A Lightweight Dual-Branch Visual Front-End for Efficient Audio-Visual Target Speaker Extraction

Zhan Jin[1,2], Peijun Yang[3,4], Juan Liu[3,4], and Ming Li[2,3✉⋆]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] School of Artificial Intelligence, Chinese University of Hong Kong, Shenzhen, China
[3] School of Artificial Intelligence, Wuhan University, Wuhan, China
[4] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
mingli369@cuhk.edu.cn

Abstract. Audio-Visual Target Speaker Extraction (AVTSE) uses face-related visual cues to extract a target speaker from a speech mixture. In many AVTSE systems, visual features are obtained from convolutional image encoders, and their computation may become a non-negligible part of the overall system when lightweight separation backbones are used. This work proposes a dual-branch visual front-end that reorganizes visual information according to its spatial characteristics. A compact lip region of interest is used to model local articulation cues through convolutional processing, while facial landmark coordinates are used to represent global geometric motion with a lightweight temporal convolutional network. The two representations are integrated into IIANet and SpatialNet with limited fusion overhead. Experiments on LRS3-2mix under a unified reduced training protocol show that the proposed method obtains a performance–computation trade-off close to the lightweight image-based baseline while reducing the module-level computation of the visual front-end. Compared with the ResNet18-based visual front-end evaluated under the same protocol, the proposed front-end reduces the additional computation inside the separation model by more than 93% on both evaluated backbones, with a moderate performance gap. These results suggest that low-dimensional geometric representations can complement compact local appearance modeling for efficient AVTSE systems. The reported computation does not include face detection or landmark estimation; therefore, the efficiency conclusion applies to settings where landmark coordinates are already available or can be obtained as part of the preprocessing pipeline. Demos are available at https://anonymous.4open.science/r/Lightweight-Visual-Frontend-E46E.

Keywords: Audio-visual speech separation · Target speaker extraction · Visual front-end · Computational cost.

⋆ Corresponding author: Ming Li

1 Introduction

Audio-Visual Target Speaker Extraction (AVTSE) is a task that uses face-related visual cues to assist deep learning networks in extracting the target speaker's speech from a speech mixture. This task is inspired by human multimodal perception, where visual information such as lip movements and facial motions often helps improve speech understanding in complex acoustic environments. With the growing demand for audio-visual perception systems in scenarios such as complex interaction, intelligent meetings, hearing-assistive devices, and robotic audition, AVTSE has received increasing attention in recent years and has achieved continuous progress in model architecture and separation performance.

Wu et al. introduced a time-domain architecture into audio-visual speech separation and proposed an end-to-end time-domain audio-visual separation network. In two-speaker and three-speaker mixture scenarios, their method reported higher signal-to-noise ratio improvement than an audio-driven Conv-TasNet baseline [1]. Gao and Grauman proposed the VisualVoice framework, which adopts a multi-task learning strategy. While learning audio-visual speech separation from unlabeled videos, it jointly learns cross-modal speaker embeddings from lip motion and facial appearance, thereby improving generalization to real-world video scenarios [2]. Inspired by the neuroanatomical structure of mammalian cortico-thalamo-cortical circuits, Li et al. proposed CTCNet [3] which performs multi-round audio-visual fusion through hierarchical representation learning and top-down connections in the thalamic subnetworks, achieving better separation results than several existing methods on multiple benchmark datasets with fewer parameters. Pan et al. proposed SAV-TFGridNet based on the time-frequency-domain separation model TF-GridNet [4]. Their method incorporates the target speaker's facial recording as conditional information into the separation process and introduces an interference-type identification mechanism to adapt expert models to different scenarios, achieving competitive results in the COG-MHEAR audio-visual speech enhancement challenge [5]. AV-MossFormer2, as one of the core components of the ClearerVoice-Studio open-source toolkit, implements end-to-end multimodal speech processing for target speaker extraction based on an improved MossFormer2 architecture [6]. However, from the perspective of practical deployment, performance improvements are often accompanied by increased computational and storage costs, which still limits the application of AVTSE models on low-computing-power devices.

In recent years, several studies have begun to explore network architectures with lower computational complexity. Pegg et al. proposed RTFS-Net [7], which operates on complex time-frequency bins obtained from the short-time Fourier transform. It models the time and frequency dimensions separately using multi-layer recurrent neural networks and introduces an attention-based audio-visual fusion mechanism. According to the original report, RTFS-Net achieves better separation results with a smaller fraction of the parameters and multiply-accumulate operations of the baseline model. IIANet [8] further employs intra-modal attention and cross-modal attention for efficient audio-visual feature fusion. Its fast version achieves favorable separation quality with lower multiply-

accumulate operations and faster CPU inference. These methods mainly reduce the overall computational complexity from the perspective of the separation backbone. However, in most AVTSE systems, the visual front-end still accounts for a considerable proportion of image feature extraction computation. As separation backbones become increasingly lightweight, the proportion of the visual front-end in the total computation may further increase, making it an important limiting factor for low-computing-power deployment.

To address the computational cost of visual front-ends, existing studies have started to design visual front-ends with lower computational loads. Martel et al. proposed AVLIT, which adopts a progressive learning strategy and an asynchronous fully recurrent convolutional neural network. With a lightweight iterative architecture, it achieves better results than several audio-only and audio-visual baselines in noisy audio-visual speech separation tasks [9]. Li et al. proposed Dolphin, which designs DP-LipCoder, a dual-path lightweight video encoder for visual feature extraction. It converts lip motion into audio-aligned discrete semantic tokens and embeds global-local attention modules in the separation network to capture multi-scale dependencies. According to the original report, Dolphin improves separation quality while reducing the number of parameters and achieving notable GPU inference acceleration [10]. Pan et al. proposed an online audio-visual autoregressive speaker extraction model, which combines a lightweight visual front-end with an autoregressive acoustic encoder in streaming scenarios and achieves performance close to relevant baselines with reduced computational scale [11]. These studies indicate that reducing the computational cost of the visual front-end is practically meaningful for AVTSE systems. Nevertheless, as lower-computation separation backbones continue to emerge, the computational load of the visual front-end still needs to be further controlled.

In this work, we use attention heatmap visualization as a preliminary observation tool to inspect how an AVTSE model distributes attention over different locations in the facial region of interest. The visual examples suggest that the inner lip region tends to receive concentrated responses, whereas several peripheral facial regions, including the outer lip, subnasal area, and jaw, may provide more spatially sparse auxiliary cues. Based on this observation, we consider a trade-off between two visual modeling strategies. Modeling a large facial region with two-dimensional convolutions introduces computation over areas that may be weakly related to the extraction task, whereas modeling only a compact lip region may remove auxiliary facial motion. Considering that facial landmarks are commonly used for face alignment and mouth localization, and that they provide a low-dimensional geometric representation, we propose a dual-branch visual front-end. The proposed front-end uses a small lip ROI to model local appearance and articulation cues, and uses landmark sequences to model global facial dynamics. The objective is not to improve the upper bound of separation performance, but to reduce the additional computation of the visual front-end under a controlled comparison with existing visual front-ends.

2 Related Work

Visual front-ends based on pretrained lip-reading models are widely adopted in AVTSE. A typical architecture consists of a 3D convolutional layer followed by ResNet18, whose weights are usually pretrained on lip-reading datasets before being integrated into speech separation networks [13]. Although this architecture provides effective visual representations, its computational cost is non-negligible. As separation backbones become increasingly lightweight, the computation introduced by the visual front-end accounts for a growing proportion of the overall system, limiting deployment on resource-constrained platforms.

To reduce the computational burden of visual feature extraction, several lightweight visual front-ends have been proposed. BlazeNet64 adopts the efficient convolutional design of BlazeFace to reduce the computation required for visual encoding while maintaining competitive separation performance [11]. Dolphin further introduces the lightweight DP-LipCoder, which represents lip motion using compact discrete semantic tokens to reduce the dependence on high-dimensional visual features [10]. Although these methods substantially improve the efficiency of image-based visual encoding, the visual front-end still contributes a considerable fraction of the total computation in lightweight AVTSE systems.

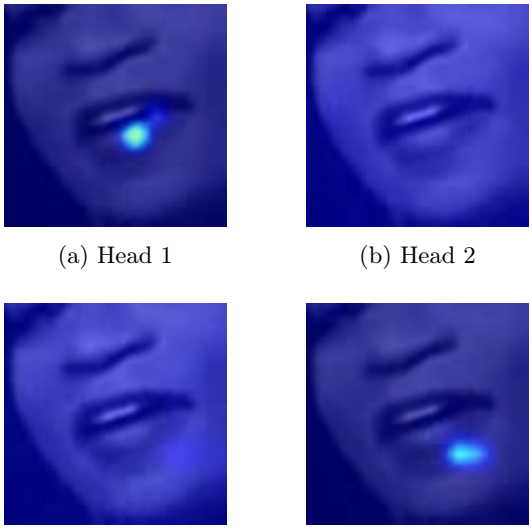
Besides designing lightweight image encoders, alternative visual representations have also been explored. PIAVE incorporates an additional facial representation generated by SADRNet into the separation framework [14]. During this process, SADRNet estimates facial landmarks as intermediate outputs for face alignment and facial reconstruction [15]. While PIAVE mainly utilizes the synthesized facial view, the landmark coordinates provide a compact geometric description of facial motion. Different from methods that rely on a single image-based visual representation, this work explicitly separates local lip appearance modeling from global geometric motion modeling. The proposed design is therefore positioned as a representation allocation strategy for reducing the computation of the visual front-end inside AVTSE systems.

3 Methods

3.1 Preliminary Observation and Motivation

To examine the spatial distribution of visual responses in AVTSE models, a visual self-attention module is inserted between the ResNet18 visual front-end and the audio-visual fusion module. Different from the original setting, where global average pooling is applied before feature fusion, the intermediate spatial feature map is preserved for visualization. For an 88×88 face ROI, ResNet18 produces a feature map with spatial resolution 11×11. After flattening the spatial dimensions, the multi-head self-attention module computes

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right), \tag{1}$$



(a) Head 1 (b) Head 2

(c) Head 3 (d) Head 4

Fig. 1: Spatial attention maps from 4 attention heads. The 11×11 feature maps are bilinearly upsampled to the original 88×88 lip ROI for visualization.

where $\mathbf{A} \in \mathbb{R}^{121\times121}$ denotes the attention relationship among spatial locations. For visualization, the attention responses are averaged over all query positions, reshaped into an 11×11 map, and upsampled to the original ROI resolution by bilinear interpolation.

Figure 1 shows attention maps produced by four attention heads. The visual examples suggest that some responses are concentrated around the inner lip region, whereas other responses are distributed over sparse peripheral regions around the mouth, including the subnasal region and the jaw. These qualitative observations motivate the hypothesis that visual cues used by AVTSE models may have heterogeneous spatial characteristics. The inner lip region contains dense local articulation information, while surrounding facial areas may mainly provide lower-dimensional geometric motion cues. Since attention visualization alone does not establish causal region importance, these observations are used only as design motivation. The effectiveness of the proposed representation is evaluated through controlled comparisons and ablation studies in Section 4.

Based on this hypothesis, a dual-path visual front-end is designed. The first path models the information-dense inner lip region using a compact lip ROI, while the second path represents sparse global facial dynamics using low-dimensional landmark coordinates.

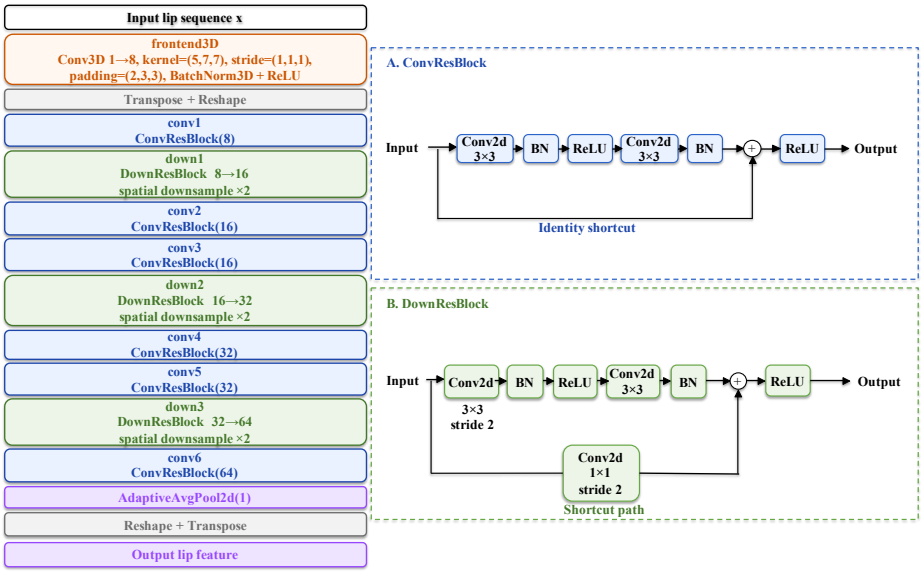


Fig. 2: Architecture of the Lip Encoder.

3.2 Lip Motion Modeling Branch

In the lip motion modeling branch, a local lip image patch is cropped from the original 96×96 facial ROI. The mean coordinate of the lip contour landmarks is used as the crop center, and a local window of 36×36 pixels is first extracted. During training, a 32×32 lip ROI is randomly cropped from this local window, and horizontal flipping is applied as data augmentation. During validation and testing, center cropping is used. This compact ROI is adopted to retain the main local lip motion while reducing computation over peripheral facial regions. In the ablation study, a horizontally extended 32×64 ROI is also evaluated. Here, 32×64 denotes height by width. This setting is included because a 32×32 crop may miss horizontally extended mouth shapes, whereas the wider crop can cover the complete lip contour along the horizontal direction.

The cropped lip image patch is processed by the Lip Encoder shown in Figure 2. The implementation contains an initial three-dimensional convolutional layer for short-term spatio-temporal modeling, followed by batch normalization and nonlinear activation. The resulting features are reshaped into frame-level feature maps and passed through a ResNet-style two-dimensional convolutional stack composed of spatial downsampling residual blocks and channel-preserving residual blocks. After spatial aggregation, a temporal projection layer produces a 64-channel visual feature sequence. In compact form, the structure is

$$\begin{aligned} &\text{Conv3D} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\quad \rightarrow \text{2D ResNet-style blocks} \\ &\quad \rightarrow \text{Spatial Pooling} \rightarrow \text{Temporal Projection}. \end{aligned} \tag{2}$$

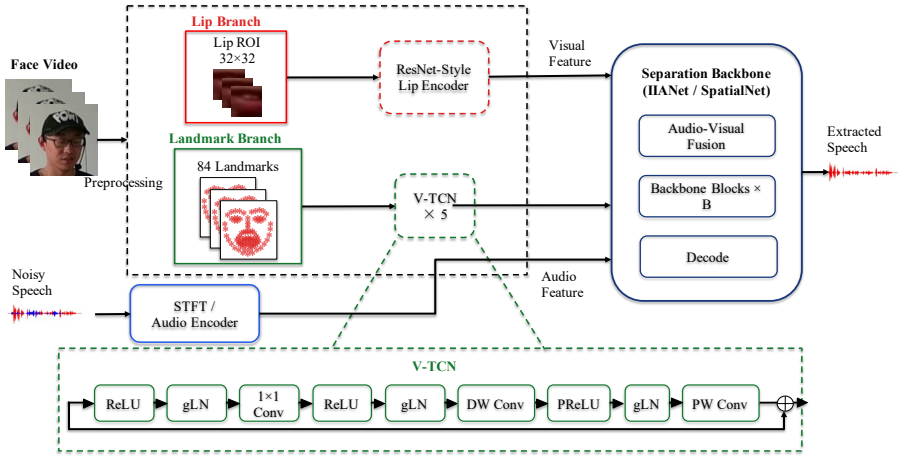


Fig. 3: Overall architecture of the proposed method.

The module is trained end to end for target speaker extraction. Compared with conventional 88×88 ROI-based encoders, this branch reduces convolutional computation mainly by limiting the input spatial range rather than by changing the basic convolutional modeling principle.

3.3 Facial Landmark Modeling Branch

Compared with image-based visual representations, facial landmark coordinates provide a compact description of facial geometry and motion. Since facial landmarks are commonly estimated during preprocessing for face alignment and mouth ROI localization, they can be reused as an additional visual representation when such preprocessing is already present. Under this assumption, the landmark branch introduces little additional computation inside the separation model. If landmark detection is not part of the original pipeline, its cost should be counted in the end-to-end system analysis.

The effectiveness of landmark-based modeling depends on localization quality. Conventional AVTSE pipelines commonly employ FAN-2D [16] for landmark estimation. In the proposed method, however, landmark coordinates are directly used as network input. Therefore, SADRNet [15], which was adopted in PIAVE [14], is used for landmark extraction. Under our THOP-based measurement protocol, processing a single 256×256 facial image requires 10.4 GMACs for SADRNet and 25.4 GMACs for FAN-2D. These values are reported to clarify the preprocessing assumption, but they are not included in the module-level visual front-end GMACs in Tables 1–3.

Motivated by the attention visualization in Section 3.1, the conventional 68-point landmark template is extended to 84 landmarks by introducing 16 additional points around the upper lip and jaw. The two-dimensional coordinates of

all landmarks are concatenated frame by frame and processed by V-TCN blocks [19] to model temporal facial dynamics.

Let $\mathbf{p}_{n,t,k} \in \mathbb{R}^2$ denote the coordinate of the k-th landmark in the t-th frame of the n-th training sample. Landmark normalization is performed independently for each landmark index and each coordinate dimension. The mean and standard deviation are computed over all frames in the training set as

$$\mu_{k,c} = \frac{1}{N_{\mathrm{tr}}} \sum_{n,t} p_{n,t,k,c}, \tag{3}$$

$$\sigma_{k,c} = \sqrt{\frac{1}{N_{\mathrm{tr}}} \sum_{n,t} \left(p_{n,t,k,c} - \mu_{k,c}\right)^2 + \epsilon}, \tag{4}$$

where $c \in \{x, y\}$ and N_{tr} denotes the total number of frames in the training set. The normalized coordinate is then computed as

$$\tilde{p}_{n,t,k,c} = \frac{p_{n,t,k,c} - \mu_{k,c}}{\sigma_{k,c}}. \tag{5}$$

The statistics $\mu_{k,c}$ and $\sigma_{k,c}$ are computed only from the training set and are reused for validation and testing.

During data augmentation, horizontal flipping is applied consistently to both the lip image and the corresponding landmark coordinates. The landmark transformation is performed before normalization. Specifically, the horizontal coordinate is mirrored according to the ROI width, and the left–right symmetric landmark indices are exchanged accordingly. This permutation is applied to both the original 68 facial landmarks and the additional 16 landmarks, so that each landmark channel preserves a consistent semantic meaning after augmentation.

3.4 Integration with Separation Backbones

The proposed dual-path visual front-end is evaluated on IIANet and SpatialNet. IIANet performs hierarchical audio-visual interaction through intra-modal and cross-modal attention mechanisms [8]. In this backbone, the proposed visual front-end replaces the original visual front-end while the remaining network architecture and fusion strategy are kept unchanged. SpatialNet is a time-frequency-domain separation network composed of alternating narrow-band and cross-band modules for modeling temporal, spectral, and spatial correlations [20]. Following the implementation in [21], the proposed visual front-end is incorporated as conditional visual information without modifying the original backbone architecture. These two backbones are used to examine whether the proposed visual representation can be integrated into different separation frameworks under a unified evaluation setting.

3.5 Visual Feature Fusion

The proposed visual front-end produces complementary features from the lip motion branch and the landmark branch. For IIANet, both branches produce

64-channel visual features. The two features are concatenated along the channel dimension and projected back to the original 64-dimensional feature space using a single 1D convolution before being fed into the original audio-visual fusion module. For SpatialNet, the visual feature dimensions are aligned with the frequency dimension of the input spectrogram. The lip feature is projected from 64 to 129 channels, while the V-TCN is configured to output 129-channel landmark features. The two visual features are then concatenated with the complex spectrogram along the introduced visual dimension, enabling the backbone to use both visual representations without changing its main separation architecture.

4 Experiments

4.1 Datasets

The experiments are conducted on the public LRS3-2mix dataset, following the CTCNet-related setting [3] to construct the training, validation, and test sets. The dataset contains 50,000 training mixtures, 5,000 validation mixtures, and 3,000 evaluation mixtures. Each mixture is generated by mixing two clean speech signals of 2 seconds at a sampling rate of 16 kHz with a random signal-to-noise ratio ranging from -5 dB to 5 dB. The corresponding original video is a face video with a frame rate of 25 FPS and a spatial size of 224×224. After preprocessing, a mouth patch window of 96×96 is obtained and further used to construct the inputs for different visual front-ends.

4.2 Training Configuration

The proposed visual front-end is compared with two representative image-based visual front-ends, namely the conventional pretrained ResNet18 encoder [13] and the lightweight BlazeNet64 encoder [11]. Both baselines use an 88×88 mouth ROI as input. During training, the input is randomly cropped from a 96×96 mouth patch with horizontal flipping applied for data augmentation. During validation and testing, center cropping is adopted.

For the proposed dual-branch visual front-end, the lip branch first extracts a 36×36 local region centered on the mouth area, which is further randomly cropped to 32×32 during training. The same horizontal flipping operation is applied consistently to both the lip image branch and the landmark branch. As described in Section 3.3, landmark channels are also permuted after flipping to preserve left–right semantic consistency. All mixture speech signals are pregenerated offline, and no dynamic mixing is used.

The training objective follows a scale-invariant waveform-level loss with an auxiliary scale-adjusted L_1 term. Given the target signal $\mathbf{s} \in \mathbb{R}^T$ and the estimated signal $\hat{\mathbf{s}} \in \mathbb{R}^T$, the scale factor is defined as

$$\alpha = \frac{\hat{\mathbf{s}}^\mathsf{T}\mathbf{s}}{\hat{\mathbf{s}}^\mathsf{T}\hat{\mathbf{s}}}, \tag{6}$$

and the loss is computed as

$$\mathcal{L}_{\text{SI-SDR}} = -10 \log_{10} \frac{\|\mathbf{s}\|_2^2}{\|\alpha\hat{\mathbf{s}} - \mathbf{s}\|_2^2}, \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{\text{SI-SDR}} + \lambda \left\|\alpha\hat{\mathbf{s}} - \mathbf{s}\right\|_1, \tag{8}$$

where T denotes the number of samples. The weighting factor λ is set to 1 in all experiments.

All models are trained on 8 V100 GPUs with a batch size of 3. The optimizer is Adam with an initial learning rate of 0.001. The learning rate is halved when the validation metric does not improve within a predefined tolerance, and training is stopped when the early stopping criterion is reached.

Due to computational constraints, a reduced training protocol is adopted instead of strictly following the original IIANet setting. Specifically, one of the two clean speech signals is randomly selected as the training target for each mixture, and two training schedules with different tolerance/early-stopping configurations are used (3/10 and 5/15). All methods in the same comparison group are trained under identical conditions. Therefore, the reported results are intended for relative comparison among visual front-ends under a unified setting, rather than for refreshing benchmark performance on LRS3-2mix.

4.3 Evaluation Metrics

The evaluation metrics include SI-SNR improvement, PESQ, and multiply-accumulate operations. Let $\mathbf{y}$ denote the mixture signal, $\mathbf{s}$ denote the clean target signal, and $\hat{\mathbf{s}}$ denote the estimated target signal. Following common speech separation evaluation practice [23], SI-SNR improvement is computed as

$$\text{SI-SNRi}(\mathbf{y}, \mathbf{s}, \hat{\mathbf{s}}) = \text{SI-SNR}(\mathbf{s}, \hat{\mathbf{s}}) - \text{SI-SNR}(\mathbf{s}, \mathbf{y}). \tag{9}$$

For two signals $\mathbf{u}$ and $\mathbf{v}$, where $\mathbf{u}$ denotes the reference signal and $\mathbf{v}$ denotes the evaluated signal, SI-SNR is defined as

$$\text{SI-SNR}(\mathbf{u}, \mathbf{v}) = 10 \log_{10} \frac{\|\mathbf{v}_{\text{target}}\|_2^2}{\|\mathbf{e}_{\text{noise}}\|_2^2}, \tag{10}$$

where

$$\mathbf{v}_{\text{target}} = \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\|\mathbf{u}\|_2^2}\mathbf{u}, \qquad \mathbf{e}_{\text{noise}} = \mathbf{v} - \mathbf{v}_{\text{target}}. \tag{11}$$

PESQ is used as a perceptual quality metric [24]. Computational complexity is measured using the `profile` function of the THOP library and reported in GMACs. The reported GMACs reflect theoretical module-level computation of the visual front-end inside the separation model and exclude face detection, video decoding, ROI extraction, and landmark estimation.

Table 1: Performance comparison of different visual front-ends on the LRS3-2mix test set using the IIANet backbone. The reported GMACs are represented as separation backbone computation + visual front-end computation. The base IIANet backbone requires 17.1 GMACs. Landmark extraction is excluded.

Tolerance	Early stop	Visual front-end	SI-SNRi	PESQ	GMACs
5	15	ResNet18(pretrained) [13]	17.5	2.90	17.1+8.0
		BlazeNet64 [11]	17.0	2.80	17.1+2.4
		32×32 + lmk84	17.1	2.80	17.1+0.3
3	10	ResNet18(pretrained)	16.2	2.64	17.1+8.0
		BlazeNet64	16.0	2.58	17.1+2.4
		32×32 + lmk84	16.0	2.59	17.1+0.3

Table 2: Performance comparison of different visual front-ends on the LRS3-2mix test set using the SpatialNet backbone. All models are trained with $tolerance = 3$ and $early_stop = 10$. The reported GMACs are represented as separation backbone computation + visual front-end computation. The base SpatialNet backbone requires 17.4 GMACs. Landmark extraction is excluded.

Visual front-end	SI-SNRi	PESQ	GMACs
ResNet18 (pretrained)	15.6	2.79	17.4+8.0
BlazeNet64	15.5	2.77	17.4+2.9
32×32 + lmk84	15.5	2.81	17.4+0.5

4.4 Experiment Results

Table 1 presents the experimental results of different visual front-ends using the IIANet backbone. Under the unified reduced training protocol, the proposed $32 \times 32 + lmk84$ dual-branch visual front-end reduces the module-level computation of the visual front-end while maintaining a separation performance close to BlazeNet64.

Under the $tolerance = 5$ and $early_stop = 15$ setting, the proposed method achieves the same PESQ as BlazeNet64 and a slightly higher SI-SNRi value. Since no multi-seed evaluation is conducted, this small numerical difference should not be interpreted as a statistically verified improvement. The main observation is that the proposed method provides a similar level of separation performance to the lightweight image-based baseline while reducing the measured visual front-end GMACs from 2.4 to 0.3 under the same THOP-based protocol.

Compared with the pretrained ResNet18 visual front-end, the proposed method shows a performance gap. This result suggests that pretrained image-based encoders still retain stronger capability in modeling fine-grained visual appearance. Nevertheless, the proposed method reduces the module-level visual front-end computation from 8.0 to 0.3 GMACs under the IIANet setting. The

Table 3: Ablation study of the proposed visual front-end on the LRS3-2mix test set using the IIANet backbone. All models are trained with $tolerance = 3$ and $early_stop = 10$. The reported GMACs are represented as separation backbone computation + visual front-end computation. The base IIANet backbone requires 17.1 GMACs. Landmark extraction is excluded.

Configuration	SI-SNRi	PESQ	GMACs
32×32 ROI + lmk84	16.0	2.59	17.1+0.3
32×32 ROI only	14.9	2.38	17.1+0.3
32×64 ROI only	15.0	2.32	17.1+0.6
lmk84 only	15.3	2.59	17.1+0.01
lmk68 only	15.2	2.58	17.1+0.01

result therefore reflects a trade-off between computation and separation quality, rather than an improvement in the upper bound of performance.

Under the more restricted $tolerance = 3$ and early stopping = 10 setting, all visual front-ends show lower scores. The proposed method remains numerically close to BlazeNet64, but the small differences among lightweight encoders should be interpreted as empirical trends under the current setup. Since the experiments adopt fixed mixture pairs, random target selection, and a reduced training schedule, the results are not intended to reproduce the original best results of IIANet.

Table 2 reports results on the SpatialNet backbone. A similar tendency is observed: the proposed method obtains lower computation than the two image-based visual front-ends, while its SI-SNRi and PESQ remain close to BlazeNet64 and moderately below ResNet18. This indicates that the proposed representation can be integrated into both evaluated backbones. However, broader architectural generality should be further examined with additional models and training runs.

Overall, the experimental results support the relative comparison objective of this work. The proposed method mainly improves the performance–computation trade-off of the visual front-end inside the separation model. The results should not be interpreted as establishing a new benchmark on LRS3-2mix.

4.5 Ablation Study

Table 3 presents the ablation results of the proposed dual-branch visual front-end. Compared with using only the 32×32 ROI, the full configuration combining the 32×32 ROI and lmk84 improves SI-SNRi from 14.9 dB to 16.0 dB and PESQ from 2.38 to 2.59. This indicates that, when the lip image region is strongly compressed, landmark coordinates provide complementary geometric information at a low computational cost.

The 32×64 ROI-only configuration is included to examine whether the performance loss of the 32×32 ROI is mainly caused by insufficient horizontal coverage. Here, 32×64 denotes height $\times$ width. The vertical size is kept the

same as that of the compact lip ROI, while the horizontal size is enlarged to cover the full mouth contour more reliably. This setting is motivated by the observation that a 32×32 crop may omit the mouth corners or part of the horizontal lip shape in frames with large articulation movements. The 32×64 ROI-only model slightly improves SI-SNRi over the 32×32 ROI-only model, but it does not reach the performance of the dual-branch configuration and introduces higher visual computation. This result suggests that simply enlarging the image crop is not as effective as combining compact local appearance with landmark-based geometric cues under the current setting.

When only the lmk84 branch is used, the model achieves 15.2 dB SI-SNRi and 2.58 PESQ. This indicates landmark sequences contain useful motion information. However, the remaining performance gap compared with the full configuration suggests that low-dimensional geometry does not fully replace local lip appearance. The two branches are therefore complementary in this experiment: the lip ROI branch provides local articulation details, while the landmark branch encodes broader facial motion with limited additional computation.

The ablation results are consistent with the design hypothesis that some peripheral facial cues can be represented more compactly by structured geometric coordinates than by dense image convolution. Since the ablation is conducted under a limited training protocol and without multi-seed evaluation, the conclusion is restricted to empirical trends observed in the current setting.

5 Discussion

The experimental results suggest that the computation of AVTSE visual front-ends can be reduced by reorganizing visual representations, rather than only by designing more compact image convolutional networks. A ResNet18-based visual front-end can model a relatively large mouth ROI and achieves stronger separation performance in the reported experiments. However, its module-level computation is non-negligible when it is paired with lightweight separation backbones. The proposed method instead decomposes visual information into local appearance cues and global geometric cues. The inner lip region is modeled using a compact ROI, while broader facial motion is represented by landmark coordinates. This design reduces the amount of dense image convolution used inside the separation model.

Compared with lightweight image-based encoders such as BlazeNet64, the proposed method differs mainly in the representation strategy. The inner lip region contains fine-grained texture and articulation patterns and is therefore modeled by a small image branch. Peripheral regions such as the outer lip, subnasal area, and jaw may provide auxiliary motion information, but processing them through a larger image ROI also increases convolutional computation. Landmark coordinates offer a compact alternative for representing such geometric variations. The ablation study supports this interpretation under the evaluated setting, since the combination of 32×32 ROI and lmk84 outperforms either branch alone.

The experiments on IIANet and SpatialNet indicate that the proposed visual representation can be inserted into two different audio-visual separation backbones. IIANet uses cross-modal attention for audio-visual interaction, whereas SpatialNet relies on time-frequency-domain modeling with visual conditioning. The similar trend observed on these two backbones suggests architectural compatibility, but it does not by itself establish broad generality across AVTSE systems. Further evaluation on additional architectures and datasets would be required for a stronger conclusion.

It is also important to clarify the computation scope. The GMACs reported in the tables measure the visual front-end modules inside the separation model and exclude video decoding, face detection, ROI extraction, and landmark estimation. This choice is made because the proposed method assumes that facial landmarks are already available from the preprocessing pipeline used for alignment and cropping. In practical systems where landmark estimation must be newly introduced, the cost of the landmark detector and its robustness should be evaluated as part of the complete pipeline.

The effectiveness of the landmark branch may also depend on the accuracy and stability of facial landmark localization. Although SADRNet is adopted in this work for pose-aware landmark extraction, the current experiments do not evaluate landmark noise, occlusion, large pose variation, or failure cases of face alignment. Therefore, the deployment relevance of the method should be interpreted with these limitations in mind. Hardware-specific latency is also not reported in this work, and the THOP-based GMACs should be regarded as a theoretical computation estimate rather than direct evidence of real-time performance.

Overall, the proposed method should be viewed as a trade-off design for low-computation AVTSE visual encoding. Its main contribution lies in combining compact local image modeling with low-dimensional geometric modeling. The current results support this direction under a unified LRS3-2mix evaluation setting, while broader robustness and end-to-end efficiency require further study.

6 Conclusion

This paper presents a dual-branch visual front-end for efficient Audio-Visual Target Speaker Extraction. The proposed design combines a compact lip ROI branch for local appearance modeling with a landmark branch for low-dimensional facial motion modeling. Experiments on LRS3-2mix under a unified reduced training protocol show that the proposed front-end achieves a performance–computation trade-off close to the evaluated lightweight image-based visual front-end while reducing the module-level computation introduced inside the separation model. The ablation results suggest that lip appearance and landmark-based geometric motion provide complementary cues. Since the current evaluation does not include multi-seed training, cross-dataset testing, or full end-to-end latency analysis including landmark detection, the results should be interpreted as evidence

of an efficient visual representation design under the evaluated setting rather than as a benchmark-level performance claim.

Acknowledgements This research is funded in part by the National Natural Science Foundation of China (62571223) and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100).

References

1. Wu, J., Xu, Y., Zhang, S.X., Chen, L.W., Yu, M., Xie, L., Yu, D.: Time domain audio visual speech separation. In Proc. of ASRU, pp. 667–673 (2019)
2. Gao, R., Grauman, K.: VisualVoice: Audio-visual speech separation with cross-modal consistency. In Proc. of CVPR, pp. 15495–15505 (2021)
3. Li, K., Xie, F., Chen, H., Yuan, K., Hu, X.: An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(10), 6637–6651 (2024)
4. Pan, Z., Wichern, G., Masuyama, Y., Germain, F.G., Hori, C., Le Roux, J.: Scenario-aware audio-visual TF-GridNet for target speech extraction. In Proc. of ASRU, pp. 1–8 (2023)
5. Blanco, A.L.A., Valentini-Botinhao, C., Klejch, O., Gogate, M., Dashtipour, K., Hussain, A., Bell, P.: AVSE Challenge: Audio-Visual Speech Enhancement Challenge. In Proc. of SLT, pp. 465–471 (2023)
6. Zhao, S., Pan, Z., Ma, B.: ClearerVoice-Studio: Bridging advanced speech processing research and practical deployment. arXiv preprint (2025)
7. Pegg, S., Li, K., Hu, X.: RTFS-Net: Recurrent time-frequency modelling for efficient audio-visual speech separation. In Proc. of ICLR (2024)
8. Li, K., Yang, R., Sun, F., Hu, X.: IIANet: An intra- and inter-modality attention network for audio-visual speech separation. In Proc. of ICML (2024)
9. Martel, H., Richter, J., Li, K., Hu, X., Gerkmann, T.: Audio-visual speech separation in noisy environments with a lightweight iterative model. In Proc. of INTERSPEECH, pp. 1–5 (2023)
10. Li, K., Gao, K., Hu, X.: Efficient audio-visual speech separation with discrete lip semantics and multi-scale global-local attention. arXiv preprint arXiv:2509.23610 (2025)
11. Pan, Z., Wang, W., Zhao, S., Zhang, C., Zhou, K., Ma, Y., Ma, B.: Online audio-visual autoregressive speaker extraction. In Proc. of INTERSPEECH (2025)
12. Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., Grundmann, M.: BlazeFace: Sub-millisecond neural face detection on mobile GPUs. arXiv preprint arXiv:1907.05047 (2019)
13. Martínez, B., Ma, P., Petridis, S., Pantic, M.: Lipreading using temporal convolutional networks. In Proc. of ICASSP, pp. 6319–6323 (2020)
14. Liu, Q., Ge, M., Wu, Z., Li, H.: PIAVE: A pose-invariant audio-visual speaker extraction network. In Proc. of INTERSPEECH, pp. 3719–3723 (2023)
15. Ruan, Z., Zou, C., Wu, L., Wu, G., Wang, L.: SADRNet: Self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction. IEEE Transactions on Image Processing **30**, 5793–5806 (2021)
16. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In Proc. of ICCV, pp. 1021–1030 (2017)

17. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: RetinaFace: Single-shot multi-level face localisation in the wild. In Proc. of CVPR, pp. 5203–5212 (2020)
18. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
19. Pan, Z., Ge, M., Li, H.: USEV: Universal speaker extraction with visual cue. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 3032–3045 (2022)
20. Quan, C., Li, X.: SpatialNet: Extensively learning spatial information for multi-channel joint speech separation, denoising and dereverberation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 1310–1323 (2024)
21. Gogate, M., Dashtipour, K., Hussain, A.: Towards personalised audio visual speech enhancement. In Proc. of INTERSPEECH, pp. 4853–4857 (2025)
22. Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S.: TF-GridNet: Integrating full- and sub-band modeling for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 3221–3236 (2023)
23. Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R.: SDR–half-baked or well done? In Proc. of ICASSP, pp. 626–630 (2019)
24. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs. In Proc. of ICASSP, vol. 2, pp. 749–752 (2001)