Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 6, No. 1, 2026

MS-VSSM: Multiscale Enhanced Visual State Space Model for Facial Expression Recognition

Qian Zhang

Southwest Minzu University, Chengdu 610000, China

Abstract: Facial Expression Recognition (FER) is a key technology in fields such as human-computer interaction and mental health assessment. However, its performance is constrained by the subtleties and dynamics of expressions, as well as the complexity arising from individual differences and environmental variations. To overcome the limitations of traditional Convolutional Neural Networks (CNNs), such as their restricted receptive fields, and the high computational complexity of Transformers, this paper proposes a Multiscale Enhanced Visual State Space Model (MS-VSSM) based on the Visual State Space Model (VSSM), aiming to improve the accuracy and robustness of FER. The model introduces three core improvements upon VSSM: (1) integrating a Path-aware Channel Attention mechanism (SE-SS2D) into the SS2D module to enhance the targeted capture of critical local facial features; (2) embedding a Dense Spatial Pyramid Pooling module (DSPP) at the beginning of each network stage to achieve multi-scale contextual information fusion; and (3) employing a Layer Scale mechanism to finely adjust the scale of deep features, thereby improving training stability and representational flexibility. In a four-category emotion recognition task constructed on the DEAP dataset, MS-VSSM achieved an accuracy of 97.31% and a weighted average F1-score of 97.56%, significantly outperforming the original VSSM and various mainstream visual backbone networks. These results validate the effectiveness and advancement of the proposed method. This study provides a new solution for efficient and accurate fine-grained facial expression recognition.

Keywords: Facial Expression Recognition, Visual State Space Model, Multi-scale Feature Fusion.

1. Introduction

They contribute to the awakening of self-awareness and, in concert with cognitive processes, foster the continuous development and refinement of personal behavior [5]. The concept of "affective computing" was introduced by Picard et al. [4] in 1995. Facial expressions, as one of the most direct and natural non-verbal signals for conveying human emotions and psychological states, have made their automatic recognition technology hold broad application prospects and significant research value in fields such as human-computer interaction, mental health assessment, intelligent driving, and security monitoring. The DEAP dataset, widely adopted as a benchmark in multimodal emotion analysis, provides a solid foundation for in-depth research on vision-based emotion recognition through its high-precision physiological signals and synchronously recorded facial videos. However, due to the subtlety, dynamic nature of facial expressions, and the complexity influenced by factors like individual differences and lighting conditions, how to construct efficient and robust models to accurately capture their essential features remains a core challenge in the field. Early research primarily relied on handcrafted features and machine learning algorithms, yet their representational capacity was limited. In recent years, deep learning methods represented by Convolutional Neural Networks (CNNs) and Transformers have greatly advanced the field. Nevertheless, they respectively suffer from limited receptive fields and high computational complexity. Against this backdrop, the Visual State Space Model, as an emerging architecture, demonstrates great potential in visual tasks by combining the dual advantages of a global receptive field and linear computational complexity, offering a new research direction for efficient and accurate facial expression recognition.

In recent years, deep learning (DL) techniques have advanced rapidly and attracted significant attention across various research fields. Numerous network architectures have been proposed, such as Deep Belief Networks (DBNs) [6], Convolutional Neural Networks (CNNs) [7], and Recurrent Neural Networks (RNNs) [8]. These models exhibit superior advantages in terms of computational efficiency and model performance, and have made substantial impacts in many domains, including image classification [9], speech recognition [10], and time-series forecasting [11, 12]. Regarding the research landscape both domestically and internationally, facial expression recognition technology has evolved from traditional methods to deep learning approaches. Studies based on traditional image methods primarily focused on designing robust feature descriptors, such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), combined with classifiers like Support Vector Machines (SVM) for emotion classification. These methods achieved certain success in controlled environments, but their handcrafted features lacked generalization capability for subtle expression variations and complex backgrounds. With breakthroughs in deep learning, deep learning-based methods for facial expression recognition have gradually become mainstream. Convolutional Neural Networks (CNNs) significantly improved recognition performance through their powerful local feature extraction capabilities. Subsequently, Vision Transformer (ViT) [2] and its variants further enhanced model performance by capturing long-range dependencies via the self-attention mechanism. However, CNNs require deep stacking to indirectly achieve global modeling, while Transformer's quadratic computational complexity imposes a heavy burden when processing high-resolution images. To address these bottlenecks, the Visual State Space Model (VSSM) [1] has emerged. Inspired by state space sequence modeling, VSSM achieves global context modeling for 2D images through structured state space

sequence layers and an efficient scanning mechanism, while maintaining linear computational complexity relative to input size. This characteristic enables an excellent balance between efficiency and accuracy across various visual tasks, marking an important innovative direction in visual modeling paradigms. It also provides a highly attractive foundational architecture for tasks like facial expression recognition, which requires balancing local details with global semantic information.

The main work and contributions of this paper lie in the proposal of an enhanced Visual State Space Model specifically tailored for the facial expression recognition task. The key aspects are outlined as follows:

(1) A path-aware channel attention mechanism is introduced into the SS2D module of the VSSM. This mechanism performs feature calibration on the outputs of the four scanning paths separately, thereby enhancing the model's ability to directionally capture key local cues of facial expressions.

(2) A multi-scale spatial pyramid pooling module is embedded in the early stages of the model. It explicitly fuses contextual information at different granularities, equipping the network with more comprehensive multi-scale representation capabilities.

(3) By incorporating the Layer Scale mechanism [3] to dynamically adjust the residual signals, the training stability and representational flexibility of the deep network are effectively improved. These enhancements work synergistically to collectively boost the model's recognition accuracy for complex expressions.

Experimental results on the DEAP dataset show that the improved model significantly outperforms the original VSSM and various mainstream visual backbone networks. This validates the effectiveness and advancement of the proposed innovations, offering a new solution for high-precision and efficient facial expression recognition.

2. Related Technologies and Theoretical Foundations

2.1. Overview of Facial Expression Recognition

Facial Expression Recognition (FER) aims to automatically identify and understand discrete emotional states (such as happiness, sadness, anger, etc.) or continuous dimensional emotions (such as valence and arousal) conveyed by human facial expressions through computational models. Its core challenge lies in the highly non-rigid, subtle, and subjective nature of facial expressions, which are also significantly affected by external factors such as individual differences, pose, lighting, and occlusions. Traditional methods primarily relied on feature engineering, using handcrafted descriptors like Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), or geometry-based features from key points to capture facial texture or shape variations, combined with classifiers such as Support Vector Machines (SVM) for discrimination. The performance of these methods heavily depended on the completeness of the feature design, resulting in limited generalization capability. The rise of deep learning has fundamentally transformed this field. Deep models, represented by Convolutional Neural Networks (CNNs), can learn hierarchical feature representations end-to-end from vast amounts of data, ranging from edges and textures to higher-level semantic expression patterns, significantly improving recognition robustness and accuracy. In recent years, Vision Transformers have further pushed the performance boundaries by modeling global dependencies through self-attention mechanisms. However, a key research problem remains: how to accurately capture the dynamic patterns in expressions—which involve both subtle local muscle movements and global coordination across facial regions—under the constraint of efficient computation.

2.2. Fundamentals of the Visual State Space Model

The Visual State Space Model (VSSM) is a novel visual backbone network architecture that integrates classical continuous state-space systems with modern deep learning methodologies. Its theoretical foundation originates from the Structured State Space Sequence Model (S4), a system that utilizes a latent, continuous "state" to memorize and integrate historical information for predicting subsequent outputs, demonstrating exceptional efficiency and performance in long-sequence modeling. VSSM adapts its core operator, the State Space Model (SSM), to the domain of two-dimensional vision. Specifically, its key innovation lies in the SS2D module. This module employs a "Cross-Scan" operation to unfold the 2D image features into one-dimensional sequences along four specific directions (top-left to bottom-right, bottom-left to top-right, top-right to bottom-left, and bottom-right to top-left), thereby mapping the 2D spatial structure into multiple 1D paths that can be processed by the state-space system. Subsequently, each path is processed by an S6 block—a parameterized and discretized state-space model—capable of modeling long-range dependencies with linear complexity. Finally, the processed sequences are folded back into a 2D spatial structure via a "Cross-Merge" operation, which integrates information from different scanning directions. This design enables VSSM to effectively capture global contextual information without incurring the quadratic computational complexity of self-attention mechanisms, while maintaining a hierarchical structure similar to Convolutional Neural Networks. The success of VSSM provides an efficient and powerful new paradigm for visual tasks requiring dense, long-range modeling (such as image classification and semantic segmentation) and offers a highly promising foundational model choice for facial expression recognition, a task that necessitates the comprehensive analysis of both local details and the overall facial expression configuration.

2.3. Attention and Multi-scale Feature Fusion Mechanisms

The core idea of attention mechanisms is to mimic the human cognitive process, enabling models to dynamically focus on the more important parts of the input. In computer vision, channel attention (e.g., the Squeeze-and-Excitation module) recalibrates the feature map along the channel dimension by learning importance weights for each channel, thereby enhancing useful features and suppressing redundant ones. Spatial attention focuses on critical locations within the feature map. Path-aware attention represents a more fine-grained application, performing independent or interactive attention computations across different processing paths or branches to capture more nuanced patterns. In facial expression recognition, attention mechanisms help models concentrate on regions most relevant to emotional expression (such as the eyes and mouth) while ignoring irrelevant

background interference. Multi-scale feature fusion mechanisms aim to address the problem where target objects exhibit different characteristics at varying scales. Facial expressions are constituted by both the coordination of global muscle groups and local subtle distortions, making multi-scale information crucial. Spatial Pyramid Pooling (SPP) is a typical representative of this mechanism, which captures multi-granularity information from fine local details to macro-global context through parallel pooling layers with kernels of different sizes, subsequently fusing or concatenating feature maps from these different scales. Densely connected multi-scale designs (e.g., DSPP) further enhance the reuse and interaction of multi-scale features. Combining attention mechanisms with multi-scale fusion allows the model to first extract rich contextual features at different scales and then adaptively select and integrate the most discriminative parts from these features, thereby constructing a representation that is more robust to scale variations and local details. Building upon this concept, this study integrates path-aware attention and a dense multi-scale pyramid fusion mechanism into the VSSM framework, with the goal of modeling complex facial expression features more comprehensively and accurately.

3. The MS-VSSM Model Architecture

3.1. Overall Model Architecture

To address the characteristics of facial expression features in the recognition task, such as their multi-scale nature, local subtlety, and significant inter-channel importance differences, this paper proposes an enhanced model built upon the Visual State Space Model (VSSM) that integrates multi-scale context and channel attention. This model is named the Multi-Scale Visual State Space Model (MS-VSSM), as illustrated in Figure 1. The overall model adopts a hierarchical encoder architecture, consisting of an input layer, four feature extraction stages, and a classification head. The core improvements lie in the introduction of a Dense Spatial Pyramid Pooling (DSPP) module at the beginning of each feature extraction stage, and the enhancement of the fundamental Visual State Space Block (VSS Block) with channel attention and feature scale normalization.

Given an input facial image with dimensions $H \times W \times 3$, it first passes through a Stem module composed of convolutional layers for preliminary downsampling and feature mapping, yielding an initial feature map with spatial dimensions of $H/4 \times W/4$ and a channel count of C. Subsequently, the feature map sequentially passes through four feature extraction stages that share similar structures but progressively deepen and widen. The core of each stage consists of a stack of improved VSS Blocks. Additionally, a specially designed DSPP module is integrated at the beginning of each stage to adaptively aggregate and fuse multi-scale contextual information related to expressions before deep state-space modeling. Between stages, feature map downsampling is performed via convolutional layers with a stride of 2 or patch merging operations, while simultaneously expanding the channel dimensions, thereby constructing a pyramidal feature representation where resolution progressively decreases and semantic information is gradually enhanced.

Specifically, the first stage does not perform downsampling after the Stem. It directly processes the initial features through the DSPP module, followed by a stack of multiple improved VSS Blocks, outputting a feature map with spatial dimensions of $H/8 \times W/8$. The second stage begins by downsampling the output from the previous stage, then similarly processes it through the DSPP module and multiple VSS Blocks, reducing the output size to $H/16 \times W/16$. Serving as the core computational layer of the network, the third stage, after similar downsampling and DSPP processing, stacks the deepest VSS Blocks, outputting a feature map of size $H/32 \times W/32$. The fourth stage, after further downsampling, employs only a small number of VSS Blocks for final refinement of the high-dimensional features, maintaining the spatial size of $H/32 \times W/32$. Finally, at the network's tail, global average pooling is applied to compress each channel's feature map into a single scalar value, forming a high-dimensional feature vector. This is then fed into a fully connected classification layer, which outputs a probability distribution corresponding to different expression categories. This overall architecture retains the advantages of VSSM, namely linear computational complexity and a global receptive field, while significantly enhancing the model's capability to capture subtle expression details and the efficacy of multi-scale feature fusion through the systematic embedding of the DSPP module and the enhanced VSS Blocks.

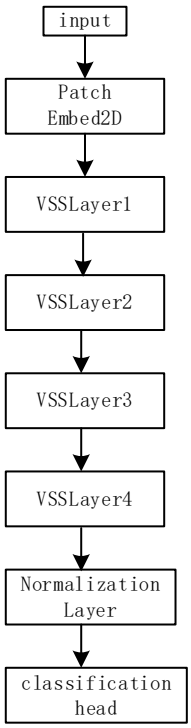

Figure 1. The Overall Model Structure Diagram

3.2. Basic VSSM Modules

3.2.1. Stem and Downsampling Layers

The input processing of the MS-VSSM model begins with the Stem module, which is responsible for performing preliminary feature extraction and spatial downsampling on the original facial image in an efficient manner that preserves rich detail. Specifically, an input image of size $H \times W \times 3$ first passes through a Stem structure composed of two consecutive 3×3 standard convolutional layers. The first convolutional layer uses a stride of 2, halving the spatial dimensions of the feature map. This is followed by Layer Normalization (LayerNorm) and the Gaussian Error Linear Unit (GELU) activation function for feature normalization and non-linear transformation. The second convolutional layer also employs a stride of 2, further compressing the feature map size to 1/4

of the input. Ultimately, the Stem module outputs a feature map with a spatial resolution of H/4 × W/4 and a channel count of C, the initial design dimension. This gradual downsampling strategy using small convolutional kernels, compared to single-step large-stride downsampling, allows for a more gentle reduction in spatial resolution. This helps preserve the subtle textures and contour information crucial for facial expressions in the early stages, laying a solid foundation for subsequent deep feature modeling.

Between the model's four hierarchical feature extraction stages, spatial downsampling is required to expand the receptive field and increase channel dimensions, thereby constructing a multi-scale feature pyramid. MS-VSSM employs structurally simple and parameter-efficient strided convolution as the core downsampling method. At each stage transition, a 3×3 convolutional layer with a stride of 2 is introduced. This convolutional operation reduces the height and width of the feature map by half while doubling the number of feature channels compared to the previous stage (e.g., from C to 2C) by increasing the number of convolutional filters. Layer normalization is also applied after the convolution to stabilize training. This downsampling method avoids complex operations. Leveraging the inherent parameter-sharing property of convolution, it can effectively aggregate local spatial information and achieve feature dimension transformation. This ensures that the model maintains effective encoding capability for the spatial structure of expression-related regions during the process of deepening semantic understanding. It provides feature inputs of appropriate scale and information density for the subsequent multi-scale fusion in the DSPP module and the sequential modeling in the VSS Blocks.

3.2.2. Basic VSS Block Structure

The basic VSS Block is the core unit of the entire VSSM model for feature transformation and contextual information modeling. Its design is inspired by the Mamba block and has been adapted for two-dimensional visual data. This module adopts a dual-branch residual structure, sequentially performing sequence modeling based on the state space model and feature mixing based on a feed-forward network, with a clear and efficient computational flow. Given an input feature map X, the basic VSS Block first applies Layer Normalization to it, followed by a linear projection layer that expands the channel count to ssm_ratio × C. This expansion factor ssm_ratio is intended to enhance the module's expressive capacity. Immediately after, a 3×3 depthwise separable convolution is applied to the features. This step is crucial as it injects local spatial inductive bias into the model, helping the network establish a fundamental perception of local image structures (such as facial feature edges and textures) in the early stages, compensating for the lack of local correlation modeling in the initial phase of pure sequence-based approaches.

Subsequently, the features are fed into the core of the module—the 2D Selective Scan (SS2D) module. SS2D unfolds the 2D feature map into sequences along specific scanning paths and utilizes a state space model with a selection mechanism (S6) to model global dependencies with linear complexity for each path. After processing by SS2D, the features pass through another linear projection layer that compresses the channel count back to the original C. The operations described above constitute one main processing branch. Its output is added to the block's initial input X via a residual connection, completing the first feature fusion and information preservation. This intermediate output then passes through a second Layer Normalization layer and is fed into a classic Feed-Forward Network (FFN). This FFN typically consists of two fully connected layers with an activation function in between and is responsible for inter-channel information interaction and non-linear transformation. The output of the FFN is finally added to the intermediate output through a second residual connection, producing the final output of this VSS Block.

This structural design cleverly combines two complementary feature processing mechanisms: the SS2D branch focuses on capturing long-range, global spatial dependencies with linear complexity, which is crucial for understanding the relationships between expression components distributed across different facial regions; while the FFN branch focuses on per-position, channel-wise feature transformation and fusion. Both are normalized by Layer Normalization and stabilized through residual connections to ensure smooth gradient flow. Together, they form a powerful and stackable basic feature transformation unit, providing a solid foundation for subsequent attention-based enhancements.

3.2.3. Principles of the SS2D Module

The SS2D module is the core computational unit of the basic VSS Block. Its design aims to efficiently and appropriately adapt the selective state space model—originally designed for one-dimensional sequences—to two-dimensional image data, which has a non-sequential structure. Traditional selective scanning mechanisms require data to have a clear sequential order, whereas pixels in an image are arranged equally in space without an inherent, unique traversal order. To resolve this fundamental conflict, SS2D introduces a "cross-scanning" strategy. This strategy constructs global contextual information from multiple directions for each position in the image without introducing quadratic computational complexity.

Specifically, given a 2D feature map $X \in R^{H \times W \times C}$, SS2D first performs the Cross-Scan operation. This operation unfolds the feature map into four independent 1D sequences along four predefined, complementary paths: from top-left to bottom-right (row-major), from bottom-right to top-left (reverse row-major), from top-right to bottom-left (a column-major variant), and from bottom-left to top-right (a reverse column-major variant). Each scanning path assigns a unique, path-dependent sequential position to every pixel in the feature map, thereby transforming the 2D spatial neighborhood relationships into proximity relationships on a 1D sequence. This process converts the input feature map into four sequences with a shape of [L, C], where L = H × W is the sequence length.

Subsequently, these four sequences are fed into four independent S6 blocks for processing. The S6 block is the core of the selective state space model, whose parameters (such as the state transition matrix) can be dynamically adjusted based on the current input (i.e., "selectivity"). This enables adaptive weighting of the importance of different contextual information. Each S6 block performs linear-time complexity, recurrent global modeling on its input 1D sequence, effectively capturing long-range dependencies along that specific scanning direction. This process can be understood as information being propagated and integrated sequentially along each path, where later positions in the sequence can implicitly aggregate information from all preceding positions.

Finally, the SS2D module executes the Cross-Merge operation. This operation rearranges the four processed output sequences from the S6 blocks back into the original two-dimensional spatial grid structure based on the reverse mapping of their respective original scanning paths, resulting in four feature maps each with dimensions of H × W × C. Each of these feature maps encapsulates global contextual information from a distinct scanning direction. By aggregating them through element-wise summation (or concatenation followed by linear fusion), the SS2D module ultimately outputs a single feature map that integrates multi-directional global information.

Therefore, the ingenuity of the SS2D module lies in its use of four meticulously designed scanning paths to approximate information interaction between any two pixels in the image (since for any pair of pixels, there always exists a path where one precedes the other in the sequence). This equivalently establishes a global receptive field. Moreover, thanks to the properties of the selective state space model, the entire process achieves computational complexity that is only linear with respect to the number of pixels. This overcomes the quadratic complexity bottleneck inherent in traditional self-attention mechanisms, making it particularly suitable for facial expression recognition tasks that require processing high-resolution inputs.

3.3. Design of Attention-Enhanced VSS Blocks

3.3.1. SE-SS2D: Channel-Selective Enhancement for Expression Features

In the basic SS2D module, the features from the four scanning paths are merged directly after global modeling, which implicitly assumes that all channel features contribute equally to the final expression representation. However, in facial expression recognition tasks, different channels often correspond to different feature detectors (e.g., some channels may be sensitive to subtle changes in the eye region, while others respond more strongly to mouth contour shapes). Moreover, the key feature channels dynamically vary for different input expressions. To endow the model with the ability to dynamically calibrate the importance of channel features, enabling it to focus on the semantic information most relevant to the current expression, we integrate a Squeeze-and-Excitation (SE) module into the SS2D module, forming the enhanced SE-SS2D unit.

Specifically, in the improved design, after the feature sequence from each scanning path completes its selective state-space modeling via its dedicated S6 block, it is not immediately sent to the Cross-Merge stage. Instead, the output sequence of each path is first independently reshaped back into a two-dimensional spatial structure (H × W × C) and then processed through a lightweight SE module. This SE module first performs global average pooling on the features across all spatial positions for each channel, generating a compressed channel descriptor vector that represents the global activation intensity of each channel. Subsequently, this descriptor passes through a bottleneck structure composed of two fully connected layers. The first fully connected layer reduces dimensionality and introduces non-linearity (typically using a ReLU activation), and the second fully connected layer restores the original channel dimension. Finally, a Sigmoid activation function outputs a set of channel weights between 0 and 1. These weights are then multiplied channel-wise with the original path features, thereby adaptively recalibrating the feature maps of each channel: suppressing the responses of channels that contribute less to the current expression and enhancing the responses of key channels.

After completing the independent channel weight adjustment for the four paths, the Cross-Merge operation is performed to fuse the four feature maps that have been modulated by channel attention. This improvement offers significant advantages: First, it seamlessly embeds the expression-sensitive channel attention mechanism into the global context modeling workflow, allowing the model to perform fine-grained, channel-level feature selection while establishing long-range dependencies. Second, since the SE module is applied separately to each scanning path, it can perform differentiated calibration based on the feature distributions obtained from modeling different directions, further enhancing the flexibility of feature selection. Finally, the entire SE-SS2D module adds only minimal parameters and computational overhead yet significantly improves the model's ability to capture discriminative expression features, enabling the model to more effectively extract key features strongly correlated with emotional states from complex facial backgrounds.

3.3.2. Layer Scale: Depth-wise Fine-grained Modulation for Expression Features

As network depth increases, the scale of feature magnitudes during forward propagation may drift, affecting training stability and convergence efficiency. This is particularly crucial when processing subtle and sensitive facial expression signals. To further finely regulate the scale of the feature flow and introduce an implicit, layer-wise adaptive modulation mechanism in deep networks, we integrate the Layer Scale technique into the improved VSS Block. This technique introduces a set of learnable, channel-wise scaling parameters before the output of specific residual branches, enabling lightweight, adaptable calibration of feature magnitudes.

In the VSS Block design of this paper, Layer Scale is placed after the SE-SS2D module and before the linear layer that projects the features back to the original channel dimension. Specifically, assuming the intermediate feature processed by the SE-SS2D module is $F \in R^{H \times W \times (ssm\_ratio \cdot C)}$, the Layer Scale operation is defined as element-wise multiplication of this feature with a learnable diagonal scaling matrix Λ, i.e., $F_{out} = F\Theta\Lambda$, where Λ is initialized as a small constant vector close to zero (e.g., 1e-5). This design implies that in the early stages of training, the contribution of the main residual path computed by the SS2D and SE modules is significantly suppressed. The output of the entire block is predominantly governed by the identity mapping (the residual connection), which greatly ensures the stability of gradient flow and a smooth initial training phase. As training progresses, the network autonomously learns and updates the scaling parameters Λ through gradient descent, thereby gradually unlocking and modulating the contribution of this complex non-linear transformation path to the final output.

Applying the Layer Scale mechanism to expression recognition models offers a dual advantage. Firstly, as an adaptive form of deep regularization, it introduces feature transformations in a controllable manner, effectively mitigating potential issues such as gradient anomalies or feature magnitude explosion in deep networks. This enhances the robustness of model training, which is crucial for a task that requires capturing subtle emotional variations from high-

dimensional facial data. Secondly, it equips the network with the ability for fine-grained, channel-wise regulation of feature representations. Different channels may correspond to feature detectors for distinct expression components (e.g., raised corners of the mouth, furrowed brows). Layer Scale allows the model to dynamically adjust the relative intensity of these detectors within the final fused features, achieving more precise and in-depth calibration of expression characteristics.

Complementing the aforementioned SE module (which focuses on cross-channel attention re-weighting), Layer Scale focuses on modulating the scale of the feature magnitudes themselves. Working in synergy, they together form a comprehensive enhancement framework for expression features—spanning from importance selection to intensity fine-tuning—significantly boosting the model's capability to model complex facial expression representations.

3.4. Multiscale Expression Fusion Module

The expression of facial emotions is a classic problem of cross-scale information fusion: macro-level overall facial contour and muscle group orientation convey the basic categorical information of emotions, meso-scale regional feature variations (such as in the eye-brow and mouth-nose areas) are key to distinguishing similar emotions, and micro-level local textures and edge details (such as fine lines around the eyes or subtle curvature of the mouth corners) are crucial for fine-grained emotion discrimination. While the base VSSM model, with its global receptive field, excels at modeling long-range dependencies, it has inherent limitations in explicitly and structurally capturing and fusing such multi-scale spatial contextual information. To compensate for this deficiency and endow the model with powerful multi-scale perception capabilities from the early feature extraction stages, we embed a lightweight yet efficient Multiscale Expression Fusion Module at the beginning of each feature extraction stage. Its core is an improved Dilated Spatial Pyramid Pooling structure.

This module takes the input feature map $X_{in} \in R^{H \times W \times C}$ of the current stage as its processing target. First, a 1×1 convolutional layer adjusts the channels and performs preliminary feature transformation on the input. Subsequently, the features are fed into four parallel branches for multi-scale context extraction. The first branch employs global average pooling to compress the feature map to a spatial size of 1×1, followed by a 1×1 convolution for non-linear transformation. This branch aims to capture global contextual features related to the overall expression state. The second, third, and fourth branches employ average pooling layers with fixed kernel sizes of 3×3, 5×5, and 7×7 respectively, coupled with dilated convolutions of corresponding dilation rates (e.g., 3×3 dilated convolutions with dilation rates of 1, 2, and 3, respectively). This simulates receptive fields of different sizes while maintaining the spatial resolution (H×W) of the feature map. These branches are responsible for extracting contextual information corresponding to local subtle variations, regional patterns, and broader facial region associations. The output of each branch is then processed through a 1×1 convolution for uniform dimensionality alignment. Finally, the feature maps from the four branches are concatenated along the channel dimension, fusing expression representations from global to local contexts across varying granularity levels. A final 1×1 convolutional layer is applied to the concatenated features. Its role is to adaptively reweight and integrate the information from different scales, learning the optimal multi-scale feature combination for the current expression recognition task, and producing a feature map with the same channel count as the input.

Embedding the DSPP module before the stack of VSS Blocks in each stage offers significant design advantages. It provides rich and structured multi-scale prior information for the subsequent VSS Blocks, which excel at global sequence modeling, enabling the model to simultaneously "see the forest and the trees" of an expression. This design paradigm effectively addresses the limitations of a single-scale receptive field, enhances the model's robustness to inputs of varying resolutions and facial regions of different sizes, and allows it to more comprehensively capture the full semantic spectrum of expressions—from the macroscopic emotional tone to microscopic muscular movements. Consequently, this facilitates more accurate and robust emotion recognition in complex real-world scenarios.

4. Experiments

4.1. Experimental Environment and Configuration

This paper implemented the proposed MS-VSSM model using the PyTorch framework. All experiments were conducted on a server, and the specific experimental environment configuration is detailed in Table 1.

Table 1. Experimental Environment Configuration

Item	Specification
GPU	NVIDIA RTX 4090(24GB) 1
RAM	120G
CPU	16 vCPU Intel(R) Xeon(R) Gold 6430
IDE	PyCharm 2025.1.1.1
Language Environment	Python 3.12(ubuntu22.04)
Deep Learning Framework	PyTorch 2.3.0

4.1.1. Dataset and Preprocessing

This study utilizes the facial video data and their synchronized affective dimensional labels from the DEAP (Database for Emotion Analysis using Physiological Signals) dataset to construct a fine-grained visual emotion recognition task. The DEAP dataset records multimodal signals from 32 participants while they watched 40 one-minute music video clips. This includes frontal face videos for each participant, as well as their self-assessed post-experiment dimensional ratings for Arousal (representing the degree of physiological activation) and Valence (representing the degree of pleasantness) for each video clip (continuous values, both ranging from 1 to 9). To build a more discriminative categorical model for emotional states, this study comprehensively considers both the Arousal and Valence dimensions to partition the data into four emotion categories, aiming to thoroughly evaluate the model's ability to perceive complex emotional states from facial expressions. The raw video data underwent a systematic preprocessing pipeline to generate high-quality image samples suitable for deep network training. The specific steps are as follows:

(1) Image Extraction and Keyframe Sampling: Each facial video was decoded. To obtain frames where emotional expressions were relatively stable and fully developed, the initial adaptation phase of each video was discarded. Starting

uniformly from the 3rd second, video frames were extracted at a rate of 1 frame per second, forming a sequence of key images representing the emotional content of that video clip.

(2) Face Detection and Region Cropping: A pre-trained frontal face detector from the `dlib` library was used to automatically locate the face bounding box in each frame. To ensure capture of the complete facial expression region (including the forehead, chin, and cheek contours), the detected bounding box was uniformly expanded by 20% of its original dimensions on all sides. Subsequently, standardized facial region images were cropped based on the expanded bounding box.

(3) Dual-Dimensional Label Fusion and Four-Class Construction: Each source video clip corresponds to a set of Arousal (A) and Valence (V) ratings. To construct a classification task, the continuous ratings for both dimensions were first binarized using the median value of 5 as the threshold: High Arousal (HA, $A \geq 5$) vs. Low Arousal (LA, $A < 5$), and High Valence (HV, $V \geq 5$) vs. Low Valence (LV, $V < 5$). These two binarized dimensions were then combined to form four mutually exclusive emotion categories: High Valence High Arousal (HVHA), High Valence Low Arousal (HVLA), Low Valence High Arousal (LVHA), and Low Valence Low Arousal (LVLA). All facial images extracted from the same video inherited this four-class label and were stored according to their category, thereby establishing an affective image dataset based on the two-dimensional Arousal-Valence model.

(4) Data Augmentation: During the model training phase, a series of random augmentation operations were applied to the input images to improve generalization performance. These included geometric transformations, such as random horizontal flipping, minor rotation (±10°), and translation; as well as photometric transformations, such as random adjustments to brightness, contrast, and saturation. These operations effectively increased the diversity of the training data while preserving the semantic content of the expressions.

(5) Size Standardization and Normalization: Finally, all cropped facial images were resized to a fixed resolution required for model input. Prior to being fed into the network, the pixel values of the images were further normalized. Through the above pipeline, the raw multimodal DEAP data was transformed into a well-structured, rigorously labeled four-class facial expression image dataset. This provides a reliable data foundation for the subsequent training and evaluation of deep learning-based fine-grained emotion recognition models.

4.1.2. Evaluation Metrics

To systematically evaluate the performance of the proposed model on the four-class emotion recognition task, this study employs a comprehensive evaluation framework encompassing both overall accuracy and fine-grained balance. The core metrics include Accuracy and the Weighted Average F1 Score, aiming to fully reflect the model's classification effectiveness from different dimensions. Accuracy, as the most intuitive metric, measures the overall proportion of correctly predicted samples, providing an initial overview of the model's global classification capability.

Given that the constructed four-class dataset in this study may exhibit an imbalanced class distribution, relying solely on accuracy cannot sufficiently reveal the model's specific performance differences across categories. Therefore, this study emphasizes the introduction of the Weighted Average F1 Score as a core evaluation metric. The F1 Score is the harmonic mean of Precision and Recall, effectively balancing the focus on prediction correctness (precision) and coverage completeness (recall). Its weighted average form further accounts for differences in the number of samples per class. By assigning weights proportional to the sample count of each class, it calculates a single numerical value that better reflects the model's comprehensive performance on the real data distribution. The calculation formula for the Weighted Average F1 Score is as follows:

$$F1 = \sum_{i=1}^{N} \omega_i \cdot F1_i$$

Where N represents the total number of classes (in this study, N = 4), $F1_i$ is the F1 score for the i-th class, and ω_i is the weight for that class, typically set as the proportion of the number of samples in that class to the total number of samples, i.e $\omega_i = \frac{\text{sample size\_}i}{\text{Total sample size}}$. This metric's value is more susceptible to the performance of classes with larger sample sizes, thereby providing a robust overall performance measure that closely aligns with real-world application scenarios. By reporting both Accuracy and the Weighted Average F1 Score concurrently, this study can not only demonstrate the model's overall classification correctness but also provide an in-depth evaluation of its robustness and generalization capability when confronted with imbalanced class distributions. This enables a more comprehensive and reliable assessment of the model's performance.

4.1.3. Selection of Baseline Models

To ensure that the proposed architecture is built upon a robust foundation, this study first conducted systematic comparative experiments on the DEAP four-classification task using several state-of-the-art visual backbone networks. The candidate models encompass representative efficient architectures: RepViT, a lightweight Vision Transformer optimized for efficiency via structural reparameterization techniques; DilatedFormer, a Transformer variant that incorporates dilated convolutions to expand the receptive field and enhance multi-scale modeling capabilities; and the state-space model-based VSSM. Under identical experimental settings, preprocessing pipelines, and training strategies, the performance of each model on the test set is shown in Table 2.

Table 2. Performance Comparison of Baseline Models

Model	ACC	F1
RepViT	0.9100	0.9079
DilatedFormer	0.9390	0.9352
VSSM	0.9403	0.9403

The experimental results demonstrate that the VSSM model achieved the best performance on both metrics—Accuracy and the Weighted Average F1 Score. Although DilatedFormer also showed strong competitiveness due to its explicit multi-scale design, VSSM still held a slight lead. More notably, VSSM's Weighted Average F1 Score was exactly equal to its Accuracy, both at 0.9403. This indicates that VSSM possesses a more balanced discriminative capability across the four emotion categories, mitigating performance bias potentially caused by class imbalance. In contrast, RepViT's performance showed a certain gap

compared to the former two models. The superior performance exhibited by VSSM can be attributed to the linear-complexity global modeling capability provided by its core SS2D module. This enables it to efficiently capture the widely distributed, long-range spatial dependencies within facial expressions, which is crucial for integrating emotional signals scattered across different facial regions. Based on its comprehensive advantages in both performance and efficiency, this study selects VSSM as the baseline model for subsequent enhancements and improvements, aiming to further explore its potential in fine-grained emotion recognition tasks.

4.2. Experimental Results and Analysis

4.2.1. Experimental Results

After training and validation on the DEAP four-category facial expression dataset, the improved model proposed in this paper achieved excellent performance. On the test set, the model achieved an overall accuracy of 0.9731 and a weighted average F1 score of 0.9756. These two closely matched metrics, both exceeding 0.97, fully demonstrate that the model possesses an exceptionally high overall classification correctness and maintains outstanding and balanced discriminative capability even when facing potential class imbalance.

To deeply analyze the model's fine-grained classification behavior, we examined its normalized confusion matrix, as shown in Figure 2. The diagonal elements of the matrix represent the correct classification rates for each category. Among them, the High Valence High Arousal (HVHA) category achieved a recognition rate as high as 99.35%; the recognition rates for the High Valence Low Arousal (HVLA), Low Valence High Arousal (LVHA), and Low Valence Low Arousal (LVLA) categories also reached 96.82%, 96.28%, and 97.47%, respectively. This indicates that the model has developed a strong ability to distinguish all four emotion categories. The primary misclassifications occurred between categories with similar arousal or valence dimensions. For example, 1.13% of HVLA samples were misclassified as HVHA, and 1.75% of LVHA samples were misclassified as LVLA. This aligns with the continuous nature of emotional perception and reveals that the model's decision boundaries are primarily focused on distinguishing the two core dimensions: "high/low arousal" and "positive/negative valence."

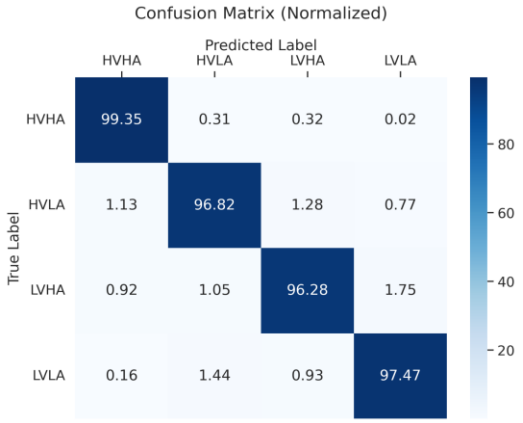

Figure 2. Confusion Matrix Results

The loss and accuracy curves are shown in Figure 3. The training accuracy started from an initial value of approximately 0.21 and increased steadily and monotonically with training epochs, eventually stabilizing above 0.93, demonstrating the model's strong learning and fitting capability. The validation accuracy exhibited a more rapid initial improvement, surpassing 0.81 after about 20 epochs, followed by a slower growth rate and final convergence at a high value of 0.9731. Concurrently, both the training loss and validation loss showed a consistent downward trend, ultimately stabilizing at very low levels. Crucially, the validation loss showed no rebound throughout the training process, and the validation accuracy consistently kept pace with, and in some epochs even exceeded, the training accuracy. This phenomenon strongly indicates that the model did not suffer from overfitting, and its generalization capability was continuously and effectively enhanced as training progressed. In summary, the experimental data—from final performance, detailed category discrimination, to the learning process—comprehensively validates the effectiveness and robustness of the proposed model architecture.

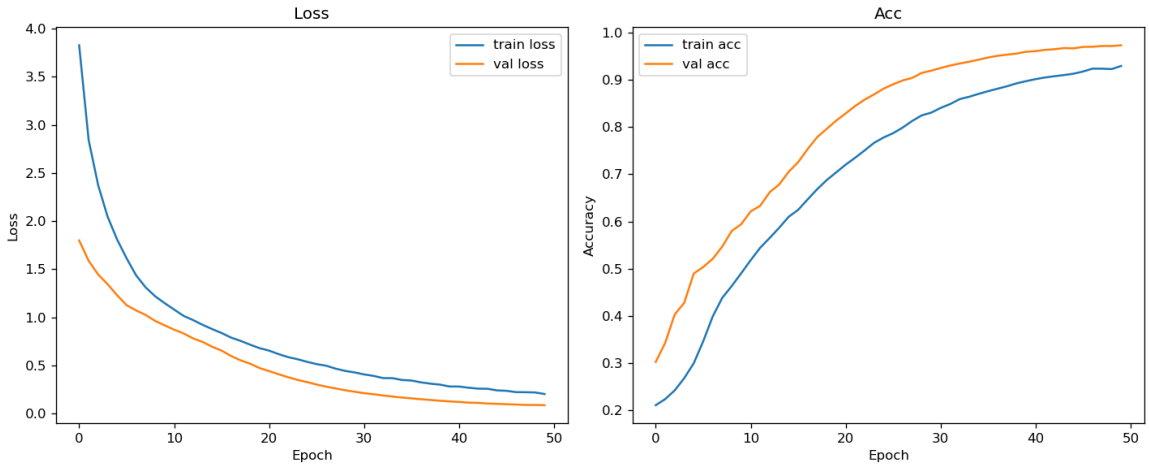

Figure 3. Loss and Accuracy Curves

4.2.2. Ablation Study

To quantitatively evaluate the individual contributions and synergistic effects of each improved component in the model, we designed and conducted a systematic ablation study. Starting from the baseline VSSM model, we incrementally integrated the SE module, the Layer Scale mechanism, and the DSPP module. Their performance was then compared under the same DEAP dataset and experimental settings, with the results presented in Table 3 below.

Table 3. Results of the Ablation Study

Model	ACC	F1
VSSM	0.9403	0.9403
VSSM+SE	0.9589	0.9617
VSSM+SE+Layer Scale	0.9672	0.9684
VSSM+SE+Layer Scale+DSPP	0.9731	0.9756

From the experimental results, it can be clearly observed that each improvement is orthogonal and effectively enhances the model's performance. First, introducing the SE module to the baseline VSSM led to an increase in both Accuracy and F1 Score by approximately 0.0186 and 0.0214, respectively. This significant gain validates the effectiveness of the channel attention mechanism. By dynamically recalibrating the weights of feature channels, it enables the model to adaptively focus on the semantic information most relevant to the discrimination of the current expression, thereby enhancing the discriminative power of the features.

Secondly, adding the Layer Scale mechanism on top of the model already containing the SE module brought about a further performance improvement, with Accuracy and F1 Score reaching 0.9672 and 0.9684, respectively. This enhancement indicates that by performing learnable, channel-wise fine-grained scaling on the deep feature flow, it not only stabilizes the training process but also allows the network to fuse the outputs of complex transformation paths in a more controllable manner. This facilitates learning a better feature representation, contributing to an Accuracy gain of about 0.0083.

Finally, integrating the DSPP module to form the complete MS-VSSM model resulted in the final performance leap, achieving a final Accuracy of 0.9731 and an F1 Score of 0.9756. The introduction of the DSPP module, which explicitly fuses multi-scale contextual information in the early stages of the network, greatly enhances the model's perception of cross-scale facial expression features—from local subtle changes to the global emotional tone. This provides key support for addressing the issue of scale diversity in expression recognition, contributing to an Accuracy improvement of about 0.0059.

In summary, the ablation study strongly demonstrates that the SE module, the Layer Scale mechanism, and the DSPP module each bring substantial performance improvements to the model, and the three are complementary. Their progressive integration ultimately allowed the complete MS-VSSM model to achieve a significant improvement of over 3 percentage points in Accuracy compared to the original VSSM baseline, fully validating the necessity and effectiveness of the improvement scheme proposed in this paper.

4.2.3. Comparative Experiments

To further validate the superiority and competitiveness of the proposed MS-VSSM model on the fine-grained emotion recognition task, we conducted a comprehensive performance comparison with several state-of-the-art visual backbone networks. The models included in the comparison are: the efficient re-parameterized Vision Transformer, RepViT; the model with an explicit dilated multi-scale design, DilatedFormer; and the original VSSM, which serves as the baseline for the improvements in this paper. All models were trained and evaluated under identical conditions—using the same DEAP four-category dataset, preprocessing pipeline, and training strategy. The final comparison of Accuracy and Weighted Average F1 Score is presented in Table 4 below.

Table 4. Comparative Experimental Results

Model	ACC	F1
RepViT	0.9100	0.9079
DilatedFormer	0.9390	0.9352
VSSM	0.9403	0.9403
MS-VSSM	0.9731	0.9756

The experimental results indicate that the proposed MS-VSSM model achieved significantly leading performance on both core metrics, with its accuracy and F1 score reaching 0.9731 and 0.9756, respectively. Compared to the second-best baseline model, VSSM, MS-VSSM achieved a substantial improvement of over 3.2 percentage points in accuracy. This strongly demonstrates that the enhanced design scheme—integrating the SE module, Layer Scale mechanism, and DSPP module—can deeply unleash and optimize the potential of the foundational VSSM architecture for expression recognition tasks from multiple dimensions: channel attention, fine-grained feature scale modulation, and multi-scale context fusion.

In the horizontal comparison among baseline models, VSSM, leveraging the linear-complexity global modeling capability provided by its selective state space model, slightly outperformed DilatedFormer. This result highlights the importance of efficiently capturing long-range spatial dependencies for integrating scattered expression signals. DilatedFormer itself, with its well-designed dilated multi-scale module, also achieved better performance than RepViT, further confirming the value of multi-scale perception in visual tasks. However, models relying solely on a single structural improvement still face performance bottlenecks. The success of MS-VSSM lies not in introducing an entirely new architecture but in applying a systematic, modular enhancement strategy to "finely polish" a foundational model (VSSM) that has already demonstrated potential. This approach enables a breakthrough in recognition accuracy while maintaining its efficient computational characteristics. The comparative experiments fully demonstrate that MS-VSSM possesses the current state-of-the-art comprehensive performance on the four-class emotion recognition task, establishing a powerful new benchmark for vision-based fine-grained affective analysis.

5. Conclusion

Focusing on the characteristics of facial expression features in emotion recognition tasks—such as multi-scale nature, local subtlety, and dynamic variation in channel importance—this paper proposes an enhanced model, MS-VSSM, based on the Visual State Space Model (VSSM), which integrates multi-scale context and channel attention. The core innovations of this model lie in the systematic integration of three key designs: 1) Introducing a path-aware channel attention mechanism (SE-SS2D) into the SS2D module to dynamically calibrate feature channels across different scanning paths, focusing on expression-discriminative information; 2) Embedding a Layer Scale mechanism within the VSS Block to perform fine-grained scale modulation on deep feature flows, enhancing training stability and the flexibility of feature representation; and 3) Embedding a Dense Spatial Pyramid Pooling (DSPP) module at the beginning of each network stage to explicitly fuse multi-scale

context from global to local levels, thereby comprehensively capturing both the macroscopic and microscopic details of expressions. In the four-class emotion recognition experiments on the DEAP dataset, MS-VSSM achieved an accuracy of 0.9731 and a weighted average F1 score of 0.9756, significantly outperforming the original VSSM and comparative models such as RepViT and DilatedFormer. Ablation studies further validated the independent contributions and synergistic effectiveness of each improved component. This research confirms the feasibility of enhancing the base VSSM through channel attention, fine-grained feature scale modulation, and multi-scale fusion, providing a new approach for building efficient and accurate fine-grained visual emotion recognition models.

Although MS-VSSM demonstrated superior performance in the experiments, several directions warrant future exploration. First, the model's generalization capability needs to be validated on more diverse and complex public expression datasets (e.g., AffectNet, RAF-DB) and in continuous dimensional prediction tasks. Second, the current work focuses on static images, while the dynamic evolution of emotions over time is a significant information source. Future work could explore extending the scanning mechanism of SS2D to the spatiotemporal domain or designing efficient video architectures to process temporal expression signals. Furthermore, the DEAP dataset itself contains rich physiological signal modalities. Exploring the effective fusion of MS-VSSM with physiological signal processing networks to construct a multimodal emotion recognition framework is a potential avenue to enhance model robustness and discriminative power. Finally, for practical deployment, further optimizing the model's computational efficiency and memory footprint, and researching its adaptation and acceleration schemes for different edge devices will be key to advancing this technology toward practical application.

Acknowledgements

This project is supported by the Innovative Research Project for Graduate Students of Southwest Minzu University (Project No.YCYB2024069).

References

[1] Liu Y, Tian Y, Zhao Y, et al. VMamba: Visual State Space Model [J]. 2024.

[2] Touvron H, Cord M, El-Nouby A, et al. Three things everyone should know about Vision Transformers [J]. 2022. DOI:10.48550/arXiv.2203.09795.

[3] Touvron H, Cord M, Sablayrolles A, et al. Going deeper with Image Transformers [J]. 2021. DOI:10.48550/arXiv.2103.17239.

[4] Picard R W, Healey J. Affective wearables [J]. Personal Technologies, 1997, 1(4): 231-240.

[5] Zhihang T, Xiaming C, Dazhi J. An Artificial Emotion Model for the Mutual Mapping Between Discrete State and Dimensional Space [J]. Journal of System Simulation, 2021, 33(5): 1062-1069.

[6] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[7] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, 2015, pp. 1–9.

[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[10] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[11] X. Shi et al., “Deep learning for precipitation nowcasting: A benchmark and a new model,” in Proc. Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017, pp. 5622–5632.

[12] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human trajectory prediction in crowded spaces,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, Las Vegas, NV, USA, 2016, pp. 961–971.