北京理工大學

留学硕士学位论文
开题报告

选题名称 **Monaural Audio-Visual Semantic Segmentation**

学　　号 **(宋体，三号，30 磅)**

姓　　名

导　　师

学　　科

研究方向

学　　院

年　　月　　日

填　表　说　明

1.只有学籍状态为注册或暂缓注册的研究生才允许开题。但学籍状态为暂缓注册的研究生只有在完成注册手续之后，开题报告及其评审结果才能被认可。

2．开题报告为 A4 大小，于左侧装订成册。硕士研究生应逐项认真填写，各栏空格不够时请自行加页。

3.开题报告经指导教师审阅通过后，由留学硕士研究生公开宣读，并接受专家组质疑、评议。考核小组由 5 名及以上本学科或相关学科的副高级及以上职称专家或硕士生导师组成，导师可为考核小组成员，但不得担任考核小组组长。评审合格后，由研究生装订后交指导教师留存备查，同时把含签字的完整开题报告电子版上传研究生管理系统。

4．留学硕士研究生应在选题前阅读至少 50 篇本研究领域及方向的国内外文献，了解、学习本领域的前沿和研究进展，并写出不少于 5000 字（中文）/词（英文）的文献综述报告。文献综述报告应反映国际和国内本领域的研究历史、现状和发展趋势。文献综述报告是开题报告的必要附件，开题报告通过后，由研究生装订后交指导教师留存备查，同时把含签字和评阅意见的完整文献综述电子版上传研究生管理系统。

5.“参考文献”著录按照 GB7714-87 文参考文献著录规则执行。书写顺序为：序号·作者·论文名或著作名·杂志或会议名·卷号、期号或会议地点·出版社·页号·年。

6. 本件只接受非涉密内容

一 简表

<table>
<tr><td rowspan="4">研究生简况</td><td>姓名</td><td></td><td>性别</td><td></td><td>出生年月</td><td></td></tr>
<tr><td>学号</td><td></td><td>入学时间</td><td></td><td>身份证号</td><td></td></tr>
<tr><td colspan="2">学科、专业</td><td colspan="4"></td></tr>
<tr><td colspan="2">本科毕业时间</td><td></td><td>本科毕业学校</td><td colspan="2"></td></tr>
<tr><td colspan="2">指导小组</td><td>姓名</td><td>职称</td><td colspan="2">工作单位</td><td>签字</td></tr>
<tr><td colspan="2">导师</td><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td colspan="2" rowspan="3">小组成员</td><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td rowspan="8">研究课题</td><td rowspan="2">名称</td><td>中文</td><td colspan="4"></td></tr>
<tr><td>英文</td><td colspan="4"></td></tr>
<tr><td colspan="2">开题状态</td><td colspan="4">首次开题（ ）； 再审核后开题（ ）</td></tr>
<tr><td>类别</td><td colspan="5">国家项目（ ）； 部（省）项目（ ）； 企业项目（ ）；
自拟项目（ ）； 是否兵器类项目（ ）；</td></tr>
<tr><td>性质</td><td colspan="5">基础研究（ ）； 应用基础研究（ ）； 应用技术研究（ ）</td></tr>
<tr><td colspan="2">与导师课题研究课题的关系</td><td colspan="4">是导师研究课题的一部分（ ）
与导师研究课题无关（ ）</td></tr>
<tr><td colspan="6">摘 要</td></tr>
<tr><td colspan="6">Audio-visual segmentation (AVS) is a task which the goal is to localize the sounding objects in pixel-wise level, and audio-visual segmentation with semantic (AVSS) aims to localize and classify the sounding objects. AVSBench and AVSS dataset are the first significant benchmark in evaluating the study. However, there exposes limitations to the robustness due to the non-availability of rich examples. We proposed a new method that uses an audio-visual attentive feature fusion module to inject audio semantics as guidance for the visual segmentation process over a widely used encoder-decoder based model. Two model variant were introduced: ResNet based and Swin Transformer based models. A custom regularization technique is deployed to encourage higher robustness to unseen data while maintained a better performance. Extensive experiments demonstrate the effectiveness of our method as well as the significance of cross-modal perception and dependency modeling for this task.</td></tr>
<tr><td rowspan="3">关键词</td><td colspan="5">1.关键词限 3~5 个；2.关键词之间用“；”分隔。</td></tr>
<tr><td>中文</td><td colspan="4"></td></tr>
<tr><td>英文</td><td colspan="4">Audio−Visual Segmentation; Multimodality</td></tr>
</table>

二 选题依据

简述该选题的研究意义、国内外研究概况和发展趋势。

Audio-visual segmentation (AVS) is a new task in which the goal is to output a pixel-level map of the objects that produce sound at the time of the image frame. On the other hand, audio-visual segmentation with semantics (AVSS) is built on top of AVS which the goal is not just localizing the sounding objects but also to identity the objects. Both of them are important because they can help us better understand and analyze audio-visual scenes, which is useful in many applications such as surveillance, robotics, and human-computer interaction. The challenge is to accurately identify and classify the sounding objects in a given audio-visual scene, which is a difficult task due to the complex and dynamic nature of audio-visual data. The authors in [15] proposed a method that uses a temporal pixel-wise audio-visual interaction module to inject audio semantics as guidance for the visual segmentation process. This method is evaluated on a benchmark dataset called AVSBench [14] and AVSS [15], which provides pixel-wise annotations for the sounding objects in audible videos. The AVSBench includes two subsets: semi-supervised audio-visual segmentation with a single sound source and fully-supervised audio-visual segmentation with multiple sound sources. While AVSS further extend the previous AVSBench by introducing fully-supervised audio-visual segmentation with semantic labels. However, despite the effort, the public available AVS related datasets are still insufficient to reflect the real-world scenarios, and the small sizes of the datasets are then resulting in poor performance of recent works. We proposed a novel segmentation model trained with our custom regularization technique to tackle AVS and AVSS problems.

Background

Researchers have studied audio-visual relation problem thus far using a few overly simplistic examples. First, the audio-visual correspondence (AVC) problem where the goal is to seek ascertain whether an auditory signal and a visual image describe the same scenario, is being studied by several academics [[1], [2], [3]]. AVC associating audio signals with visual signals in a meaningful way, particularly when there is a time delay between the two signals or when they are corrupted by noise or other distortions. In other words, the AVC problem is about determining which sounds in an audio signal correspond to which visual events in a video signal, and vice versa. For example, in a video conference call, the AVC problem involves synchronizing the audio and video signals to ensure that the speaker’s lips and the sound of their voice are properly

aligned. In a robot navigation scenario, the AVC problem involves identifying the location of a sound source based on the robot's visual observations.

Others focused on audio-visual event localization (AVEL) [[4], [5], [6], [7], [8]], which maps video clips into pre-established event labels. In other words, given an audio-visual recording of a scene, AVEL aims to identify the spatial and temporal locations of events of interest in the scene, such as the sound of a car passing by or a person speaking. Similar to this, others have also looked at audio-visual video parsing (AVVP), which aims to separate a video into various events and categorize them as audible, visible, or both. All these instances are limited to the frame/temporal level due to a lack of pixel-level annotations, which reduces the problem to audible image categorization.

Sound source localization (SSL), a similar problem, seeks to identify the visual regions inside the frames that match the sound [[9], [10], [11], [12]]. In contrast to AVC/AVEL/AVVP, SSL seeks patch-level scene understanding; as a result, results are typically displayed as heat maps created either by visualizing the similarity matrix between the audio feature and the visual feature map or by class activation mapping (CAM) [13], which ignores the actual shape of the sounding objects.

In compare to general audio-visual localization, the effort in creating a system that capable to output pixel-wise semantic map of sounding object is quite poor. Among the publicly available datasets, for instance, the AVE and LLP datasets are gathered for tasks involving the audio-visual events localization and video parsing. They can't be utilized for pixel-level segmentation because they can only annotate video frames with categories. Researchers often employ the Flickr-SoundNet and VGG-SS datasets for the sound source localization problem, where the videos are samples from the massive Flickr and VGGSound datasets, respectively. The authors include bounding boxes that show where the intended sound source is located; these boxes might be used as patch-level supervision. Since the sounding items are typically asymmetrical in shape and certain areas within the bounding box are in fact unrelated to the actual sound source, this will surely produce inaccurate evaluation findings. This is the cause of the limitations of existing sound source localization techniques, which are unable to learn the precise forms of sounding objects and hence prevent the mapping of audio signals to fine-grained visual clues. AVS segments the sounding visual objects, with a fully-supervised trained model that recognizes every sounding visual object pixel in a scene. To tackle the problem, AVSBench [[14], [15]] dataset introduced. The AVSBench [15] is an essential AVS benchmark that has enabled the testing of

the first AVS algorithms. Although it took more time to create the pixel-level binary annotations of sounding objects [16], they still represent a more cost-effective option than a more intricate but more desirable pixel-level multi-class annotation [15].

Reference

[1] Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 609–617 (2017)

[2] Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. Advances in Neural Information Processing Systems 29 (2016)

[3] Arandjelović, R., Zisserman, A.: Objects that Sound. In: *arXiv*. https://doi.org/https://arxiv.org/abs/1712.06651v2. (2017)

[4] Lin, Y.B., Li, Y.J., Wang, Y.C.F.: Dual-modality seq2seq network for audio-visual event localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2002–2006. IEEE (2019)

[5] Lin, Y.B., Wang, Y.C.F.: Audiovisual transformer with instance attention for audio-visual event localization. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2020)

[6] Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 247–263 (2018)

[7] Zhou, J., Zheng, L., Zhong, Y., Hao, S., Wang, M.: Positive sample propagation along the audio-visual event line. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8436–8444 (2021)

[8] Ramaswamy, J., Das, S.: See the sound, hear the pixels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2970–2979 (2020)

[9] Wu, X., Wu, Z., Ju, L. and Wang, S.: Binaural Audio-Visual Localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. 35, pp. 2961-2968 (2021)

[10] Masuyama, Y., Bando, Y., Yatabe, K., Sasaki, Y., Onishi, M., Oikawa, Y.: Self-supervised Neural Audio-Visual Sound Source Localization via Probabilistic Spatial Modeling. In: arXiv. https://doi.org/https://arxiv.org/abs/2007.13976v1. (2020)

[11] Geng, T., Wang, T., Duan, J., Cong, R., Zheng, F.: Dense-Localizing Audio-Visual Events in Untrimmed Videos: A Large-Scale Benchmark and Baseline. In: arXiv. https://doi.org/https://arxiv.org/abs/2303.12930v2. (2023)

[12] Rachavarapu, K. K., Aakanksha, A., Sundaresha, V., N, R. A.: Localize to Binauralize: Audio Spatialization from Visual Sound Source Localization. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1910-1919 (2021)

[13] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016)

[14] Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-Visual Segmentation. In: *arXiv*, abs/2207.05042v3. (2022)

[15] Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-Visual Segmentation with Semantics. In: ArXiv, abs/2301.13190. (2023)

[16] Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 12674-12684. (2020)

三 研究内容

论证提出课题研究的总体思路，拟开展的主要研究内容及内涵，说明各部分研究内容之间的相关关系。

Research Problem Statement

Previous works have faced limitations and challenges in building unbiased audio-visual semantic segmentation benchmarks. One of the main challenges is the biased nature of the training sets, which contain audio-visual data that may not be representative of the real-world scenarios. Additionally, previous works have relied on matching images and audio based on the semantic classes of visual objects, which can introduce biases and inaccuracies in the resulting datasets. They used a method of matching images and audio based on the semantic classes of visual objects to create audio-visual datasets. This method involves associating audio with visual objects based on their semantic labels, such as “dog,” “car,” or “person.” However, this approach can introduce biases and inaccuracies in the resulting datasets because it assumes that the audio and visual objects are always related in a one-to-one manner. In reality, there may be multiple sounds or objects present in a scene, and they may not always correspond to the same semantic label. Furthermore, the size of the public available datasets are still insufficient to represent real-world scenarios. In numbers, AVS and AVSS dataset contains over 13k and 80k annotated frames respectively (AVSS contains all samples from AVS), while COCO dataset contains over 200k annotated images. These limitations should be addressed, since it can lead to poor generalization of models beyond the training set and hinder the development of accurate audio-visual semantic segmentation systems.

Research Questions

Q1. How well AVSBench dataset generalizes audio-visual segmentation task? How to increase the robustness of the proposed model?

Q2. What efforts can be done in increasing the robustness? Revisit AVSBench dataset.

Q3. How well the proposed model perform and the impact bring by each proposed elements?

Aims and Objective

Aim: Modeling audio-visual segmentation problem using U-Net architecture and attentive feature fusion module.

Objective 1 (O1): Designing a method to data augmentation.

Objective 2 (O2): Careful designing an objective function to avoid models being prone to overfitting.

Objective 3 (O3): Providing qualitative and quantitative analysis.

The research question **Q1**, **Q2** and **Q3** are addressed with **O1**, **O2** and **O3** respectively.

四 研究方案

拟采用的研究方法、技术路线、实验方案及可行性分析。

Related Works

In this section, we firstly take a look at related tasks from single modularity perspective, e.g. image (semantic) segmentation and some audio-related tasks. Second, we introduce and discuss the merit and importance of multimodality in generally.

Image (Semantic) Segmentation

Image segmentation and semantic segmentation are both important tasks in computer vision that involve dividing an image into segments and assigning meaningful labels to those segments. However, they serve different purposes and have distinct characteristics. Image segmentation is a general term for the process of partitioning an image into multiple segments, each representing a coherent object or region of interest. It doesn't necessarily involve assigning semantic labels to those segments. An example of image segmentation is when an image is divided into segments based on color, texture, or other visual properties. Semantic segmentation, on the other hand, goes a step further by assigning each pixel in an image a meaningful class label. The goal is to label every pixel with the corresponding object or category it belongs to. This type of segmentation is used to create detailed and accurate maps of objects within an image. For instance, in a street scene, semantic segmentation would label each pixel as road, sidewalk, building, car, tree, etc., providing a pixel-level understanding of the scene.

One commonly used architecture for image segmentation, including semantic segmentation, is the Fully Convolutional Network (FCN), which was introduced by Long et al. in "Fully Convolutional Networks for Semantic Segmentation" [26]. FCNs use convolutional layers to generate dense pixel-wise predictions, enabling them to handle images of arbitrary sizes. Another popular architecture is the U-Net, proposed by Ronneberger et al. in "U-Net: Convolutional Networks for Biomedical Image Segmentation" [37]. U-Net consists of a contracting path to capture context and a symmetric expanding path to enable precise localization. In recent years, deep learning architectures based on encoder-decoder structures with skip connections have been commonly used. These include architectures like SegNet [27] and DeepLab [28].

For backbones (or encoders), architectures pre-trained on large-scale image classification datasets, such as ResNet [44], ResNeXt [29], and EfficientNet [30], are often used as a starting

point. These pre-trained models are then fine-tuned on segmentation tasks to leverage the feature extraction capabilities they have learned.

Audio-Related Tasks

Audio-related tasks like audio classification and audio event detection are vital in analyzing and understanding audio signals. Just as in computer vision where deep learning architectures revolutionized image tasks, similar advancements have been made in audio tasks using neural networks.

Audio Classification involves categorizing audio clips into predefined classes, such as identifying the genre of music or recognizing spoken words. One popular architecture for audio classification is the Convolutional Neural Network (CNN), which has shown success when spectrogram representations of audio are treated as images. Hershey et al. introduced the use of CNNs for sound classification in "Cnn architectures for large-scale audio classification" [38]. Transfer learning using pre-trained models like VGG or ResNet on large-scale audio datasets like AudioSet has also proven effective [31].

Audio Event Detection is concerned with identifying specific sound events within audio streams, such as detecting the sound of a dog barking or a doorbell ringing. Recurrent Neural Networks (RNNs) and their variants, like Long Short-Term Memory (LSTM) networks, are widely used due to their ability to model sequential dependencies in audio data.

Relating to image encoders, it is worth noting that audio data can be represented as spectrograms, analogous to images. Spectrograms capture the frequency content of audio signals over time, which makes them amenable to processing by convolutional neural networks (CNNs), similar to how CNNs are used for image analysis. This approach has been effectively utilized for both audio classification and audio event detection tasks.

What is Multimodality

Multimodality refers to the integration of information from multiple distinct data modalities, such as text, images, audio, or sensor data, to improve the performance of various tasks. The benefits of multimodality are manifold. By fusing information from diverse sources, models can leverage complementary strengths, leading to more robust and accurate predictions. This approach enhances the overall understanding of data, capturing nuances that might be missed

by analyzing individual modalities. Additionally, multimodal models often exhibit improved generalization, as they can learn from broader contextual cues.

Several tasks lend themselves naturally to the concept of multimodality. Multimodal Sentiment Analysis involves analyzing text, images, and audio to understand the emotions expressed by individuals, leading to a more comprehensive understanding of sentiment. Visual Question Answering (VQA) combines image and text data to answer questions about the content of images. Audio-Visual Speech Recognition merges lip movement visual information from videos with corresponding audio to enhance speech recognition accuracy, especially in noisy environments. Multimodal Translation involves translating text between languages while taking into account corresponding images or other visual context. Human-Computer Interaction tasks can leverage multimodality to enable more intuitive communication between humans and machines, such as using gestures, voice, and touch together.

In scientific research, multimodality can aid in Medical Diagnostics, where combining medical images with patient data can lead to more accurate disease detection. Autonomous Driving systems benefit from fusing data from sensors like cameras, LIDAR, and radar for a holistic perception of the environment. In Social Media Analysis, multimodal models can analyze text, images, and videos to better understand trends, user behavior, and interactions.

Overall, multimodality harnesses the power of diverse data sources to tackle complex tasks, leading to richer insights and enhanced performance across a wide range of applications.

The size and class distribution of a dataset act as a main factor in determining the robustness of a trained model towards unseen data. Hence, there are researches made in this particular issue, including to devise a better resample techniques in the imbalanced dataset or to devise a better objective loss to promote robustness. The follows introduce some of the works.

Class-Imbalanced Dataset

When confronted with a class-imbalanced dataset, where certain classes have a substantially smaller representation than others, several concerted efforts can be made to rectify the imbalance and ameliorate model performance. One fundamental approach involves leveraging various data augmentation techniques to artificially generate additional instances for the underrepresented classes, thus leveling the playing field in terms of class distribution. Moreover, tactics like oversampling, wherein existing samples from minority classes are duplicated, or

under-sampling, which involves reducing instances from the overrepresented classes, can be employed to achieve a more equitable distribution of classes.

In the realm of loss functions, the focal loss [17] emerges as a powerful tool for addressing class imbalance. This specialized loss function serves to down-weigh easily classifiable examples, thereby intensifying the focus on challenging instances and consequently mitigating the undue influence of dominant classes during training. Another effective strategy involves the utilization of class-weighted loss [[18], [19]], whereby greater weights are assigned to minority classes during training, effectively affording them heightened significance in the learning process. For segmentation tasks, the Dice loss [20] gains prominence, considering the pixel-level agreement between predicted and ground truth masks, imparting distinct penalties for false negatives and false positives. Expanding beyond loss functions, the integration of ensemble methods, the judicious application of transfer learning [21], and the crafting of tailored neural architectures can further contribute to effectively managing class imbalance and elevating model performance across all classes.

Small-Sized Dataset

Addressing challenges posed by small-sized datasets requires a strategic approach to ensure effective model training and generalization. Researchers have proposed several techniques to combat the limitations of small datasets. Data augmentation methods, such as rotation, flipping, and cropping, can artificially expand the dataset, enhancing model performance. Bayesian approaches, like Monte Carlo Dropout [22], help capture model uncertainty in predictions, crucial in scenarios with limited data. Transfer learning, exemplified by [23], leverages pre-trained models on larger datasets to extract valuable features, which are then fine-tuned on the smaller dataset. Meta-learning methods, like [24], enable models to quickly adapt to new tasks with limited data by learning from a variety of similar tasks. Ensemble techniques, such as [25], combine predictions from multiple models trained on different small datasets, mitigating the risk of overfitting. These strategies collectively empower models to effectively learn from scarce data, minimizing the challenges of small-sized datasets.

Mathematical Modelling

Given some video frames $I \in \mathbb{R}^{N\times C\times H\times W}$ and their corresponded cropped audio clips $A \in \mathbb{R}^{N\times L}$, it is required to design a model that output mask $M \in \mathbb{R}^{N\times K\times H\times W}$, where *N* is the number of frames, *C*, *H* and *W* is the dimension of the image, *L* is the number of sampled audio data dependent on its sample rate and *K* is the number of classes. In case of AVS, *K* is set as 1. Follow the previous works as mentioned in section above, audio clips are transformed into mel-spectrogram $A_{mono} \in \mathbb{R}^{N\times F\times T}$, where *T* and *F* is the number mel bands and frames respectively, as describe in *Eq. 1*.

Eq. 1: The transformation from a raw audio waveform to a mel spectrogram

---

1. Frame the Audio:
   - Divide the audio waveform into overlapping frames: $x(t) = x[n] \cdot w(t - nT_s)$
2. Calculate the Short-Time Fourier Transform:
   - Compute the Discrete Fourier Transform (DFT) for each frame: $X[k,m] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi kn/N}$
3. Calculate Power Spectrum:
   - Compute the magnitude squared of the DFT: $S[k,m] = |X[k,m]|^2$
4. Apply Mel Filterbanks:
   - Apply the mel filterbank for each frame and each filter: $E[m,i] = \sum_{k=0}^{N/2} H[i,k] \cdot S[k,m]$
   - $H[i,k]$ is the value of the *i*-th mel filter at frequency index *k*.
5. Log Compression:
   - Apply the logarithm to the filterbank energies: $M[m,i] = \log(E[m,i] + \epsilon)$
   - ϵ is a small constant to prevent taking the logarithm of zero.

Visual cues $V_{fea} \in \mathbb{R}^{N\times c\times h\times w}$ and audio cues $A_{fea} \in \mathbb{R}^{N\times d}$ are extracted by passing both inputs into image encoder $f(\cdot)$ and audio encoder $g(\cdot)$, where *c* is the number of output channels and *d* is the audio feature dimension. Considering the dimensionality of both cues does not align well, A_{fea} should be expanded as V_{fea} before the fusion of two cues by a fusion mechanism $h(\ ,\ )$ to produce enhanced visual cues $\widehat{V_{fea}}$. Finally, decoder $k(\cdot)$ decodes and upsamples $\widehat{V_{fea}}$ to output *M*. To conclude:

$$V_{fea} = f(I),\ A_{fea} = g(A_{mono}) \tag{2}$$

$$M = k\left(\widehat{V_{fea}}\right) = k \cdot h\left(V_{fea}, A_{fea}\right) \tag{3}$$

Reference

[17] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[18] Akil, M., Saouli, R., & Kachouri, R. (2020). Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. Medical image analysis, 63, 101692.

[19] Furtado, P. (2021, June). Experiments with loss function for improvement of multi-class segmentation of diabetic retinopathy lesions. In Thirteenth International Conference on Digital Image Processing (ICDIP 2021) (Vol. 11878, pp. 178-186). SPIE.

[20] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA).

[21] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[22] Gal, Y., & Ghahramani, Z. (2015). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. ArXiv. /abs/1506.02142

[23] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR).

[24] Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ArXiv. /abs/1703.03400

[25] Xie, Q., Luong, M., Hovy, E., & Le, Q. V. (2019). Self-training with Noisy Student improves ImageNet classification. ArXiv. /abs/1911.04252

[26] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[27] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[28] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[29] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[30] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. International Conference on Machine Learning (ICML).

[31] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Transfer learning for music classification and regression tasks. ArXiv. /abs/1703.09179

Methodologies

As an overview of our proposed method, to tackle the AVS problem, we build an image classifier that identifies key objects in visual frame and an audio event classifier that retrieve informative audio cues out from audio data. The two classifiers are linked at certain point, enabling the routing of the data containing the prediction logits and the weak spatial information from the later to the former. More specifically, we propose a UNet [37] -like architectural multimodal model with attentive feature fusion module. The model encodes audio and frames into semantic features, merge the features via the fusion module, then to pixel-wise mask *w.r.t.* the sounding object. The proposed base model is presented in *Fig. 1*. Unlike the single skip connections in UNet, we formulated double skip connections with ASPP and GAB module. Furthermore, our model leverages deep supervision [68] to generate mask predictions of multiple scales, which are utilized for loss function and serve as one of the inputs to GAB.

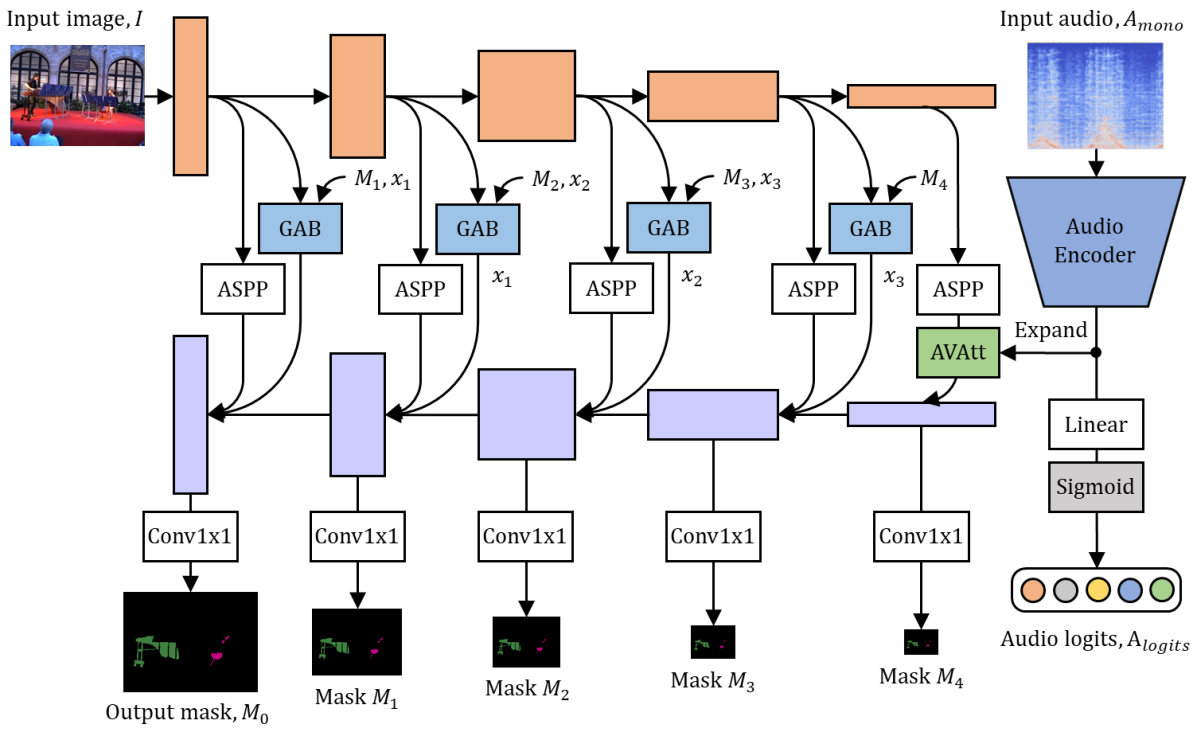

Fig. 1: The proposed base model

From a high level-perspective, the system primarily consists of two input subnetwork, *viz.*, the *audio feature encoder* and *visual subnetwork*. We deploy a multimodal fusion module, AVAtt to merge audio and visual features. We apply the technique of transfer learning by using the pre-trained weights as initial weights to our model.

Audio Feature Encoder

This is a pre-trained VGGish [38] trained on AudioSet [39]. VGGish is a convolutional neural network (CNN), which is a variant of the VGG [64] model. The encoder accept log mel spectrogram audio inputs $A_{mono} \in \mathbb{R}^{N\times F\times T}$, and outputs audio features $A_s \in \mathbb{R}^{N\times D}$, where D = 256 is the feature dimension. VGGish consists only four groups of convolution-maxpool layers instead of five by dropping the last group of convolutional and maxpool layers. The 1000-wide fully connected layer at the end is replaced by a 256-wide fully connected layer. This acts as a compact embedding layer. In case of AVSS setting, for probability prediction, the model apply a linear classifier, followed by a sigmoid activation on top of A_S:

$$P_S = sigmoid\big(\mathrm{g}_1(A_S)\big) \tag{4}$$

where $g_1 \in \mathbb{R}^{N\times K}$ is a linear classifier, K is the number of classes.

Visual Subnetwork

The subnetwork is a UNet [32] -like architectural CNN, made up by the *image feature encoder* and the *decoder*, which is an architectural that is commonly used in many image segmentation tasks [[48]-[53]]. Down through the encoder, spatial information is progressively downsamples while feature information is increased, capturing the global image context, and produces hierarchical visual feature maps during the encoding process. We denote the features as $F_i \in \mathbb{R}^{N\times C_i\times h_i\times w_i}$, where h_i, w_i is the dimension of the output features at stage *i*, depends on specific backbone.

The network incorporates a pair of skip connections, which consists of ASPP and GAB modules respectively, between the encoder and decoder. Atrous Spatial Pyramid Pooling (ASPP) modules [69] post-process the visual features F_i to $V_i \in \mathbb{R}^{N\times C\times h_i\times w_i}$, where C = 256. These modules encourage the recognition of visual objects in different receptive fields by employ multiple parallel filters with different rates. Group Aggregation Bridge (GAB) modules [70] outputs $J_i \in \mathbb{R}^{N\times C\times h_i\times w_i}$. The module plays a crucial role in capturing contextual information from both low-level and high-level features, ultimately enhancing the model's capacity for making accurate and detailed segment predictions. The GAB module operates by taking three primary inputs: low-level features, high-level features, and a mask. Firstly, the high-level features are adjusted to match the dimensions of the low-level features using techniques like depth-wise separable convolution and bilinear interpolation. Then, both sets of features are grouped into four distinct units along the channel dimension. For each group, a low-level feature group is concatenated with a high-level feature group, creating four fused feature groups. Importantly, the corresponding mask is concatenated to each fused feature group, ensuring that the spatial information is incorporated effectively. To capture information at varying scales, dilated convolutions with different rates are applied to the distinct groups of fused features. These convolutions help extract details across different receptive fields. Subsequently, the outputs of these convolutions are concatenated together along the channel dimension, facilitating the integration of multi-scale information. To allow for seamless interaction between features at different scales, a final plain convolution with a kernel size of 1 is applied. This stage refines the captured information, contributing to the module's ability to produce comprehensive and refined predictions.

At *j*-th stage, where *j* = 1, 2, 3, 4, the outputs from stage $V_{5\text{-}j}$, $J_{5\text{-}j}$ of the encoder and the last stage $Z_{6\text{-}j}$ are utilized for decoding process.

$$Z_{6-j} = \begin{cases} AVAtt(V_5, A_S), & j = 0 \\ g_2\left(concat\left(V_{5-j},\ J_{5-j},\ Z_{6-j}\right)\right), & j > 0 \end{cases} \quad (5)$$

where $g_2 \in \mathbb{R}^{C_1 \times C}$ is a 1×1 channel-wise convolution, C_1 = 768 and C = 256. The decoded features are then upsampled to the next stage. The output of the decoder is a list of mask predictions of multiple scales, $M_i \in \mathbb{R}^{N \times h_i \times w_i}$. M_i is scaled up bilinearly to match the size of ground truth mask M_{true}.

Audio-Visual Attentive Feature Fusion (AVAtt)

Till this point, $A_v = V_5$ contains abundant visual context information that is essential for image segmentation, while A_S provides audio categorical information towards what objects are detected given the input audio. The aim is to merge the feature information from both audio and image encoder, where the audio feature act as an auxiliary guidance to the visual feature so that the predicted segmentation mask agrees with the sounding object in the frame. As a result, we apply multi-head attention mechanism to fuse them while reducing computation cost as possible.

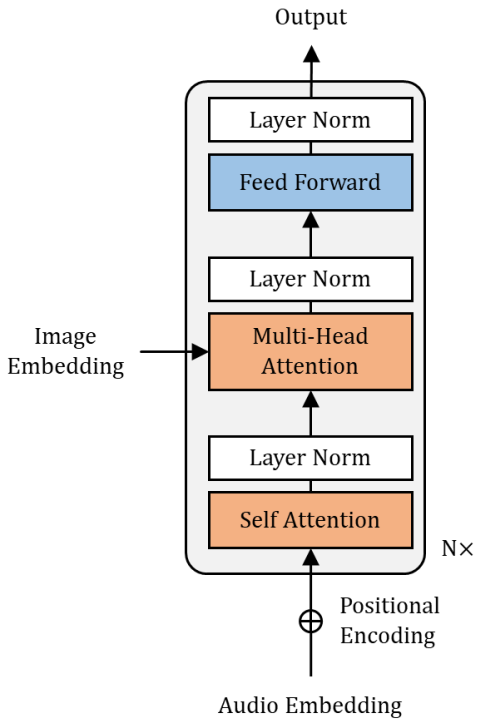

Fig. 2: Audio-Visual Attentive Feature Fusion.

Specifically, the multi-modal fusion module consists of a stack of L = 3 Transformer Decoder layers [47], as depicted in *Fig. 3*. To align both audio and visual features, A_S is expanded to match the dimensions of A_v: $A_v \in \mathbb{R}^{N \times D \to N \times 1 \times D \to N \times C_i \times D}$. In i-th layer, AVAtt takes visual

features $A_v \in \mathbb{R}^{N\times W\times H}$ and audio features $A_s \in \mathbb{R}^{N\times C_i\times D}$ as input, and outputs new improved visual features $\widehat{A_v}$, by following the procedure as stated below.

First, AVAtt uses Self Attention (SA) and Multi-head Attention (MHA) to find sounding regions in A_V by A_S:

$$\widehat{A_S} = \mathrm{LN}\big(\mathrm{SA}\big(Q_{A_S}, K_{A_S}, V_{A_S}\big) + A_S\big) \tag{6}$$

$$\widehat{A_v} = \mathrm{LN}\left(\mathrm{MLP}\left(\mathrm{LN}\big(\mathrm{MHA}\big(Q_{\widehat{A_S}}, K_{A_v}, V_{A_v}\big) + A_v\big)\right)\right) \tag{7}$$

where LN denotes layer normalization [67]. SA and MHA can be formulated as:

$$\mathrm{SA}(Q_A, K_A, V_A) = A + V_A^{\mathrm{T}} \times softmax\big(K_A^{\mathrm{T}} \times Q_A\big) \tag{8}$$

while MHA is an extension of this mechanism by using multiple parallel heads to capture different patterns of relationships within the input sequence.

Objective Function

The primary loss function as for AVS task is a multi-staged binary-cross-entropy-Dice (BCE-Dice) loss, while as for AVSS task we select cross-entropy-Dice (CE-Dice) loss instead, where the function at stage-*i* is as formulated as:

$$L_i = \begin{cases} BCEDice(M_{true}, M_i), & if\ \ AVS \\ CEDice(M_{true}, M_i), & if\ \ AVSS \end{cases} \tag{9}$$

where $M_i \in \mathbb{R}^{N\times H\times W}$ is the predicted segmentation map at stage-*i*, $M_{true} \in \mathbb{R}^{N\times H\times W}$ is the ground truth label, and *BCEDice*, *CEDice* are simply the summation of two losses:

$$BCEDice(y, \hat{y}) = BCE(y, \hat{y}) + Dice(y, \hat{y}) \tag{10}$$

$$CEDice(y, \hat{y}) = CE(y, \hat{y}) + Dice(y, \hat{y}) \tag{11}$$

Combined loss like BCE-Dice loss and CE-Dice loss is selected because it is expected to bring more robustness, as reported in [[71]-[74]]. It is used as the main supervision function and measures the pixel-wise difference between the predicted segmentation map and the ground truth label.

To point out, the ground truth M_{true} contains pixel-wise annotations *w.r.t.* the sounding objects, and not all the visible objects M_{all} in the frame. In common cases, there will be other objects in the image that does not sounds. By a relationship where M_{true} is a subset of M_{all}, some information is wasted when training. We try to turn the wasted information into a useful

one by designing a loss function named *Patch Classification Loss* that encourage the model to "group" pixels (where having a high chance they belong to a particular object) together. The objective of the loss is to divide an input image into multiple candidate groups, where every pixel-wise probabilities in each group should each having a same value respectively. This further promote the model to learn to group the visual representation and to uplift the confidence around the segmentation boundaries. The mathematical notation of *Patch Classification Loss* is shown below.

$$L_{patch}(M_0, M_{sam}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{m}\sum_{j=1}^{m} KL\left(\log\ softmax(P_{ij}) \middle| softmax\left(\frac{1}{n_j}\sum_{k=1}^{n_j} P_{ijk}\right)\right) \quad (12)$$

where *N* represents the total number of batches, *m* represents the number of groups in each batch, and n_j represents the number of elements in groups *j*. We generate the segmentation masks by pretrained ViT-L Segment Anything Model (SAM) [45] with settings of Automatic Mask Generator. It is a promptable segmentation model designed to segment any object by zero-shot transfer to unseen image distributions, and can be used for prompt-based segmentation tasks such as segmenting objects from a point, bounding box, and text prompt.

An additional regularization term called audio-visual mapping (AVM) loss as introduces in [14] is added in AVS task to enforce the correlation between audio and visual signals and ensure that the masked visual features have similar distributions with the corresponding audio features. Specifically, AVM loss is computed using the Kullback-Leibler (KL) divergence between the average pooled visual features and the corresponding audio features, as formula:

$$L_1 = L_{AVM} = sum\left(KL\left(avg\left(M_p \odot Z_p\right), A_p\right)\right) \quad (13)$$

where M_p is the ground truth label, Z_p is the visual feature map and A_p is the audio feature map of *p*-th pixel.

In place of AVM, we introduced a multi-label classification (BCE) loss on the audio logits A_{logits}, as we noticed AVM contributes relatively little effort in improving evaluation results as for AVSS task. Ground truth class label are retrieved from the ground truth mask, encouraging the class-label-guided audio encoder share its learnt semantic cues with the visual subnetwork.

$$L_1 = L_{BCE}\left(A_{logits}, A_{true}\right) \quad (14)$$

Deep supervision [75] is employed to calculate the loss function for different stages, in order to generate more accurate mask information. Hence, the total objective function is computed as the weighted sum of three losses, as follow:

$$L = \sum_{i=0}^{S} \lambda_i L_i + \lambda_1 L_1 + \lambda_2 L_{patch} \quad (15)$$

where $\lambda_i, \lambda_1, \lambda_2$ are constant weights. In this paper, we set λ_i to 1, 0.5, 0.4, 0.3, ... from $i = 0$ to $i = S$. Considering the expensive computation cost when training with multiple loss functions, as a balance, we set $S = 4$ and $S = 2$ in case of AVS and AVSS task respectively.

Training

To assess the efficacy of our model, we trained on AVSBench and AVSS datasets. For image encoder, three different backbones are selected: ResNet34, ResNet50 and SwinV2. Unlike the implementations of other previous works [[14], [15], [65], [66]], we firstly initialize the both encoders with their default pretrained weights (ImageNet1k for ResNet, SwinV2; and AudioSet [39] for VGGish), then we pre-train the models on all S4, MS3 and AVSS subset as a whole, using AdamW optimizer [76] with both initial learning rate and weight decay of 1e-4, $\lambda_1 = 0.25$ and $\lambda_2 = 0.1$. The training process is set to run on 20 epochs, with batch size of 16. ReduceLROnPlateau schedule is used to dynamically reduce the learning rate by factor of 0.8 with patience of 1, stepped per epoch. The pretraining process is terminated by early stop. After that, finetuning on S4 and MS3 subset separately, with initial learning rate and weight decay of 1e-6 and 1e-3 respectively, for another 20 epochs. The hardware and software configurations used during the model training and evaluation are outlined below:

Table 1: Hardware specification

GPU	NVidia Tesla A100-SXM4-40GB
CUDA Version	12.0
GPU Memory	40 GB
CPU	Intel Xeon CPU*2
AI Performance	19.5 TFLOPS (FP32)

Table 2: Software configuration

Operating System	Ubuntu 20.04.6 LTS
Deep Learning Framework	PyTorch 2.0.0
Python Version	3.8.10

Evaluation Results

To evaluate our method, we employ Mean Intersection over Union (mIoU) and F-score as metrics. We compare the results of our method with the baseline on all S4, MS3 and AVSS settings in *Table 3*.

Table 3: Evaluation Results

Metric	Setting	Baseline		Ours			
		ResNet50	PVT-v2	ResNet34	ResNet50	ResNet101	SwinV2-b
mIoU	S4	72.79	78.74	76.56	73.25	77.15	**80.53**
	MS3	47.88	54.00	55.90	59.00	58.68	**66.66**
	AVSS	20.18	29.77	30.97		50.30	**54.58**
F-score	S4	.848	.879	.827	.815	.869	.875
	MS3	.578	.645	.590	.593	.644	**.683**
	AVSS	.252	.352	.432		.616	**.657**
Trainable Params (M)		~91	~101	~115	~153	~167	~237

Audio-Visual Dataset

The dataset contains ~11,000 5 or 10 seconds videos over 71 class labels (including background as a single class) represented as semantic maps. The videos are further trimmed into roughly 1 second per sample.

The AVSBench-v2 dataset is divided into three subsets. In the first subset, there is a single sound source in the video, leading to the task call semi-supervised Single Sound Source Segmentation (S4). In the second subset, there are multiple sound sources, leading to the task of

fully-supervised Multiple Sound Source Segmentation (MS3). For these two subsets, the ground truths are binary masks indicating pixels emitting the sounds.

The third subset, Audio-Visual Semantic Segmentation (AVSS) is a semantic-labels subset that introduces semantic labels of the sounding objects, exploring by fully-supervised. Compared to the S4 and MS3 settings, AVSS requires generating semantic maps that further tell the category information of the masked sounding objects.

Data Preprocessing

Image frames was sampled for raw video by ~1.04 fps. The images are first resized to 224×224 then apply normalization using mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225) as common practice, by applying the formula: $I = (I - \mu \cdot 255)/(\sigma \cdot 255)$.

The raw audio *x* is preprocessed as follow: (1) All audio is resampled to 16 kHz, which is sufficient for audio processing [[40], [41]], and mono-channel $x_{mono}(t)$, depends on the individual input audio, if the audio is binaural, we create the monaural audio $x_{mono}(t)$ by mixing $\{x_{L(t)}, x_{R(t)}\}$ as $x_{mono}(t) = (x_L(t) + x_R(t))/2$. (2) Compute spectrogram x_{spec} using magnitudes of the Short-Time Fourier Transform (STFT) with a window size of $25ms$, a window hop of $10ms$, and a periodic Hann window. (3) Compute mel spectrogram x_{mel} by mapping the spectrogram to 64 mel bins covering the range 125-7500 Hz. (4) Compute stabilized log mel spectrogram $X_{mono} \in \mathbb{C}^{96\times64}$ by applying $log(x_{mel} + 0.01)$ where the offset 0.01 is used to avoid taking a logarithm of zero. These features are then framed into non-overlapping examples of 0.96 seconds, where each example covers 64 mel bands and 96 frames of $10ms$ each. Mel spectrogram is widely used in general audios deep learning [[42], [43]], enabling it to be processed by an image classification neural network.

Data Augmentation

Data augmentation is a technique to increase the size of the dataset by applying various transformations to the original data. It helps the deep learning model generalize better, especially when the original dataset size is too small. We apply several transformations to both image and audio files in the original data. The specific transformations applied are outlined in *Table 4*. We apply 20 random combinations of transformations to each sample. All the transformations are made in a very small amount to prevent the potential removal of key features render the augmented data useless.

Table 4: Data augmentation

Type	Transformation	Description
(a) Audio	Gaussian noise	Add noise to the samples
	Gain	Increase/decrease the audio volume
	Gain transition	Gradual change in volume over a specific time span
	Loudness normalization	Apply a constant amount of Gain
	Pitch shift	Increase/Decrease the pitch
	Resample	Resample the audio
	Time stretch	Increase/decrease the audio speed
(b) Image	Blur	Blur the image
	Brightness and contrast	Apply changes in brightness and contrast
	Gaussian noise	Overlay random noise

Reference

[32] Wang, L., Luc, P., Recasens, A., Alayrac, J., & Oord, A. V. (2021). Multimodal Self-Supervised Learning of General Audio Representations. *arXiv*. https://doi.org/https://arxiv.org/abs/2104.12807v2

[33] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv*. https://doi.org/https://arxiv.org/abs/2103.00020v1

[34] Chen H, Xie W, Vedaldi A, Zisserman A.: Vgg-Sound: A large-scale audio-visual dataset. In: ICASSP, 2020

[35] Xu C, Kang J.: Soundscape evaluation: Binaural or monaural?. In: J Acoust Soc Am, 2019; 145 (5): 3208–3217

[36] Butler, R. A., Humanski, R. A., & Musicant, A. D. (1990). Binaural and Monaural Localization of Sound in Two-Dimensional Space. *Perception*. https://doi.org/10.1068/p190241

[37] Ronneberger O, Fischer P, Brox T. UNet: Convolutional networks for biomedical image segmentation[A]. In: International Conference on Medical image computing and computer-assisted intervention[C]. Springer, Cham, 2015. 234-241.

[38] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: CNN architectures for large-scale audio

classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 131–135 (2017)

[39] Gemmeke J F, Ellis D P, Freedman D, Jansen A, Lawrence W, Moore R C, Plakal M, Ritter M.: Audioset: An ontology and human-labeled dataset for audio events. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)

[40] Alex J S R, Kumar M A, Swathy D V.: Deep Learning Approaches for Fall Detection Using Acoustic Information. In: Advances in Smart Grid Technology. Springer, 479–488. 2021

[41] Baevski A, Zhou Y, Mohamed A, Auli M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 12449-12460, 2020

[42] Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley M D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 28, 2880–2894 (2020)

[43] Gong Y, Chung Y A, and Glass J.: Ast: Audio spectrogram transformer. In: arXiv preprint arXiv:2104.01778 (2021)

[44] He K, Zhang X, Ren S, Sun J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016)

[45] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., & Girshick, R. (2023). Segment Anything. ArXiv. /abs/2304.02643

[46] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database[A]. In: 2009 IEEE conference on computer vision and pattern recognition[C]. IEEE, 2009. 248-255.

[47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I..: Attention Is All You Need. *arXiv*. https://doi.org/https://arxiv.org/abs/1706.03762v5. (2017)

[48] Kohl S A, Meyer C, De Fauw J, Ledsam J R, Eslami S M, Rezende D J, Ronneberger O.: A Probabilistic UNet for Segmentation of Ambiguous Images. *arXiv*. https://doi.org/https://arxiv.org/abs/1806.05034v4. (2018)

[49] Ronneberger O, Fischer P, Brox T. UNet: Convolutional networks for biomedical image segmentation[A]. In: International Conference on Medical image computing and computer-assisted intervention[C]. Springer, Cham, 2015. 234-241.

[50] Men K, Dai J, Li Y.: Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. In: Med. Phys. 44, pp. 6377–6389 (2017)

[51] Wang J Z, Lu J Y, Qin G, Shen L J, Sun Y Q, Ying H M, Zhang Z, Hu W G.: Technical Note: A deep learning-based autosegmentation of rectal tumors in MR images. In: Med. Phys. 45, pp. 2560–2564 (2018)

[52] Neven R, Goedemé T.: A Multi-Branch UNet for Steel Surface Defect Type and Severity Segmentation. In: Metals. 11. 870. (2021)

[53] Zhang R, Du L, Xiao Q, Liu J.: Comparison of Backbones for Semantic Segmentation Network. In: J. Phys.: Conf. Ser. 1544 (2020)

[54] Carreira J, Zisserman A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proc. CVPR (2017)

[55] Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T, Ronneberger O.: 3D UNet: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv*. https://doi.org/https://arxiv.org/abs/1606.06650v1. (2016)

[56] Wang S, Wang K, Yang T, Li Y, Fan D.: Improved 3D-ResNet sign language recognition algorithm with enhanced hand features. In: Sci Rep 12, 17812 (2022)

[57] Wang, T., Liu, M., Zhu, J., Liu, G., Tao, A., Kautz, J., & Catanzaro, B. (2018). Video-to-Video Synthesis. *arXiv*. https://doi.org/https://arxiv.org/abs/1808.06601v2

[58] Torfi, A., Iranmanesh, S. M., Nasrabadi, N. M., & Dawson, J. (2017). 3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition. arXiv. https://doi.org/10.1109/ACCESS.2017.2761539

[59] Jordan M I.: Serial order: a parallel distributed processing approach. Technical report. United States, (1986)

[60] Sutskever I, Vinyals O, Le Q V: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, 3104-3112 (2014)

[61] Garg, A., Noyola, J. and Bagadia, S., 2016. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report.

[62] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*. https://doi.org/https://arxiv.org/abs/2010.11929v2

[63] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C. (2021). ViViT: A Video Vision Transformer. *arXiv*. https://doi.org/https://arxiv.org/abs/2103.15691v2

[64] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv. https://doi.org/https://arxiv.org/abs/1409.1556v6

[65] Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., & Carneiro, G. (2023). A Closer Look at Audio-Visual Semantic Segmentation. ArXiv. /abs/2304.02970

[66] Ling, Y., Li, Y., Gan, Z., Zhang, J., Chi, M., & Wang, Y. (2023). Hear to Segment: Unmixing the Audio to Guide the Semantic Segmentation. ArXiv. /abs/2305.07223

[67] Jimmy Lei Ba, Jamie Ryan Kiros, & Geoffrey E Hinton. (2016). Layer normalization. arXiv preprint arXiv:1607.06450

[68] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer.

[69] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A.L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4), 834–848.

[70] Ruan, J., Xie, M., Gao, J., Liu, T., & Fu, Y. (2023). EGE-UNet: An Efficient Group Enhanced UNet for skin lesion segmentation. ArXiv. /abs/2307.08473

[71] Rajput, V. (2021). Robustness of different loss functions and their impact on networks learning capability. ArXiv. /abs/2110.08322

[72] L. Yang, C. You (2018). Instance-U-Net and Watershed: Improved Segmentations for breast cancer cells, Fall 2018 CS230 Standford, student paper.

[73] L. Zhou, C. Zhang and M. Wu. (2018). D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018, pp. 192-1924, doi: 10.1109/CVPRW.2018.00034.

[74] Yeung M, Sala E, Schönlieb CB, Rundo L. (2022) Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput Med Imaging Graph. 2022 Jan;95:102026. doi: 10.1016/j.compmedimag.2021.102026.

[75] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. (2015) Deeply-supervised nets. In: Artificial Intelligence and Statistics, pages 562–570, 2015.

[76] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

五 研究工作进度安排

理论研究：应包括文献调研，理论推导，数值计算，理论分析，撰写论文等；实验研究和工程技术研究：应包括文献调研，理论分析，实验设计，仪器设备的研制和调试，实验操作，实验数据的分析处理，撰写论文等。

The research work is planned to be completed in 7 months, from September 2023 to March 2024.

The research work schedule is shown in *Table 5*.

Table 5 Work schedule

Time	Task
July – October 10, 2023	Literature research, theoretical analysis
October 11 – December, 2023	Experimental design
January 3 – 6, 2024	Experimental operations, debugging
January – March 2024	Summarize the research results, write thesis, prepare for defense

六 预期研究成果

详细说明预期研究取得的成果，包括但不限于新理论、新方法、新技术以及新装置、新方案等, 以及预期发表的论文、申请的专利、参加的学术会议等。

Publish a journal paper.

七 本课题创新之处

论证说明研究内容、拟采用的研究方法、技术路线或预期成果中有哪些创新之处。

Publish a journal paper.

八 研究基础

1.与本项目有关的研究工作积累和已取得的研究工作成绩。

2.已具备的实验条件，尚缺少的实验条件和解决的途径（包括利用国家重点实验室和部门开放实验室的计划与落实情况）。

3. 研究经费预算计划和落实情况。

北京理工大学

留学硕士学位论文开题报告——导师意见

学籍状态	□ 注册 □ 暂缓注册
导师对开题报告的审阅意见： 导师 签字： 年 月 日	

说明： 本页全部由指导教师填写。

北京理工大學

留学硕士学位论文开题报告评审表

<table>
<tr><td>学号</td><td></td><td>姓名</td><td></td><td>导师姓名</td><td></td></tr>
<tr><td>所在学院</td><td colspan="2"></td><td>学科、专业</td><td colspan="2"></td></tr>
<tr><td>课程学习情况</td><td colspan="2">已修课程学分</td><td></td><td>待修课程学分</td><td></td></tr>
<tr><td>选题名称</td><td colspan="5"></td></tr>
<tr><td>课题经费来源</td><td colspan="5"></td></tr>
<tr><td>开题报告时间</td><td colspan="5"></td></tr>
<tr><td rowspan="6">评审组成员</td><td></td><td>姓　名</td><td>职　称</td><td colspan="2">工作单位</td><td>签　字</td></tr>
<tr><td>组长</td><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td rowspan="4">组员</td><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td colspan="6">评审组意见：

要求针对论文选题的背景、意义以及拟定研究内容的深度、完整性，拟采取研究方案/方法的可行性，研究进度安排的合理性等给出明确的意见。对后续研究工作中可能遇到的难点等给出具体的建议。

给出是否同意开题的明确意见。

组长签字：　　年　　月　　日</td></tr>
<tr><td colspan="6">学院审核意见：

主管院长签字：　　年　　月　　日</td></tr>
</table>