SPEECH GUIDED MASKED IMAGE MODELING FOR VISUALLY GROUNDED SPEECH

Jongbhin Woo, Hyeonggon Ryu, Arda Senocak, Joon Son Chung

Korea Advanced Institute of Science and Technology, South Korea

ABSTRACT

The objective of this study is to investigate the learning process of Visually Grounded Speech (VGS) models through joint learning that combines contrastive learning and masked image modeling. Typically, VGS models aim to establish audio-visual alignment between images and their spoken captions within a contrastive learning framework. Building upon this seminal concept, in this work, we explore whether visual reconstruction with the help of cross-modality can enhance alignment, given that spoken captions describe visual appearances. To achieve this, we extend the contrastive learning-based VGS models by incorporating a masked autoencoder that utilizes cross-attention in the decoder. Through this cross-modal interaction in the decoder, spoken caption features guide the model to reconstruct the masked patches and capture correspondence between the two modalities. Our findings suggest that integrating cross-modal reconstruction within the contrastive learning framework enhances audio-visual feature alignment. Consequently, our proposed method gives comparable performance to existing models that utilize prior knowledge or other modalities, such as object region proposals or Contrastive Language-Image Pretraining (CLIP).

Index Terms— Visually Grounded Speech, Self-supervised Learning, Masked Autoencoder, Contrastive Learning

1. INTRODUCTION

Infants initially struggle to connect spoken words with objects but gradually learn through repeated exposure to unsegmented visuals and sounds. This process forms the basis for Visually Grounded Speech (VGS) models, which aim to replicate this learning mechanism. VGS models establish semantic relationships between words and visual representations, mimicking how infants acquire language.

VGS models establish feature alignment between two modalities: spoken caption and paired images, without relying on text information. Prior research [1, 2, 3, 4, 5, 6] have guided models to understand semantic information from spoken utterances, using visual information as a supervisory signal, and to learn the shared feature space. Such models build visual-spoken language understanding without necessitating text-level annotation. VGS research has focused on the correspondence between speech and image modalities. In this domain, the predominant approach has been contrastive learning [2]. Previous studies enhance their models using either off-the-shelf object detector [7] or Contrastive Language-Image Pretrained (CLIP) models [8]. Earlier research leverage visual signals to employ the spoken utterance encoder for various tasks, such as subword detection [9], image captioning [10], and zero-shot speech segmentation [6]. A different research trajectory examines VGS in the context of multilingual spoken language learning, with the goal of aligning features across diverse languages. [11, 12]

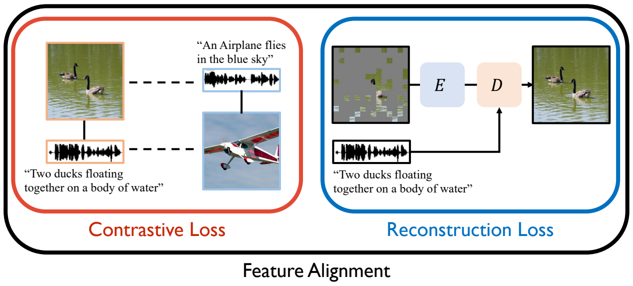
"An Airplane flies in the blue sky"
"Two ducks floating together on a body of water"
Contrastive Loss
E
D
"Two ducks floating together on a body of water"
Reconstruction Loss
Feature Alignment

Fig. 1. Speech-Guided Masked Image Modeling (MIM) assists the model to enhance the audio-visual correspondence for Visually Grounded Speech (VGS).

As an essential characteristic of VGS, the model learns the audio-visual correspondence in a self-supervised manner. Given this, it is natural to explore another stream of self-supervised training: masked image modeling (MIM). In the vision-language field, which is closely related to VGS and distinguished by the use of either spoken or written language form, studies investigate whether MIM is beneficial for contrastive language-image pretraining. These studies integrate contrastive loss with masked language modeling, adopt masked self-distillation [13], or employ cross-modal reconstruction [14]. Given this context, it is pertinent to question if these approaches remain effective for spoken captions, which are more continuous and less explicitly formed than the text. To address this, we introduce a method that combines Speech-Guided MIM and contrastive learning for VGS, where the spoken utterance features assist the model in reconstructing the masked patches.

In this paper, we investigate the potential of cross-modal reconstruction in contrastive learning for VGS. As illustrated in Figure 1, our learning process is divided into two parts: Contrastive Learn-

This work was supported by the National Research Foundation of Korea grant funded by the Korean government (Ministry of Science and ICT, RS-2023-00212845) and the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2023-00259991) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

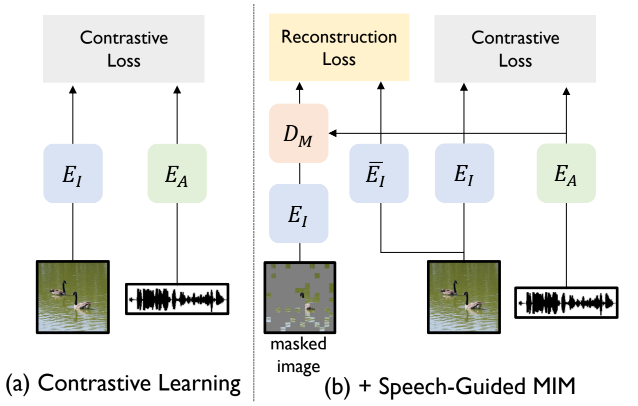

Fig. 2. **Comparison between contrastive learning and the proposed method.** Our approach exploits the correspondence between images and spoken captions using a cross-modal decoder together with Masked Image Modeling (MIM).

ing and Speech-Guided MIM. For contrastive learning, we extract audio and full image features using the audio encoder and image encoder, respectively. Concurrently, we mask a large portion of the image and subsequently extract features from this masked image. The cross-modal decoder then takes over to reconstruct the masked patches, using the spoken utterance feature to guide the reconstruction of the masked region. Under the guidance of cross-modal reconstruction, we demonstrate that our model gives on-par or better performance than existing benchmarks, including those that utilize additional prior knowledge.

2. APPROACH

An overview of our approach is depicted in Figure 2. Our design integrates the contrastive learning-based VGS model with MIM, targeting enhanced alignment within the audio-visual feature space. The subsequent section provides details on the training process.

2.1. Preliminaries

Our proposed model comprises four primary components: an image encoder $E_I(\cdot;\theta_I)$, an audio encoder $E_A(\cdot;\theta_A)$, a cross-modal decoder $D_M(\cdot;\theta_M)$, and a momentum image encoder $\bar{E}_I(\cdot;\bar{\theta_I})$. The momentum encoder's parameters, $\bar{\theta_I}$, are updated via $\bar{\theta_I} = \alpha\bar{\theta_I} + (1-\alpha)\theta_I$, where α denotes the momentum coefficient. All these components are based on the transformer architecture [15].

For a given image-spoken caption pair, (I,A), the input image is defined as $I \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the image's height, width, and number of channels, respectively. We partition I into N non-overlapping patches, expressed as $I = \{i^1,...,i^N\}$, with each feature vector $i^n \in \mathbb{R}^{P^2C}$. Here, N is calculated as $H\times W/P^2$, and P indicates the height and width of each patch. The input audio A, initially a raw waveform, is transformed into a set of feature vectors, $A = \{a^1,...,a^T\}$, using a convolutional block of E_A. Both feature vector sets, I and A, are augmented with a [CLS] token and subsequently processed by the image encoders E_I and audio encoder E_A, respectively:

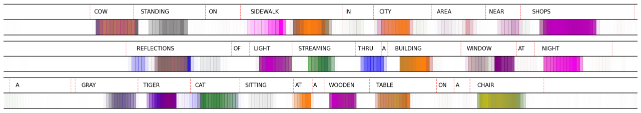

Fig. 3. Illustration of attention weights allocated to each frame by the [CLS] token, derived from the last layer of the audio encoder. Each color in the figure represents a different attention head.

$$\begin{aligned} E_I(I) &= Z_I = \{z^I_{cls}, z^I_1, z^I_2, ..., z^I_N\} \\ E_A(A) &= Z_A = \{z^A_{cls}, z^A_1, z^A_2, ..., z^A_T\}. \end{aligned} \tag{1}$$

For image reconstruction, we sample the set of masked patch indices, $\mathcal{M}$, resulting in $I_m = \{i^k \mid k \in \mathcal{M}\}$ and $I_v = \{i^k \mid k \notin \mathcal{M}\}$.

2.2. Training

We present the two distinct components of our methods: Contrastive Learning and Speech-Guided MIM.

Contrastive Learning. The alignment between spoken captions and image features comes from assigning the paired features close to each other. The prevailing trend in the VGS field establishes the alignment within the framework of contrastive learning. We input the mean feature of Z_I and z^A_{cls} into their respective projection layers, yielding global representations denoted as z^I and z^A. Subsequently, the inner product of these global representations is computed to derive a similarity score. The objective is to amplify this score for spoken caption-image pairs. We use InfoNCE [16] loss as our contrastive loss function $\mathcal{L}_{NCE}$ defined as:

$$\begin{aligned} \mathcal{L}_I &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z^I_i z^A_i\right)}{\sum_{j=1}^{B}\exp\left(z^I_i z^A_j\right)} \\ \mathcal{L}_A &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z^A_i z^I_i\right)}{\sum_{j=1}^{B}\exp\left(z^A_i z^I_j\right)} \\ \mathcal{L}_{NCE} &= \mathcal{L}_I + \mathcal{L}_A, \end{aligned} \tag{2}$$

where, B denotes the number of image-spoken caption pairs present in a mini-batch.

Speech-Guided Masked Image Modeling. The Speech-Guided Masked Image Modeling (MIM) learns the representation through the reconstruction of the masked visual tokens with the help of spoken features. The spoken caption involves in the MIM through the interaction of cross-attention in the cross-modal decoder. As the image-spoken caption pair datasets consist of the image and the spoken caption that describes the visual scene, the natural correspondence in the cross-modalities are guaranteed.

In the reconstruction process, the output feature of the audio encoder, denoted as Z_A, is input into the cross-modal decoder D_M. This decoder is characterized by its cross-attention layer, where Z_A plays a pivotal role, serving as both the key and value. This setup enables an interaction between audio and visual features in the cross-attention mechanism. Given that audio features can include semantically redundant elements, which are not beneficial for reconstruction, it is essential to filter and pass only the semantically significant parts to D_M. To facilitate this, we utilize the attention weights W of the [CLS] token across each temporal frame from the last layer of the audio encoder, as depicted in Figure 3 and outlined in [6]. By performing matrix multiplication of $\{z_1^A, z_2^A, ..., z_T^A\}$ with these attention weights, we generate summarized speech features whose length matches the number of heads in the transformer layer as:

$$Z^A_{summarized} = W \cdot \{z_1^A, z_2^A, ..., z_T^A\}. \tag{3}$$

During the masking phase, we randomly select the masked patch indices, denoted as $\mathcal{M}$. The image encoder E_I, which shares parameters with the contrastive learning component, is fed only the unmasked visual tokens, I_v. Subsequently, the cross-modal decoder D_M processes these visible image features along with the summarized speech features. In line with the approach in [13], our reconstruction goal is to replicate the features created by the momentum encoder $\bar{E}_I$, utilizing the entire image. We employ Mean Square Error (MSE) as the loss function in our MIM for optimization:

$$\begin{gathered} D_M(E_I(I_v), Z^A_{summarized}) = \{z_1^R, z_2^R, ..., z_N^R\} \\ \bar{E}_I(I) = \{\bar{z}_1^I, \bar{z}_2^I, ..., \bar{z}_N^I\} \\ \mathcal{L}_{MIM} = -\frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} \left\| \bar{z}_k^I - z_k^R \right\|_2^2. \end{gathered} \tag{4}$$

Our learning objective with Speech-Guided MIM is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{NCE} + \lambda \mathcal{L}_{MIM}, \tag{5}$$

where λ represents the weight for $\mathcal{L}_{MIM}$.

3. EXPERIMENTS

3.1. Datasets

We employ three primary datasets for training and testing our models: Places Audio [17], Flickr8K Audio Captions Corpus [1] (FACC), and SpokenCOCO [10]. All these datasets comprise pairs of images and their corresponding spoken captions. The Places Audio dataset offers 400K pairs, where captions were spontaneously recorded by annotators as they viewed the associated image, and for our experiments, we use the "2020" splits. SpokenCOCO provides 123K images, each paired with 5 unique spoken captions derived from the text captions of the MSCOCO [18] dataset; for this dataset, we follow the Karpathy split as outlined in [6]. Lastly, FACC contains 8K images and each image is paired with 5 distinct spoken captions which originates from Flickr8K's text captions. For the evaluation, we use the standard train and test splits.

3.2. Implementation Details

For the image encoder E_I, we employ the ViT-B/16 architecture as described in [20]. This consists of a 12-layer transformer with a width of 768 and 12 attention heads. We leverage pre-trained weights from [21]. The momentum encoder $\bar{E}_I$ has the identical architecture of the image encoder. The cross-modal decoder D_M incorporates a single cross-attention layer, which is composed of cross-attention, self-attention, and a feed-forward layer. Our audio encoder's design and initial weights are derived from HuBERT Base [22]. Following the approach in [7], we reinitialize the last 3 layers of the pretrained HuBERT. This encoder contains a 7-layer convolutional block and a 12-layer transformer, with a width of 768 and 12 attention heads similar to [20]. We project the global representations from each modality using projection layers, each being a linear layer with an output dimension of 768. We use the BertAdam [23] optimizer with an initial learning rate of 0 that linearly increases to 5×10^{-6} during the initial 10% of training process. Subsequently, the learning rate decays to 0 in line with VG-HuBERT [7]. The momentum coefficient α is gradually increased from 0.999 to 1 following a cosine scheduler. For data augmentation, images are cropped to a size of 224 x 224, and RandAugment [24] is applied. In alignment with MAE [21], we set a masking ratio of 75%. Training is conducted with a batch size of 100 over 30 epochs for SpokenCOCO, 30 epochs with a batch size of 80 for Places Audio, and 20 epochs with a batch size of 100 for FACC. We introduce the reconstruction loss after 3 epochs, setting its weight λ at 2.0.

3.3. Baselines

Before presenting the quantitative results of the comparison with other existing works and closely-related baselines, we introduce the details of these baselines below:

VG-HuBERT [6]: The architecture of VG-HuBERT corresponds to model (a) in Figure 2, employing contrastive learning as its objective function. The image encoder is based on ViT-S/8 and inherits its weights from DINO [25]. Meanwhile, the audio encoder uses the structure and weights of the HuBERT Base.

FaST-VGS [7]: This model incorporates a region proposal network before the image encoder transformer, processing the detected regions instead of the entire image. FaST-VGS adopts a coarse-to-fine approach through both feature-wise inner products and a cross-modal encoder. For a fair comparison, we restrict our evaluation to the model's coarse retrieval performance, specifically considering the retrieval score based on inner product.

SpeechCLIP [8]: SpeechCLIP is a model that leverages the well-organized Contrastive Language-Image Pretrain (CLIP) image encoder which is trained on large-scale image-text dataset.

VG-HuBERT (M): This architecture closely follows that of VG-HuBERT, with a notable exception in the image encoder. Given that ViT-S/8 is not feasible when incorporating the additional component, we opt for the ViT-B/16. We initialize this encoder with pre-trained weights from MAE [21], where M stands for MAE image encoder. We note that our proposed method utilizes the cross-modal decoder on top of this model.

3.4. Quantitative Results

In this section, we provide comparison with our method and previous approaches introduced in Section 3.3 in three different dataset and ablation study.

Comparing our methods with existing baselines. This section discusses the performance of our method in comparison with previous methods. As demonstrated in Table 1 and Table 2,

Table 1. **Quantitative results on Places Audio test-seen and test-unseen.** All models are trained on Places train set. † denotes a model which employs prior knowledge.

	Places Audio (test-seen)						Places Audio (test-unseen)					
	A → I			I → A			A → I			I → A		
Model	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ResDAVEnet[2]	35.2	67.5	78.0	30.4	63.1	74.1	38.3	68.5	78.8	31.2	65.0	75.4
MILAN[19]	58.4	84.6	90.6	53.8	83.4	90.1	62.1	86.0	90.5	58.2	85.8	90.9
FaST-VGS[7]†	60.0	86.1	92.3	60.2	85.1	92.2	62.8	88.4	92.9	62.3	89.0	93.2
VG-HuBERT (M)	60.4	86.9	92.1	59.3	87.0	92.6	63.8	89.1	93.6	63.6	88.3	93.2
Ours	**62.5**	**88.0**	**93.7**	**61.3**	**87.8**	**93.5**	**67.4**	**89.5**	**94.5**	**64.6**	**89.7**	**93.9**

Table 2. **Quantitative results on SpokenCOCO test sets.** All models are trained on SpokenCOCO train set. † denotes a model which uses prior knowledge.

	A → I			I → A		
Method	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
	SpokenCOCO (test)					
ResDAVEnet[2]	17.3	41.9	55.0	22.0	50.6	65.2
VG-HuBERT[6]	30.6	60.8	72.8	42.8	73.5	83.9
FaST-VGS[7]†	31.8	62.5	**75.0**	42.5	73.7	84.9
VG-HuBERT (M)	30.6	60.5	72.7	42.2	74.1	84.1
Ours	**32.2**	**62.8**	74.8	**45.5**	**75.8**	**85.9**

Table 3. **Quantitative results on FACC test sets.** All models are trained on FACC train set. † denotes a model which utilizes prior knowledge or other modality.

	A → I			I → A		
Method	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
FaST-VGS [7]†	26.6	56.4	68.8	36.2	66.1	76.5
SpeechCLIP [8]†	26.7	**57.1**	**70.0**	**41.3**	**73.9**	**84.2**
VG-HuBERT (M)	25.9	54.9	68.1	36.1	66.8	78.8
Ours	**27.4**	56.9	69.5	36.7	67.6	79.5

Table 4. **Ablation study results on Places Audio test-seen and test-unseen.** All models are trained on Places train set. Here, R@1, R@5, and R@10 represent the average recall rates for image-to-audio and audio-to-image retrieval tasks.

	test-seen			test-unseen		
Method	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
VG-HuBERT (M)	59.9	86.9	92.3	63.7	88.7	93.4
+ Recon	60.7	87.0	91.7	65.5	89.3	93.3
+ Speech-Guided	61.1	87.5	92.1	64.9	88.3	93.2
+ Summary (Ours)	**61.9**	**87.9**	**93.6**	**66.0**	**89.6**	**94.2**

our method show better performance, achieving state-of-the-art results on both the Places Audio and SpokenCOCO datasets. Specifically, on the Places Audio test-seen and test-unseen splits, our method consistently outperforms existing methods across all metrics (R@1, R@5, R@10), for both Audio to Image (A → I) and Image to Audio (I → A) retrieval tasks. Similarly, in the SpokenCOCO test sets, our approach surpasses other methods, demonstrating its robustness and effectiveness in varied settings. Our methods consistently demonstrate superior performance compared to VG-HuBERT (M) and FaST-VGS across different datasets. Notably, FaST-VGS relies on an off-the-shelf object detector, contributing to its performance; however, our method achieves the results without such reliance on prior knowledge.

In the context of the FACC dataset, as shown in Table 3, SpeechCLIP outperforms ours. We attribute this to the small size of the FACC dataset, where SpeechCLIP benefits significantly from leveraging a powerful vision-language foundation model.

Ablation study on different reconstruction strategies. Our ablation study, summarized in Table 4, evaluates the contributions of different components on the Places Audio test-seen and test-unseen datasets. The study compares the baseline VG-HuBERT (M) model with incremental enhancements: the addition of reconstruction (+ Recon), speech guidance (+ Speech guided), and our complete model with audio summarization (+ Summary).

The results indicate that each component incrementally improves the model. The addition of reconstruction (+ Recon) delivers an initial enhancement over the baseline VG-HuBERT (M), especially in test-unseen scenarios. The incorporation of speech guidance (+ Speech-Guided) also contributes to retrieval performance, but its impact is more subtle. While certain metrics show a slight increase, the overall advancement is moderate. However, the most significant improvement is observed with the introduction of audio summarization (+ Summary). This enhancement is evident in both test-seen and test-unseen scenarios, surpassing the VG-HuBERT (M) model. This suggests that selectively passing semantically significant parts to the cross-modal decoder, a process achieved through our audio summarization technique, effectively leverages Speech-Guided MIM.

4. CONCLUSION

In this paper, we focus on enhancing the audio-visual alignment with the motivation of visual reconstruction with the spoken utterances, as they contain the visual descriptions. We propose a model that combines Contrastive Learning with Speech-Guided MIM, where cross-modal interaction is also utilized in the decoder for reconstruction. Our experiments demonstrate on-par or better performance than existing benchmarks, including those that utilize additional prior knowledge. We believe that our proposed method, which utilizes MIM, introduces a new perspective to VGS-based models and will be beneficial to the community by paving the way for new directions.

5. REFERENCES

[1] David Harwath and James Glass, “Deep multimodal semantic embeddings for speech and images,” in *ASRU*, 2015.

[2] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in *Proc. ECCV*, 2018.

[3] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi, “Representations of language in a model of visually grounded speech signal,” in *Proc. ACL*, 2017.

[4] Herman Kamper, Aristotelis Anastassiou, and Karen Livescu, “Semantic query-by-example speech search using visual grounding,” in *Proc. ICASSP*, 2019.

[5] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al., “Avlnet: Learning audio-visual language representations from instructional videos,” in *Proc. Interspeech*, 2021.

[6] Puyuan Peng and David Harwath, “Word discovery in visually grounded, self-supervised speech models,” in *Proc. Interspeech*, 2022.

[7] Puyuan Peng and David Harwath, “Fast-slow transformer for visually grounding speech,” in *Proc. ICASSP*, 2022.

[8] Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath, “SpeechCLIP: Integrating speech with pre-trained vision and language model,” in *SLT*, 2023.

[9] David Harwath, Wei-Ning Hsu, and James Glass, “Learning hierarchical discrete linguistic units from visually-grounded speech,” in *Proc. ICLR*, 2020.

[10] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass, “Text-free image-to-speech synthesis using learned segmental units,” in *Proc. ACL*, 2021.

[11] David Harwath, Galen Chuang, and James Glass, “Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech,” in *Proc. ICASSP*, 2018.

[12] Hyeonggon Ryu, Arda Senocak, In So Kweon, and Joon Son Chung, “Hindi as a second language: Improving visually grounded speech with semantically similar samples,” in *Proc. ICASSP*, 2023.

[13] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al., “Maskclip: Masked self-distillation advances contrastive language-image pretraining,” in *Proc. CVPR*, 2023.

[14] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto, “Masked vision and language modeling for multi-modal representation learning,” in *Proc. ICLR*, 2023.

[15] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[17] David Harwath, Antonio Torralba, and James Glass, “Unsupervised learning of spoken language with visual context,” in *NeurIPS*, 2016.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in *Proc. ECCV*, 2014.

[19] Ramon Sanabria, Austin Waters, and Jason Baldridge, “Talk, don’t write: A study of direct speech-based image retrieval,” in *Proc. Interspeech*, 2021.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, “Masked autoencoders are scalable vision learners,” in *Proc. CVPR*, 2022.

[22] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *TASLP*, 2021.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. ACL*, 2019.

[24] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proc. CVPR*, 2020.

[25] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, “Emerging properties in self-supervised vision transformers,” in *Proc. ICCV*, 2021.