2022b). Then, we introduce our meta-architecture ECLIPSE that combines CLIP with our novel knowledge distillation structure and ViT acceleration for efficient training and inference.

Contrastive Language-Image Pre-training

First, we revisit basic form of contrastive language–image pretraining (Radford et al. 2021). CLIP features a dual-encoder architecture where the image encoder f_I and text encoder f_T are jointly trained with contrastive objective $\mathcal{L}_C$.

Image-Text Alignment Matrix. For convenience, we denote $A \in \mathbb{R}^{N\times N}$ as the image-text alignment matrix for a given batch of N image-text pairs $\{(x_i^I, x_i^T)\}_{i=1}^N$. Each element of the image-text alignment matrix A_{ij} is the cosine similarity between the projected representations of the i-th text and j-th image (i.e., $T_i = f_T(x_i^T)$ and $I_j = f_I(x_j^I)$, respectively), written as:

$$A_{ij} = \text{sim}(T_i, I_j), \tag{1}$$

where $\text{sim}(\cdot)$ is cosine similarity.

InfoNCE Loss. In CLIP, the encoded image features I and text features T are projected to the same dimension where the embeddings for matching image-text pairs are pulled together while embeddings for non-matched pairs are pushed apart with the InfoNCE loss (Oord, Li, and Vinyals 2018). Using Eq. (1), the InfoNCE loss $\mathcal{L}_N$ is rewritten as:

$$\mathcal{L}_N(A) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(A_{ii}/\tau\right)}{\sum_{j=1}^{N}\exp\left(A_{ij}/\tau\right)}, \tag{2}$$

where τ is a learnable temperature variable. The loss for the text encoder $\mathcal{L}_T$ and image encoder $\mathcal{L}_I$ are written as:

$$\mathcal{L}_T = \mathcal{L}_N(A), \;\; \mathcal{L}_I = \mathcal{L}_N(A^T). \tag{3}$$

The overall loss for CLIP is the average of the loss for each encoder, written as $\mathcal{L}_{\text{CLIP}}(A) = \frac{1}{2}(\mathcal{L}_T + \mathcal{L}_I)$.

Accelerating ViTs with Token Sparsification

Previous work in ViT acceleration (Rao et al. 2021; Liang et al. 2022b) mainly focused on token sparsification since the complexity of transformer attention is reduced at a quadratic scale with respect to the number of tokens that are discarded, significantly improving model throughputs. Most recent works proposed token sparsification via external models or reorganizing the patch tokens based on their attentiveness with the `[CLS]` token. In this work, we benchmark EViT (Liang et al. 2022b) since no additional parameters are introduced for acceleration. We follow their architecture design and discard a fixed ratio (1-κ) of inattentive tokens according to the attention value between the `[CLS]` token and each patch in the 4^{th}, 7^{th}, and 10^{th} transformer layers, where κ is the token keep rate.

ECLIPSE

Towards a data-efficient pretraining with uncurated image-text pairs, we propose ECLIPSE, a novel distillation pipeline that alleviates image-text misalignments. The overall architecture of ECLIPSE is illustrated in Figure 3. ECLIPSE features a text encoder, a online image encoder (EViT), and a momentum encoder (ViT). Below we provide a step-by-step description of our proposed ECLIPSE architecture.

Knowledge Distillation. Knowledge distillation, introduced by (Hinton et al. 2015), is a learning paradigm where we train the student network to mimic the "soft" labels predicted from the teacher network. Following previous intuition, we adopt a knowledge distillation framework for the token-sparsified online ViT (student) to train the output of the full ViT with momentum weights (teacher), aiming to accelerate ViT without degrading performance. However, our empirical results show that applying conventional distillation (Hinton et al. 2015; Touvron et al. 2021; Caron et al. 2021) (i.e., training the student network to directly predict the output of the teacher network) to CLIP shows minor improvement in performance when transferred to downstream tasks. Motivated by this finding, we propose a unique distillation architecture via the image-text alignment matrices denoted in Eq. (1).

Training Loss of ECLIPSE. Given the momentum encoder $\bar{f}_I$, online encoder f_I and text encoder f_T, we define a pair of alignment matrices using Eq. (1) as:

$$\bar{A}_{ij} = \text{sim}(T_i, \bar{I}_j), \;\; A_{ij} = \text{sim}(\text{sg}(T_i), I_j), \tag{4}$$

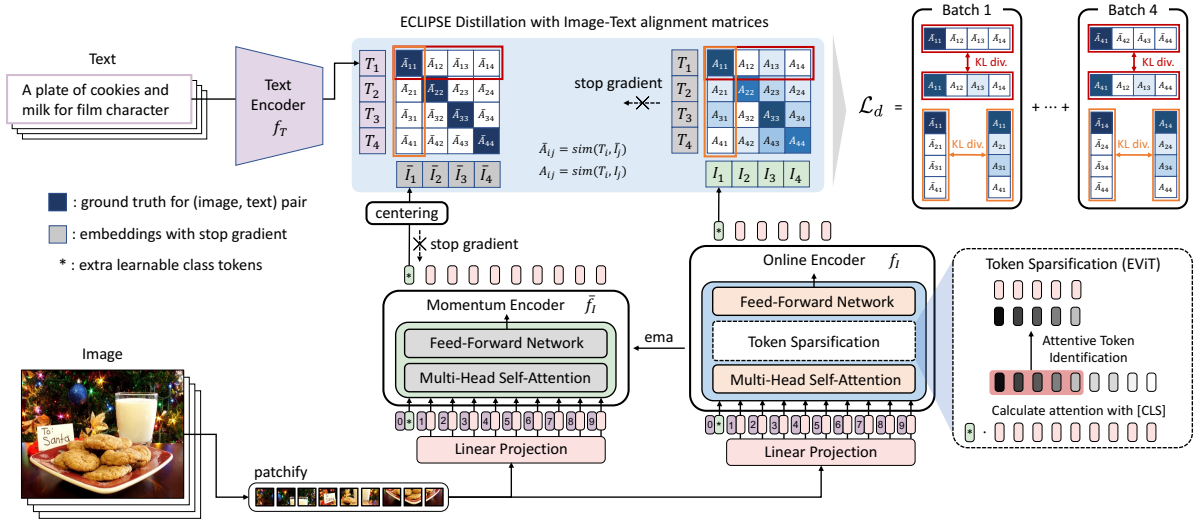

Figure 3: Overview of our proposed ECLIPSE. ECLIPSE is a meta-architecture for contrastive language-image pretraining that features a text encoder f_T, a momentum teacher encoder (Full ViT, $\bar{f}_I$), and a streamlined online encoder (ViT with token sparsification, f_I). Though the online network of ECLIPSE is compatible with any ViT acceleration method in literature (Liang et al. 2022b; Rao et al. 2021; Liang et al. 2022a), we choose EViT (Liang et al. 2022b) due to its simple architecture without introducing additional parameters. Full ViTs without any sparsification can be also adopted for the online network, in which ECLIPSE then provides a full-capacity model with enhanced performance.

where sg denotes stop-gradient and $\bar{I}_j = \bar{f}_I(x_j^I)$, $I_j = f_I(x_j^I)$ is the projected representations of j-th image with the momentum encoder and online encoder, respectively. Note that gradient is not calculated for the momentum encoder $\bar{I}$ as it is updated by EMA of the online encoder I.

We first obtain the loss for the teacher-text alignment matrix $\bar{A}$ with InfoNCE loss in Eq. (2), denoted as $\mathcal{L}_{\text{CLIP}}(\bar{A})$. Instead of training the online network to directly predict the output of the momentum network, we distill knowledge by predicting A to match $\bar{A}$. We define the distillation loss with KL divergence for each row and column between two matrices. Let σ be the softmax function, the KL divergence between A and $\bar{A}$ is rewritten as:

$$D_{\text{KL}}(\bar{A}||A) = \sum_{i=1}^{N} \sigma(\bar{A}_i) \log \frac{\sigma(A_i)}{\sigma(\bar{A}_i)}. \tag{5}$$

The overall distillation loss is the average of KL loss for row vectors and column vectors, written as $\mathcal{L}_{\text{distill}}(\bar{A}, A) = \frac{1}{2}(D_{\text{KL}}(\bar{A}||A) + D_{\text{KL}}(\bar{A}^T||A^T))$.

To accelerate training for the online network, we balance $\mathcal{L}_{\text{distill}}$ with InfoNCE loss $\mathcal{L}_{\text{CLIP}}(A)$ (Touvron et al. 2021). The final loss of the online network is then written as:

$$\mathcal{L}_{\text{online}} = \lambda \mathcal{L}_{\text{CLIP}}(A) + (1-\lambda)\mathcal{L}_{\text{distill}}(\bar{A}, A), \tag{6}$$

where λ is a parameter that balances the KL divergence loss and the InfoNCE loss. The final loss for ECLIPSE is then written as:

$$\mathcal{L} = \mathcal{L}_{\text{online}} + \mathcal{L}_{\text{CLIP}}(\bar{A}). \tag{7}$$

Momentum Update. Let θ_{f_I}, $\theta_{\bar{f}_I}$ be the parameter of the online image encoder and momentum encoder, respectively. For the t-th step, we update $\theta_{\bar{f}_I}^{(t)}$ of the momentum encoder according to the following:

$$\theta_{\bar{f}_I}^{(t)} = m\theta_{\bar{f}_I}^{(t-1)} + (1-m)\theta_{f_I}^{(t)}, \tag{8}$$

where m denotes the momentum parameter. We use $m = 0.994$ in our experiments. Momentum centering is also adopted for $\bar{f}_I$ (Caron et al. 2021) (see our supplement for further discussion with regard to the momentum parameter and centering).

Experiment

Implementation Details and Datasets

For implementation details, our work is built on top of the open-source SLIP codebase (Mu et al. 2021)[1]. For DeCLIP (Li et al. 2022), we follow the implementation details of the official code release[2]. The performance on GPU-machine runs for CLIP and SLIP follows the exact implementation details upon this codebase unless mentioned otherwise. All of our models are pretrained in 16× A100 GPUs. Further details can be found in the Appendix.

Pretraining datasets. To validate the effectiveness of ECLIPSE, we pretrain ECLIPSE on large-scale open-source datasets, CC (Conceptual Captions) 3M (Sharma et al. 2018) and YFCC (Yahoo Flickr Creative Commons) 15M (Thomee et al. 2016). Furthermore, to show the scalability of ECLIPSE, we curate 88M image-text pairs[3]. Since

[1]https://github.com/facebookresearch/SLIP
[2]https://github.com/Sense-GVT/DeCLIP
[3]Details of our curated dataset will be in our supplement

Method	VLC	SSL	MLM	Top1(%)
(a) CLIP	✓			17.10
(b) CLIP w/ EViT	✓			16.55
(c) ECLIPSE (CLIP)	✓			**19.67**
(d) SLIP	✓	✓		22.94
(e) SLIP w/ EViT	✓	✓		21.32
(f) ECLIPSE (SLIP)	✓	✓		**24.42**
(h) DeCLIP	✓	✓	✓	25.40
(i) DeCLIP w/ EViT	✓	✓	✓	23.26
(g) **ECLIPSE**	✓	✓		**26.41**

Table 1: ImageNet-1k Top 1 zero shot accuracy with models pretrained on CC3M dataset under three training configurations. Details of each configuration is denoted in Sec. . ECLIPSE outperforms previous CLIP variants (Radford et al. 2021; Mu et al. 2021; Li et al. 2022) across all training configurations. Note that naïvely adopting EViT to existing methods suffers from performance drop after acceleration.

Method	$\mathcal{L}_{\text{online}}$ in Eq. 6			Top1(%)
	λ	$\bar{A}$	A	
CLIP	-	-	-	17.1
ECLIPSE	0.5	$T \times \bar{I}$	$\text{sg}(T) \times I$	19.7
(a) ECLIPSE	1.0	$T \times \bar{I}$	$\text{sg}(T) \times I$	16.1
(b) ECLIPSE	0.5$\rightarrow$ 1	$T \times \bar{I}$	$\text{sg}(T) \times I$	18.5
(c) Output	0.5	$\bar{I}$	I	16.8
(d) Matrix	0.5	$\bar{T} \times \bar{I}$	$T \times I$	16.3
(e) Matrix	0.5$\rightarrow$ 1	$\bar{T} \times \bar{I}$	$T \times I$	18.6
(f) ECLIPSE + $\bar{T}$	0.5	$T \times \bar{I}$	$\bar{T} \times I$	18.8
(g) ECLIPSE + $\bar{T}$	0.5$\rightarrow$ 1	$T \times \bar{I}$	$\bar{T} \times I$	15.9

Table 2: ImageNet-1k Top-1 zero-shot accuracy for different $\mathcal{L}_{\text{online}}$ in Eq. 6. Different λ and ($\bar{A}$, A): (a-b) λ schedules (c) distillation of output, (d-e) distillation of momentum alignment matrix and (f-g) additional use of text momentum encoder for ECLIPSE, are tested. $T \times I$: alignment matrix between text and image embeddings; overbar indicates embeddings from momentum encoder. sg: stop-gradient.

the large-scale datasets (e.g., YFCC15M, 88M) feature extremely noisy text captions, intensive analysis is done with models pretrained on the relatively clean CC3M dataset.

Downstream datasets. Following CLIP (Radford et al. 2021), we evaluate the transferability of pretrained ECLIPSE on 11 widely used downstream datasets. We also transfer to zero-shot Image-Text retrieval tasks on Flickr30K and MS-COCO datasets. The evaluation settings for each dataset are consistent with CLIP as in the open-source implementation[1]. See more details of downstream datasets in our supplement.

Comparing ECLIPSE with CLIP variants

We first compare ECLIPSE against other state-of-the-art Contrastive Language-Image Pretraining approaches (Radford et al. 2021; Mu et al. 2021; Li et al. 2022). Table 1 shows the ImageNet zero-shot results of ECLIPSE and other CLIP variants, each grouped under identical experimental settings. All models are pretrained on the CC3M dataset with a learning rate 5e-4 for 40 epochs[4]. We use κ=0.7 for EViT with a ViT-B/16 backbone. For the first group (a,b,c), we compare models that only leverage Vision-Language Contrastive learning (VLC) between image-text pairs without any augmentation. In the second group (d,e,f), SimCLR loss (SSL) with two augmented image views is added to the aforementioned VLC. In the last group (h,i,g), we compare ECLIPSE with models trained with additional text augmentation (Wei and Zou 2019) (EDA).

In Table 1, we can observe that ECLIPSE on top of existing contrastive language-image pretraining pipelines, i.e., CLIP, SLIP, and DeCLIP, outperforms its baseline by a noticeable margin even with the online EViT encoder. On the other hand, naïvely applying EViT to existing pretraining pipelines (denoted as w/ EViT) results in lower performance. Note that ECLIPSE requires less training costs (see Sec.) and achieves 54% speed up in inference time (see Table 3). Furthermore, (g) is a simple extension of (f) where we add additional distillation loss for the augmented views (see details in our supplement). Even without leveraging language self-supervision (Masked Language Modeling), our streamlined ViT of ECLIPSE outperforms DeCLIP.

Keep Rate	CLIP Top1 Acc (%)	ECLIPSE Top1 Acc (%)	Throughput (image/s)
1.0 (=ViT)	**17.10**	**20.32** (+3.22)	564
0.9	16.82 (-0.28)	19.41 (+2.31)	662 (+17%)
0.8	16.68 (-0.42)	19.57 (+2.47)	758 (+34%)
0.7	16.55 (-0.55)	**19.67** (+2.57)	871 (+54%)
0.6	16.32 (-0.78)	18.80 (+1.70)	998 (+77%)
0.5	16.06 (-1.04)	18.37 (+1.27)	1132 (+101%)

Table 3: ImageNet-1k Top-1 zero-shot accuracy for CLIP and ECLIPSE after expediting vision encoders with different keep ratios (Liang et al. 2022b). All models were pretrained on CC3M dataset with a ViT-B/16 backbone. The relative performance difference compared to CLIP-ViT model is presented in the paranthesis.

Ablation Study

Here, we conduct ablation studies to validate how each component of ECLIPSE contributes to the final performance. All models in this section are pretrained in CC3M dataset with $\kappa = 0.7$ unless mentioned otherwise.

Variables for our Distillation. ECLIPSE is powered by a unique knowledge distillation structure where the image-text alignment matrix obtained by the online encoder and the text encoder predicts the alignment matrix jointly estimated by the momentum encoder and the text encoder. In Table 2, we ablate λ and distillation target in Eq 6. First, ECLIPSE can be trained with only hard labels without distillation ($\lambda = 1$). We observe that ECLIPSE outperforms (a)

[4]More detailed training configuration will be provided in supplement.

Method	Additional Supervision	Pets	CIFAR-10	CIFAR-100	SUN397	Food-101	Flowers	Cars	Caltech-101	Aircraft	DTD	ImageNet	**Average**
CLIP	-	19.4	62.3	33.6	40.2	33.7	6.3	2.1	55.4	1.4	16.9	31.3	27.5
SLIP	S	28.3	72.2	45.3	45.1	44.7	6.8	2.9	65.9	1.9	21.8	38.3	33.9
ECLIPSE	-	**24.7**	**67.8**	**38.8**	**44.4**	**34.0**	**6.2**	**2.8**	**56.7**	**2.1**	**19.6**	**32.7**	**30.0**
ECLIPSE	S	**31.3**	**79.5**	**46.0**	**46.4**	42.0	**7.2**	**3.3**	65.8	**2.5**	**22.5**	**39.5**	**35.1**

		Image-to-text retrieval						Text-to-image retrieval					
		Flickr30k			COCO Captions			Flickr30k			COCO Captions		
Method	Additional Supervision	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	-	34.9	63.9	75.9	20.8	43.9	55.7	23.4	47.2	58.9	13.0	31.7	42.7
SLIP	S	47.8	76.5	85.9	27.7	52.6	63.9	32.3	58.7	68.8	18.2	39.2	51.0
ECLIPSE	-	**42.6**	**71.4**	**83.8**	**24.9**	**50.6**	**62.4**	**28.9**	**53.0**	**64.2**	**15.1**	**35.4**	**47.0**
ECLIPSE	S	**50.2**	**77.4**	**87.5**	**27.9**	**53.9**	**65.9**	**33.6**	**59.6**	**70.9**	17.5	**39.6**	50.7

Table 4: Zero-shot image classification performance (single-modal) on 11 downstream datasets and image–text retrieval (multi-modal) on the test splits of Flickr30k and COCO Captions with models pre-trained on YFCC15M. Our ECLIPSE achieves competitive performance with other state-of-the-art works while resulting in 54% acceleration. Additional Supervisions other than Contrastive loss for image-text pairs are abbreviated as S: SSL between Augmentations.

learning from only hard labels, validating that our distillation loss contributes to the final performance. We also found that (b) progressively changing λ from 0.5 to 1 is worse than our default setting. We also checked the other extreme case, learning from only distillation ($\lambda = 0$), results in training failure as expected. Second, we compare ECLIPSE with previously proposed (c) feature-level distillation (Caron et al. 2021) where the student network directly predicts the output of the momentum teacher and (d-e) soft alignment matrix distillation (Li et al. 2021a; Lu et al. 2022) where image-text alignment matrix obtained from online encoders predicts the alignment matrix from momentum encoders for both image and text. We also test (f-g) replacing stop-gradient of text encoder with text momentum encoder which causes increase in training time. The result shows the supremacy of our proposed distillation over the existing distillation methods and ECLIPSE variants with additional text momentum encoder.

Token Keep Rate. Table 3 shows time[5] vs performance analysis of different keep rates (κ) for EViT (Liang et al. 2022b). We compare our proposed ECLIPSE with CLIP where EViT is directly applied. In the case of CLIP, the performance degrades as κ is lowered. On the other hand, ECLIPSE with ($\kappa = 0.7$) shows the highest performance among keep rates excluding full vision ($\kappa = 1.0$). We conjecture that the token dropping of EViT can affect contrastive learning since the partial view of the attentive tokens can be interpreted as an additional augmentation on the student network.

[5]We measure throughputs (128 batch, Avg of 100 runs) with https://github.com/youweiliang/evit/blob/master/helpers.py

Pretraining ECLIPSE on Larger Datasets

In this section, we pretrain ECLIPSE on larger scale dataset (e.g., YFCC15M)[6] and evaluate its transferability in single-modal and multi-modal downstream tasks. For simplicity, we measure the effectiveness of our ECLIPSE model with two versions: (i) ECLIPSE using only the original image-text pair ((c) in Table 1) and (ii) ECLIPSE with SimCLR (Chen et al. 2020a) loss between two augmented views ((f) in Table 1).

Zero-shot Classification. For single-modal experiments, we test the zero-shot classification performance on 11 downstream datasets. Table 4 shows the zero-shot classification accuracy of ECLIPSE pretrained on YFCC15M dataset and transferred to downstream classification datasets. For the test phase, the learned text encoder f_T synthesizes a zero-shot linear classifier by embedding the arbitrary categories of the test dataset. As classes are in the form of a single word, we use prompts including the label (e.g., "`a photo of a {label}`") as in CLIP (Radford et al. 2021). ECLIPSE with CLIP-level supervision outperforms CLIP across all 11 datasets, while ECLIPSE with additional supervision outperforms its corresponding baseline across 9 out of 11 datasets.

Image–Text Retrieval For multi-modal evaluations, we test the zero-shot image–text retrieval on Flickr30k and COCO Captions benchmarks. The image-text retrieval task can be split into two sub-tasks according to the target modality: image retrieval and text retrieval. Image-text pairs are ranked according to their similarity scores. Table 4 shows the zero-shot performance for image–text retrieval tasks of ECLIPSE pretrained on YFCC15M dataset. Our ECLIPSE

[6]More details of training configuration will be provided in supplement

Method	Supervision	3M	15M	88M
CLIP	C	17.1	31.3	57.4
ECLIPSE	C	**19.7**	**32.7**	**60.2**

Table 5: ImageNet-1k Top 1 zero shot accuracy for vision-language pretraining on different dataset scales. Both models were pretrained with a ViT-B/16 backbone for 3M and ViT-B/32 backbone for others. ECLIPSE shows a consistent tendency across various scales of pretrain datasets.

Encoder	CLIP ViT	ECLIPSE κ=1.0 + $\bar{T}$	ECLIPSE κ=1.0	ECLIPSE κ=0.7
Train time (s/batch)	0.409	0.500	0.484	0.415

Table 6: Training time comparison between CLIP and ECLIPSE variants. ECLIPSE achieves comparable training speed with CLIP even with disillation by removing text momentum encoder ($\bar{T}$) and replacing full ViT online encoder to streamline ViT.

outperforms its counterpart CLIP across all measures with a considerable gap.

Scalability In this section, we examine how ECLIPSE performs under various scales of pretraining dataset. In order to emphasize the effect of our meta-architecture ECLIPSE under image-text contrastive learning, we take the most simple form of Contrastive Language-Image Pretraining without any augmentation or self-supervision. Table 5 shows the zero-shot ImageNet Top1 accuracy of our streamlined ViT ($\kappa = 0.7$) after pretraining on each CC3M, YFCC15M, and our curated 88M dataset. Across various scales of pretraining datasets, ECLIPSE shows consistent performance improvement, thus validating the data scalability of our proposed method.

Discussion

Training Cost. In Table 6, we compare training speed of CLIP and ECLIPSE. The result shows that ECLIPSE can reach similar training speed of CLIP even with distillation. Consistent with our hypothesis, removing text momentum encoder ($\bar{T}$) and introducing expedition to the online encoder (κ=0.7) substantially boost the training speed compared to the naïve distillation with text momentum encoder. We also measure the average GPU memory usage during training[7]. With our GPU machine with 16×A100 GPUs, ECLIPSE ($\kappa = 0.7$) shows 18912 MiB/GPU average usage, showing a negligable increase compared to CLIP w/ EViT 18604 MiB/GPU, while being sufficiently efficient than CLIP trained with full ViT, 21758 MiB/GPU.

Efficient Image Self-Supervision for ECLIPSE. Previous CLIP variants (Mu et al. 2021; Li et al. 2022) have shown that incorporating self-supervised learning with augmented image views (e.g., SimCLR (Chen et al. 2020a), SimSiam (Chen and He 2021)) to the contrastive language-image pretraining can be advantageous for learning better visual representations. These works add additional forward and backward paths and MLP layers to treat the augmented views of an image. For example, SLIP-style image self-supervision can easily be applied to ECLIPSE as in Table 1(f). However, this requires two additional forward and backward computations, resulting in longer training time. Towards a more efficient self-supervised training, we here incorporate image SSL with the online and momentum branches of ECLIPSE. We introduce a shared projection head on top of the momentum and online encoder for SSL. This strategy is analogous to MoCo (He et al. 2020; Chen et al. 2020b) without a memory queue. We compare this efficient version (ECLIPSE-ES) with SLIP-style SSL and investigate the effect of augmentation in Table 7.

[7]Tested with 128 batches per GPU

Method	M	Online Encoder *view* 1	Online Encoder *view* 2	Online Encoder *view* 3	Top1 Acc. (%)	TPS (img/s)
CLIP	-	Img	-	-	17.10	**243**
SLIP	-	Img	Aug	Aug	22.94	129
ECLIPSE (SLIP)	Img	Aug	Aug	-	24.42	157
ECLIPSE-ES (a)	Img	Aug	-	-	23.91	**221**
ECLIPSE-ES (b)	Aug	Aug	-	-	**25.01**	**221**

Table 7: ImageNet-1k Top 1 zero shot accuracy with models pretrained on CC3M dataset under different number of image *views* with either simple cropping (Img) or data augmentation (Aug). M:momentum, TPS: throughput per second.

First, ECLIPSE-ES (b) surpasses the original SLIP and ECLIPSE (SLIP) even with fewer augmented views. From this result, we assume that the momentum encoder plays an essential role in the improved performance just as MoCo outperforms SimCLR in image SSL. Moreover, as the forward path of ECLIPSE is computed with the momentum encoder which does not require backward computation, ECLIPSE-ES features much shorter training time (see TPS column in Table 7). Second, we found that feeding augmented image views for both paths is better than using the original image for the teacher encoder: ECLIPSE-ES(b) vs ECLIPSE-ES(a). Consequently, ECLIPSE-ES demonstrates its training efficiency and opens up the possibility of advancement in integrating image SSL into VLP.

Conclusion

We propose ECLIPSE, a meta-architecture for streamlining vision transformers under visual–language pretraining. Our novel distillation formulation enables data-efficient training with accelerated ViTs under Contrastive Language-Image Pretraining. Our extensive experiments validate that there is a sweet spot between expedition and distillation where the partial view from the expedited online image encoder interacts complementarily with the momentum teacher. Future works will include extending ECLIPSE frameworks to other modalities.

References

Andonian, A.; Chen, S.; and Hamid, R. 2022. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Learning Representations*.

Chen, X.; Fan, H.; Girshick, R. B.; and He, K. 2020b. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020c. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born Again Neural Networks. In *International Conference on Machine Learning*.

Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; and Douze, M. 2021. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hessam Bagherinezhad, M. R., Maxwell Horton; and Farhadi, A. 2018. Label Refinery: Improving ImageNet Classification through Label Progression. *arXiv preprint arXiv:1807.03748*.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, H.; Liang, Y.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; and Zhou, M. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Jiang, Z.-H.; Hou, Q.; Yuan, L.; Zhou, D.; Shi, Y.; Jin, X.; Wang, A.; and Feng, J. 2021. All tokens matter: Token labeling for training better vision transformers. *Advances in Neural Information Processing Systems*.

Kim, B.; Jo, Y.; Kim, J.; and Kim, S. 2023. Misalign, Contrast then Distill: Rethinking Misalignments in Language-Image Pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2563–2572.

Kim, B.; Lee, J.; Kang, J.; Kim, E.-S.; and Kim, H. J. 2021. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 74–83.

Kim, B.; Mun, J.; On, K.-W.; Shin, M.; Lee, J.; and Kim, E.-S. 2022. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19578–19587.

Li, J.; Selvaraju, R. R.; Gotmare, A. D.; Joty, S.; Xiong, C.; and Hoi, S. 2021a. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021b. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-

semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*.

Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; and He, K. 2023. Scaling Language-Image Pre-training via Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *International Conference on Learning Representations*.

Liang, W.; Yuan, Y.; Ding, H.; Luo, X.; Lin, W.; Jia, D.; Zhang, Z.; Zhang, C.; and Hu, H. 2022a. Expediting large-scale vision transformer for dense prediction without fine-tuning. *Advances in Neural Information Processing Systems*.

Liang, Y.; GE, C.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2022b. EViT: Expediting Vision Transformers via Token Reorganizations. In *International Conference on Learning Representations*.

Liu, S.; Fan, H.; Qian, S.; Chen, Y.; Ding, W.; and Wang, Z. 2021a. HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lu, H.; Fei, N.; Huo, Y.; Gao, Y.; Lu, Z.; and Wen, J.-R. 2022. COTS: Collaborative Two-Stream Vision-Language Pre-Training Model for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2021. SLIP: Self-supervision meets Language-Image Pre-training. *arXiv preprint arXiv:2112.12750*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in Neural Information Processing Systems*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*.