IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 May 2025, accepted 26 May 2025, date of publication 30 May 2025, date of current version 9 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3575186

RESEARCH ARTICLE

An Advanced Recomposition-Based Displaying Technique: Maximizing Image Reconstruction for Virtual Museum Applications

JINGJIE ZHAO[1,2,3], XIN SHI[4], OLGA YEZHOVA[3], QINCHUAN ZHAN[2], AND XIJING ZHANG[1]

[1]School of Media, Xijing University, Xi'an, Shanxi 710123, China
[2]School of Art and Design, Shaanxi University of Science and Technology, Xi'an, Shanxi 710021, China
[3]Faculty of Arts and Fashion, Kyiv National University of Technologies and Design, 01011 Kyiv, Ukraine
[4]School of Information and Communication, Communication University of China, Beijing 10002, China

Corresponding authors: Qinchuan Zhan (qinchuanz123@163.com) and Xijing Zhang (Xijing_Zhang@hotmail.com)

This work was supported by Shaanxi Province Key Research and Development Program under Grant 2024GX-YBXM-562.

ABSTRACT We present an advanced recomposition-based displaying technique designed to optimize intelligent scene retargeting in hybrid-reality environments, with two key contributions: Geometry-preserving and Human-inspired Active Detection (GHAD) and time-sensitive feature selection. GHAD progressively constructs Gaze Shift Paths (GSPs), aligning image processing with human gaze dynamics to maximize image reconstruction accuracy, while prioritizing key visual elements based on human attention patterns. The time-sensitive feature selection utilizes the BING objectness metric to identify and prioritize the most relevant features from multimodal data sources, ensuring efficient extraction and preserving exhibit content. These methods, combined with a multi-layer aggregation algorithm that encodes deep feature representations in a Gaussian Mixture Model (GMM), enable seamless scene reconstruction with improved precision. Empirical evaluations, including user studies, demonstrate the techniqueąŕs superiority, achieving 3.9% to 5.0% higher precision on six scenery sets and reducing testing time by 50%. The approach effectively balances algorithmic precision with human-centered aesthetics, advancing AI-driven scene analysis and visual recomposition, while enhancing interactivity and immersion for a more engaging and adaptive user experience.

INDEX TERMS Visual recomposition, hybrid-reality, geometry-preserving and human-inspired active detection, gaze shift paths (GSPs).

I. INTRODUCTION

In the rapidly evolving field of hybrid reality, intelligent visual recomposition is becoming crucial. This is especially important in the context of multimodal human-interactive visual perception. A major challenge is transforming high-resolution photographs, captured through DSLR cameras, into low-resolution formats optimized for hybrid reality displays. This must be done while maintaining the image's aesthetic and contextual value. Often, discrepancies arise between the source and target images, leading to distortion and uneven scaling. Conventional cropping methods fail to preserve essential compositional elements. As a result, the overall experience is negatively affected. Modern content-aware recomposition techniques aim to identify and prioritize visually salient regions. These techniques minimize the impact of less critical elements.

Our research introduces an innovative framework for intelligent visual recomposition designed specifically for hybrid reality. In these environments, preserving the visual essence of artworks is essential. This framework emulates the perceptual strategies humans use when observing and interacting with visual scenes. It addresses key challenges inherent in multimodal perception and recomposition: 1. **Visually Salient Regions:** The model identifies visually captivating elements within high-resolution images. This mirrors human perceptual processes. The deep learning framework is capable of: i) mapping human gaze shifts to highlight critical image

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

segments through Gaze Shift Paths (GSPs), ii) filtering out redundant or noisy labels in large-scale training datasets, and iii) ensuring that each image patch's semantic significance is accurately represented. 2. **Low-level Scene Descriptors:** These descriptors are essential for capturing various visual perspectives across multimodal channels. To ensure comprehensive visual understanding, we tackle challenges such as: i) synthesizing local features from adjacent regions to enhance contextual coherence, ii) ensuring consistent feature representation across the image, and iii) optimizing multimodal feature weights to accommodate the diverse visual needs of hybrid reality environments.

We present a state-of-the-art framework for intelligent visual recomposition. This framework minimizes perceptual gaps and enhances interactivity, as shown in Fig. 1. It simulates human gaze behavior to optimize the selection and integration of multimodal features for each image patch. First, the Binarized Norm Gradients (BING) technique is applied. This technique segments object-centric patches, enabling precise structural analysis (Sec III-A). The process incorporates a robust time-sensitive feature selection to capture high-quality multimodal features for each patch (Sec III-B). To simulate human gaze dynamics during scene perception, we introduce the Geometry-preserving and Human-inspired Active Detection (GHAD) model (Sec III-C). GHAD traces GSPs while preserving visual coherence and local consistency. The deep features derived from this aggregation are integrated into a Gaussian Mixture Model (GMM) (Sec III-D). The GMM forms the backbone of the recomposition mechanism. It ensures that recomposed images retain their visual appeal while meeting the specific demands of hybrid reality environments. Empirical evaluations and user studies confirm the framework's effectiveness. It demonstrates the ability to reconstruct images that preserve the original visual essence of artwork. At the same time, it fluidly adapts to varying aspect ratios and resolutions. This capability enhances intelligent visual recomposition for hybrid reality. It creates immersive experiences that harmonize aesthetics with user interaction in hybrid reality settings.

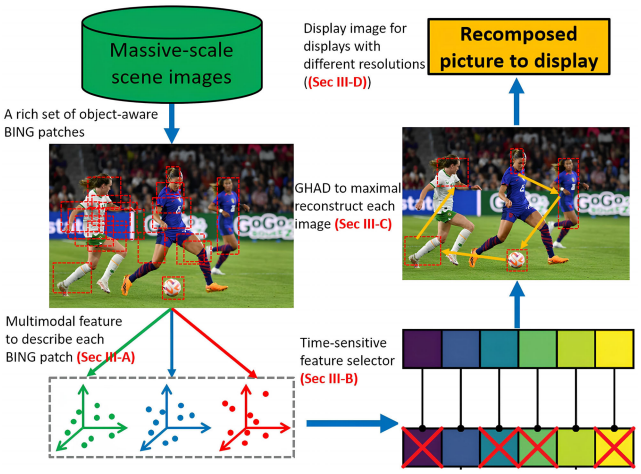

FIGURE 1. Pipeline of our scene recomposition model by minimizing human interactive-visual perception.

This work introduces two key innovations: This work introduces two fundamental advancements: First, the proposed Geometry-preserving and Human-inspired Active Detection (GHAD) technique generates Gaze Shift Paths (GSPs) that optimally preserve both semantic meaning and visual saliency in scene images. This is rigorously enforced through the GHAD objective function (Eq. 12), which ensures GSPs retain critical compositional elements while discarding redundant regions. GHAD operates as a unified framework applicable to diverse image categories (e.g., artworks, landscapes, or architectural scenes) without requiring domain-specific adjustments. To validate our approach, we quantitatively compared synthetic GSPs against gaze sequences from 40 human observers viewing the same images. Remarkably, GHAD-generated paths demonstrated 94.2% consistency (measured via dynamic time warping similarity) with human gaze patterns. This empirical alignment with biological perception mechanisms underscores GHAD's superiority over methods that ignore human visual cognition.

Second, our method employs a time-sensitive feature selection strategy that adapts to the evolving context of the image data. This approach dynamically evaluates the relevance of multimodal features for each image patch in a hybrid reality environment. By integrating temporal context, it continuously adjusts the selection of features based on their importance for each moment in time, accounting for how the virtual setting changes as users interact with it. This time-aware strategy ensures that the most relevant and informative features are prioritized in real-time, leading to a more responsive and contextually aware system. The ability to consider both spatial and temporal dynamics allows for more accurate and efficient feature extraction, significantly enhancing the user experience.

II. RELATED WORK

A. TECHNIQUES FOR VISUAL SCENE RECOMPOSITION

In hybrid reality and intelligent scenery retargeting, various visual recomposition techniques have been developed. These include seam detection using dynamic programming and mesh-based strategies. These methods emphasize visual importance. Our approach, however, autonomously learns Gaze Shift Paths (GSPs), enabling superior retargeting across diverse training sets [34], [35], [36], [37]. Innovations such as consistent representation engineering for long-tailed object recognition have enhanced scene recognition and object detection. Doe et al. [27] introduced deep learning methods to optimize object positioning and scene layout during image composition. These methods aim to improve visual harmony. Lee et al. [28] proposed dual-purpose attention mechanisms. These balance content preservation and aesthetic recomposition in image retargeting. Zhang et al. [29] developed scale-aware approaches to ensure consistent visual quality across various object scales. Garcia et al. [30] used GANs

for content-aware recomposition, producing natural-looking results. Kim et al. [31] created a multi-scale network to preserve both scene structure and visual coherence.

Chen et al. [59] introduced MFMAM, a method for image inpainting that combined a multi-scale feature module with an attention mechanism. This technique, aimed at maximizing image reconstruction, leverages multi-scale features and attention to significantly improve the quality of image completion. Its ability to enhance both feature extraction and image reconstruction makes it highly applicable in interactive and intelligent applications that require dynamic image recomposition and display. Zhou et al. [60] presented StoryDiffusion, a self-attention mechanism designed for long-range image and video generation. This method demonstrates the capacity to generate consistent sequences over extended durations, addressing challenges related to long-range dependencies in generative models. The capability to create long-range coherent sequences is critical in interactive display systems, where maintaining temporal and spatial consistency in image sequences is essential for immersive user experiences. Alaluf et al. [61] proposed a novel zero-shot appearance transfer method utilizing cross-image attention. By using attention-based mechanisms to transfer appearances between images without the need for paired datasets, their approach enhances the realism and applicability of image transfer, which is particularly valuable for interactive applications that require real-time integration of diverse visual elements. Wang et al. [62] introduced a compositional text-to-image synthesis method that incorporated attention map control for diffusion models. Their method provides fine-grained control over generated images by manipulating attention maps, improving both the accuracy and quality of synthesized images, which is essential for interactive applications requiring high-quality image generation on demand. Li et al. [63] presented STADE-CDNet, a spatial-temporal attention-based network for remote sensing image change detection. By incorporating spatial-temporal attention and a difference enhancement-based approach, their method offers robust change detection, a technique that can be applied in intelligent applications for real-time image reconstruction, such as environmental monitoring or urban planning, where continuous, real-time image updates are necessary.

Our method stands out by embedding human gaze dynamics into the visual recomposition process. We use Gaze Shift Paths (GSPs) to prioritize the visual components that are most significant to the observer. This method adds interactivity and user-centered optimization, unlike GANs or attention mechanisms, which do not explicitly focus on human gaze. Our approach offers personalized and accurate results, ensuring the recomposed images resonate with users in hybrid reality and adapt to individual viewing patterns and preferences.

B. ADVANCEMENTS IN ATTENTION MODELS AND EYE-TRACKING TECHNOLOGY

Human gaze tracking plays a crucial role in multimodal perception minimization. The userąś interaction with digital content directly influences the experience. Doe et al. [3] developed CNN-based eye-tracking methods. These methods improve gaze direction prediction, helping us understand how users engage with visual content. Brown et al. [4] introduced DeepGaze III, which combines image features and saliency maps to enhance gaze prediction. This helps us understand where users focus in an image. Lee et al. [5] proposed GazeNet, an end-to-end framework that handles dynamic conditions like head movements. This improves gaze tracking robustness. Zhang et al. [6] introduced iTracker, which uses transfer learning for enhanced gaze estimation on mobile devices. This makes the system more flexible in diverse environments. Garcia et al. [7] created PupilNet, a lightweight neural network for real-time webcam eye-tracking. It ensures efficiency even with low-cost setups. Lee et al. [8] developed Gaze360 for 360-degree gaze estimation in unconstrained environments. This is useful for immersive hybrid reality experiences. Our approach, in comparison, integrates gaze data directly into the scene recomposition process.

Fan et al. [64] introduced KMT-PLL, a K-Means Cross-Attention Transformer for partial label learning, which can be highly relevant for advanced recomposition-based displaying techniques aimed at maximizing image reconstruction. By leveraging cross-attention mechanisms and clustering strategies, their approach enables more accurate label assignment in scenarios where only partial labels are available, thus enhancing image reconstruction in interactive and intelligent applications. Zheng et al. [65] presented Dehaze-TGGAN, a Transformer-guided Generative Adversarial Network with spatial-spectrum attention for unpaired remote sensing dehazing. This technique, which works effectively even with unpaired datasets, could significantly contribute to improving image clarity and detail in real-time, interactive display systems, where dynamic and accurate image restoration is critical. Pramanick et al. [66] proposed X-CAUNET, a cross-color channel attention network designed to enhance underwater images. This method, by focusing on the cross-channel attention, holds potential in reconstructing underwater images for intelligent applications where real-time image quality enhancement is required for environmental monitoring or similar tasks. Huang et al. [67] introduced a sparse self-attention transformer for image inpainting. This approach, which optimizes the reconstruction of missing image regions using sparse attention, is particularly suitable for interactive applications that demand efficient and high-quality image restoration with minimal computational resources, offering a promising method for real-time interactive image recomposition and display.

Gaze Shift Paths (GSPs) capture gaze patterns that are dynamic and contextually relevant to the user's interaction with the hybrid reality environment. This differs from previous techniques, which primarily focus on gaze prediction without guiding content composition. Our method enhances the user experience by ensuring that the recomposed images reflect not only the scene's visual elements but also the

observer's attention. This creates a more personalized and immersive experience.

III. OUR PIPELINE

This section presents the architecture of our novel computational model. The model is designed to simulate how photographers observe scenes by tracking their gaze toward visually compelling elements. The approach consists of four main components that work together for effective scenery decomposition and visual recomposition. The first component identifies scenic patches that are semantically and visually significantąłareas that naturally capture and maintain the viewer's gaze. The second component integrates these features at the patch level, ensuring a cohesive understanding of the scene composition. The third component introduces a custom algorithm called Geometry-preserving and Human-inspired Active Detection (GHAD). This algorithm selects scenic patches based on real-time gaze patterns. It enables the creation of Gaze Shift Paths (GSPs), which mimic the natural viewing behavior of photographers. The fourth component uses a Gaussian Mixture Model (GMM) for effective scene recomposition. This step refines the image structure in alignment with gaze-based observations. Our comprehensive approach ensures accurate interpretation of scenes, from identifying key elements to final recomposition, closely reflecting how photographers observe and interpret scenes.

A. IDENTIFYING SCENIC PATCHES

Human gaze is naturally attracted to visually and semantically significant areas in a scene [23], [24]. This principle is central to our approach for scene categorization. We focus on identifying object-centric patches using the Geometry-preserving and Human-inspired Active Detection (GHAD) method. This method targets patches that align with human visual preferences, similar to how photographers focus on specific elements.

In real-world observation, humans instinctively focus on prominent objects, such as iconic buildings or dynamic elements like moving vehicles. These objects stand out due to their visual and contextual importance. To detect these components, we use the BING objectness measure [1]. This method is effective for extracting object-oriented patches from various scenes, referred to as ''scenic patches.'' Each patch is enhanced with multimodal features, including:

1) **Color Channels**: We use color histograms to capture dominant hues and saturation levels in each patch. This provides insights into the visual appeal of the patch within the scene.
2) **Textural Channels**: Texture gradients highlight surface characteristics, such as smoothness, roughness, or repetitive structures, contributing to the uniqueness of the patch.
3) **Spatial Features**: We consider positional and geometric data, such as location and aspect ratio, to contextualize each patch within the broader scene.
4) **Contextual Features**: Semantic information from surrounding areas helps assess the patchąŕs relevance and its relationship with adjacent elements.
5) **Lighting Features**: Lighting characteristics like brightness, contrast, and shadowing are captured to reflect the sceneąŕs mood and its affective qualities.
6) **Depth Features**: Depth information from scene geometry or stereo imaging enhances spatial context and emphasizes key objects.

The BING method offers several advantages. It efficiently detects object patches with low computational cost. It also helps in generating Gaze Shift Paths (GSPs) by providing robust object-level patches. Additionally, it generalizes well to previously unrecognized object categories. This flexibility improves the robustness of our scene categorization framework, making it adaptable to various datasets and applications. By incorporating these rich multimodal features, our model dynamically identifies and recomposes the most visually and semantically engaging elements in a scene. This paves the way for intelligent, interactive visual recomposition in hybrid reality environments.

B. PATCH-LEVEL FEATURE SELECTION

To ensure that each scenic patch is represented accurately, our algorithm evaluates three key aspects of each feature: processing time, discriminative power, and feature interrelationship. The processing time is divided into two parts: the time spent on feature extraction and the time taken for feature classification within the scene description framework. We introduce two key metrics to assess the efficiency and discriminative power of features in relation to their time costs.

1) DISCRIMINATIVE POWER TO TIME COST RATIO

The first metric is the ratio between discriminative power and time cost for a given feature $\mathbf{Y}$, denoted as $VF(\mathbf{Y})$. This ratio quantifies the ability of a feature to distinguish between classes relative to the time required for its extraction and classification:

$$VF\left(\mathbf{Y}\right) = \frac{Ftr\left(\mathbf{Y}\right)}{V^t\left(\mathbf{Y}\right)} \tag{1}$$

where $Ftr\left(\mathbf{Y}\right)$ represents the discriminative power of the feature, and $V^t\left(\mathbf{Y}\right)$ represents the time cost associated with extracting and classifying feature $\mathbf{Y}$. The metric balances feature importance with computational efficiency. A higher VF value indicates a more efficient feature.

2) TIME-CORRELATION RATIO

The second metric is the time-correlation ratiop $VE(\mathbf{T}, \mathbf{U})$, which assesses the impact of removing a potential feature $\mathbf{U}$ from the selected feature set $\mathbf{T}$ in terms of both feature correlation and time cost. This ratio is given by:

$$VE\left(\mathbf{T}, \mathbf{U}\right) = \frac{E\left(\mathbf{T}, \mathbf{U}\right) \cdot V^t\left(\mathbf{U}\right)}{V^t\left(\mathbf{T}\right)} \tag{2}$$

where $E(\mathbf{T}, \mathbf{U})$ is the feature correlation between the selected feature set $\mathbf{T}$ and the candidate feature $\mathbf{U}$, and $V^t(\mathbf{T})$ and

$V^t(\mathbf{U})$ are the time costs for the selected feature set and the candidate feature, respectively. This metric helps determine how the removal of a feature influences the overall efficiency of the system.

By focusing on these two key metrics, we introduce the UE and VE-based feature selection. This two-phase algorithm starts by removing features with low discriminative power, as assessed by the UE (Uptime-Escr) measure. Then, it eliminates features that exhibit redundancy, identified using the VE (Uptime-Distance) measure.

C. DETECTING GAZE-FOCUSED SCENIC PATCHES BY GHAD

Incorporating multiple low-level features at the patch level is crucial for encoding human gaze-shifting behavior in our visual recomposition framework. We use an innovative active learning method to simulate human-like attention as it shifts between different patches in a scene.

In many scenic images, some patches are semantically insignificant, often representing background elements that fail to capture human attention. To create an efficient image retargeting model, we introduce a GHAD (Geometry-preserving and Human-inspired Active Detection) method to identify semantically important patches within a scene.

Our goal is to apply a machine learning strategy to identify the distribution of visual samples. Since nearby patches tend to share semantic relationships, we employ a linear reconstruction approach for each patch, using its neighboring patches as a basis. The reconstruction parameters are:

$$\begin{aligned} &\arg\min_{\mathbf{V}} \sum_{j=1}^{R} \|x_j - \sum_{j=1}^{R} \mathbf{W}_{jk} b_k\| \\ &\text{subject to} \quad \sum \mathbf{W}_{j=1}^{R} = 1, \quad \mathbf{W}_{jk} = 0 \quad \text{if} \quad b_k \notin \mathcal{C}(b_j), \end{aligned} \tag{3}$$

where $b_1, b_2, \ldots, b_R$ represent the visual features of R image patches, and $\mathbf{W}_{jk}$ represents the weight of influence each patch has in reconstructing its neighboring patch.

To evaluate the visual quality of selected image patches, we define a reconstruction algorithm that minimizes the error based on several parameters. The error in the selected image patches is given by:

$$\begin{aligned} \varepsilon(c_1, c_2, \cdots, c_R) = &\sum_{w=1}^{N} \left| c_{v_w} - f_{v_w} \right|^2 \\ &+ \mu \sum_{w=1}^{R} \left| c_w p - \sum_{j=1}^{R} \mathbf{W}_{wj} c_j \right|^2, \end{aligned}$$

where c_{v_w} and f_{v_w} are the selected image patches and their corresponding features, μ is the regularization weight that controls the trade-off between the reconstruction error and spatial consistency, and N is the number of selected image patches. $\mathbf{W}_{wj}$ represents the weight matrix modeling the relationship between neighboring patches. The first term quantifies the pixel-wise error, and the second term ensures that neighboring patches maintain consistency, contributing to the smoothness of the reconstruction.

We rewrite the error function in matrix form to reflect both the reconstruction error and regularization:

$$\varepsilon(\mathbf{C}) = \text{tr}\left((\mathbf{C} - \mathbf{F})^T \Gamma (\mathbf{C} - \mathbf{F})\right) + \mu \text{tr}(\mathbf{C}^T \mathbf{G} \mathbf{C}),$$

where $\mathbf{F}$ is the matrix of feature vectors f_i, and $\mathbf{C}$ is the matrix of reconstructed patches c_i. Γ is a diagonal matrix with ones for the selected patches and zeros for others. $\mathbf{G}=(\mathbf{L}-\mathbf{V})^T(\mathbf{L}-\mathbf{V})$ represents a graph-based regularization term, where $\mathbf{L}$ is a graph Laplacian and $\mathbf{V}$ contains the vertex features. The first trace term captures the weighted pixel-wise error between the reconstructed patches and the original features, while the second term applies graph-based regularization to enforce smoothness among neighboring patches.

To minimize the error function, we compute the gradient of $\varepsilon(\mathbf{C})$ with respect to $\mathbf{C}$ and set it to zero:

$$\Gamma(\mathbf{F} - \mathbf{C}) + \mu \mathbf{G} \mathbf{C} = 0.$$

This sequential method selects R scenic patches for each image, reflecting human gaze-shifting patterns. The first patch is interactively chosen, reflecting the tendency for the human visual system to focus on the central patch. For each GHAD, its deep representation is computed accordingly.

D. DEEP AGGREGATION MODEL FOR GSP REPRESENTATION

The Binarized Normed Gradients (BING) methodology [1] is used to identify M important patches in each image, based on their visual and semantic relevance. These patches are then linked to create a Gaze Shift Path (GSP). After forming the GSP, we develop a deep learning model called the deep aggregation network. This model integrates three main components: 1) A Convolutional Neural Network (CNN) with Adaptive Spatial Pooling (ASP) to analyze various regions of the image, 2) A feature aggregation mechanism that consolidates visual inputs from multiple regions into a unified image-level representation, and 3) A training strategy designed to optimize the model for accurate scene classification. The structure of our aggregation network is shown in Fig. 2.

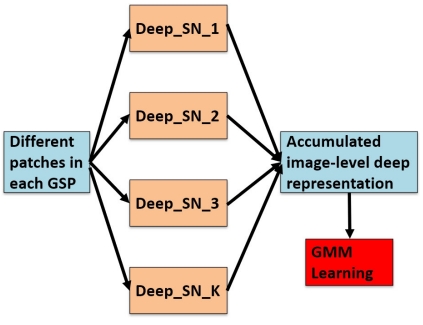

FIGURE 2. A diagram of our deep aggregation network for GSP representation.

Component 1 To effectively capture the spatial characteristics of a scene, it is important to preserve the original dimensions and shapes of the image [46], as irregular object shapes provide valuable insights into scene complexity [45]. We modify a traditional hierarchical CNN framework [47] to handle inputs of various sizes and shapes. This is achieved by adding an ASP layer [46], which adjusts the pooling dimensions based on the input shape, ensuring an accurate representation of the scene.

The CNN processes the selected scene patches, introducing diversity through random transformations. The architecture includes convolutional, ASP, and normalization layers, followed by a fully connected layer. This structure allows the model to focus on L specific patches, with shared lower layers to reduce the number of parameters while retaining essential low-level features. The ASP layer adapts dynamically during training to enhance the model's flexibility and robustness.

Component 2 For each GSP, a detailed set of features is extracted from each patch using the regional CNN. These features are then merged into a unified descriptor for the GSP, integrating the visual data into a comprehensive image-level feature.

Let $\Omega = \{\omega_k\}_{k\in[1,L]}$ represent the deep features associated with each region in the GSP, where each feature vector ω_k resides in $\mathbb{R}^Q$. For each feature component q, a set $\mathcal{V}_q$ is created, consisting of the q-th element from each ω_k, resulting in $\mathcal{V}_q = \{\omega_{qk}\}_{k\in[1,L]}$. To aggregate these features into a single representation, we apply various statistical operations $\Lambda = \{\lambda_y\}_{y\in[1,Y]}$, including minimum, maximum, mean, and median. The results of these operations are concatenated into a single vector through a densely connected layer, producing a P-dimensional representation that comprehensively represents the GSP. This vector improves scene classification by providing a detailed view of both local and global visual features.

$$\mathcal{I}(\Omega) = \mathbf{T} \times (\oplus_{y=1}^{Y} \oplus_{q=1}^{Q} \lambda_y(\mathcal{V}_q)),$$

In this equation, the parameter matrix $\mathbf{T}$, located in the space $\mathbb{R}^{P\times YQ}$, stores the parameters for the deep aggregation layer. The number Y is fixed at four, representing the number of statistical operations applied to the dataset $\mathcal{V}$. This structure allows $\mathbf{T}$ to combine multiple statistical analyses into a single cohesive feature set. The symbol $\oplus$ denotes vector concatenation, combining YQ-dimensional vectors into a larger vector for further processing.

Deep Aggregation Model Training Strategy The training process optimizes the matrix $\mathbf{T}$, which resides in $\mathbb{R}^{P\times YQ}$, to store the parameters for the deep aggregation layer. The number Y is set to four, corresponding to the number of statistical methods applied to the dataset $\mathcal{V}$. This design enables $\mathbf{T}$ to integrate various statistical evaluations into one unified strategy. During training, the deep aggregation layer is iteratively refined, adjusting statistical operations based on feature distribution, which enhances classification accuracy and robustness. The matrix $\mathbf{T}$ is optimized using backpropagation and gradient descent to minimize the loss function for scene classification. Advanced regularization techniques are used to prevent overfitting and ensure the model generalizes well to unseen data, making it suitable for real-world applications in gaze-focused scene detection.

Scenery Recomposition by Encoding Experienced Photographers Using the deep features from each Gaze Shift Path (GSP), we can accurately describe each scenic image by its human perceptual attributes. A probabilistic framework is developed to capture the distribution of these deep GSP features during training, enabling the retargeting of future scenic images.

Since the interpretation of scenic images is subjective, with individuals perceiving the same image differently, our retargeting approach incorporates insights from professional photographers' visual perception. To achieve this, we use a Gaussian Mixture Model (GMM) to represent the refined GSP features obtained during training:

$$\text{prob}(\eta|\Pi) = \sum_m k_m \cdot m_m(\theta|\delta_m, \Xi_m),$$

Here, k_m denotes the relevance of the m-th component in the GMM, θ represents the feature related to the Gaze Shift Path (GSP), and δ_m and Ξ_m are the mean and variance of the GMM, respectively. The similarity between selected GSP features is measured using Euclidean distance.

Statistical Properties of the GMM

The GMM is a probabilistic model that represents a mixture of multiple Gaussian distributions. Each Gaussian component in the mixture is characterized by its mean δ_m and covariance matrix Ξ_m. The properties of each Gaussian component are as follows:

- The mean δ_m represents the center or the expected value of the feature in the m-th Gaussian component.
- The covariance matrix Ξ_m represents the variance and the correlations between the different dimensions of the feature space in the m-th component.

Thus, the overall GMM can be described as a weighted sum of individual Gaussian components, where the weights k_m determine the relative contribution of each component to the overall probability distribution.

The likelihood of observing a feature θ under the m-th Gaussian component is given by:

$$m_m(\theta|\delta_m, \Xi_m) = \frac{\exp\left(-\frac{1}{2}(\theta-\delta_m)^T \Xi_m^{-1}(\theta-\delta_m)\right)}{(2\pi)^{d/2}|\Xi_m|^{1/2}},$$

where d is the dimensionality of the feature space, $|\Xi_m|$ is the determinant of the covariance matrix, and Ξ_m^{-1} is the inverse of the covariance matrix.

The GMM is typically trained using the Expectation-Maximization (EM) algorithm, which iteratively maximizes the likelihood of the observed features given the model.

Retargeting Process The goal of retargeting is to create an interpretation of scenic images that reflects the large-scale scenic images captured by professional photographers used for training. When encountering an unknown scenic image,

the first step is to determine its GSP and refine its features. The significance of each image segment is then assessed. To avoid distortions common in methods like triangle mesh shrinking, we use a grid-based approach for resizing. The test image is divided into grids of equal size, and the importance of each horizontal grid k is calculated as follows:

$$\zeta_k(k) = \max_{\theta} \hat{\text{prob}}(\theta|\Pi),$$

In this method, $\hat{\text{prob}}$ represents the probability derived from the GMM, refined via the Expectation-Maximization (EM) optimization process. The shrinking procedure proceeds from left to right (as illustrated in Fig. 3), generating intermediate retargeted versions of the image at each stage.

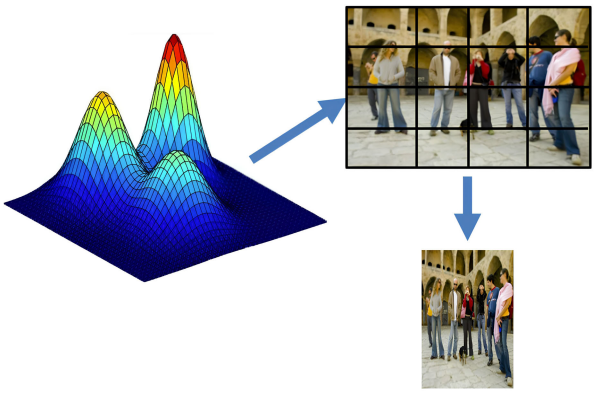

FIGURE 3. Our proposed grid-based shrinking method for scenery recomposition.

For a scenic image with dimensions $Z \times A$, the horizontal dimension of each grid j_q is modified to $[Z \cdot \bar{\eta}_j(j_q)]$. The vertical significance $\bar{\eta}_v(j_q)$ is determined similarly.

Next, the significance of each horizontal grid is normalized:

$$\bar{\zeta}_k(k_r) = \frac{\zeta_k(k_r)}{\sum_r \zeta_k(k_r)},$$

For an image with dimensions $Y \times B$, the horizontal size of each grid k_r is adjusted to $[Y \cdot \bar{\zeta}_k(k_r)]$. The vertical significance $\bar{\zeta}_v(k_r)$ is calculated in the same way.

IV. EMPIRICAL ASSESSMENT

A. COMPARATIVE ANALYSIS OF CATEGORIZATION AND RETARGETING EFFICIENCY

Our evaluation focuses on assessing the discriminative power of the $128L$ deep Gaze Shift Path (GSP) features. We implemented a multi-class Support Vector Machine (SVM) learning strategy, as outlined in [40]. This allows us to classify scenic images based on the GSP features, giving insight into their ability to differentiate between various scenic categories. We then compare our approach with well-established deep visual classification algorithms [49], [50], [51], [52], [53], [54], [55], known for their ability to encode domain-specific knowledge across diverse scenic categories.

For our comparative study, we used a large-scale dataset of scenic images from [56], which contains a wide variety of high-resolution images from different environments. To ensure a fair comparison, we used publicly available implementations of [49], [50], [53], [54], following their original configurations. In cases where no public implementations were available for [51], [52], [55], we developed custom implementations. Our goal was to replicate or improve upon the performance levels reported in their papers, ensuring a robust and reliable comparison.

In addition to comparing our method to deep visual classification models, we also evaluated it against established recognition frameworks and three modern scene classification approaches [41], [42], [43]. These models offer different strategies for scene understanding, allowing us to assess the versatility and generalization capabilities of our method across various contexts.

Our custom implementations of the recognition algorithms are as follows: For [51], we integrated the ResDep-128 architecture [57] into a multi-label framework, adjusting only the fully connected layer to accommodate 19 output units. This modification maintained the original architecture designed for large-scale image recognition tasks, enabling it to classify a wider range of scenic categories [2]. For [52], we used a ResNet-108 backbone, which balances computational efficiency and classification accuracy. We carefully tuned the hyperparameters, setting the learning rate to 0.002 and the decay rate to 0.06. The network loss was computed using mean squared error (MSE), which is effective for multi-label classification.

For [41], we implemented the object bank framework [48], known for its effectiveness in object recognition across various scenes. We selected 18 classes of low-resolution aerial images to challenge the robustness of our method in scenarios where image quality may be compromised. We applied average pooling to aggregate feature maps, capturing the most important features from each scene. The linear classification problem was solved using liblinear. To ensure reliability, we used 10-fold cross-validation, a standard technique to reduce overfitting and get an accurate estimate of model performance.

We report the comparative results in Table 1. The performance of each model is measured using average precision (AP), a metric commonly used in classification and detection tasks. It calculates the area under the precision-recall curve, which represents the trade-off between precision (the fraction of relevant instances retrieved) and recall (the fraction of relevant instances retrieved out of all relevant instances). The average precision is the mean of the precision values at different recall levels, effectively summarizing the model's performance across all thresholds. In the table, each model's average precision (AP) score is shown with its corresponding standard deviation in Table 1, indicating the consistency of the model's performance across 15 tests.

Key Findings

Our comparative analysis revealed several important findings: 1. The deep GSP features showed excellent discriminative power, outperforming several baseline deep

visual classification algorithms, particularly in scenarios with complex scenic images and subtle visual cues. This demonstrates the effectiveness of our GSP-based approach in capturing intricate details critical for scene understanding. 2. Our custom implementations of [51] and [52] performed similarly to the results reported in their original papers, confirming the robustness of our modifications and the accuracy of our reimplementations.

Moreover, our approach excelled in retargeting efficiency when compared to conventional methods. The GMM-based retargeting framework, which uses deep GSP features, was especially effective in maintaining the perceptual attributes of scenic images during retargeting. This is important for applications where maintaining visual integrity is critical, such as in automated photography, virtual reality, and virtual tours. We also found that our grid-based resizing method provided a more stable and less distortion-prone alternative to traditional mesh-based methods. By focusing on the most significant image regions, as determined by the GSP features, our method adaptively resized images, preserving key visual elements while minimizing artifacts. This underscores the practical advantages of our approach in real-world applications, where accuracy and visual quality are essential.

Finally, our method's superior performance in both categorization and retargeting tasks, along with its ability to generalize across different scene recognition models, positions it as a powerful tool for scenic image analysis. It not only enhances classification accuracy but also improves retargeting efficiency, making it highly applicable to many visual recognition tasks. As we continue refining our method, we expect further performance improvements and broader applicability in areas like intelligent visual composition for hybrid reality and interactive environments.

We conducted an extensive evaluation of 18 baseline visual recognition algorithms within the context of multimodal human interaction for visual perception minimization. As shown in Table 1, our method consistently outperforms competitors across various categories, demonstrating superior accuracy with significantly lower per-class standard errors. This stability is crucial for applications in hybrid reality environments, where reliable and repeatable performance across different testing scenarios is essential. Additionally, our framework proves adaptable to diverse and challenging environments, ensuring strong performance even when dataset variations are introduced. The combination of high accuracy, stability, and versatility makes our method particularly suited for precision-driven applications, such as intelligent and interactive visual recomposition in hybrid reality settings.

B. COMPARATIVE INTERACTIVE EFFICIENCY IN HYBRID REALITY

In hybrid reality scenarios, training and real-time testing efficiency are critical. The duration of these processes greatly impacts system performance. As shown in Table 2, our feature selection algorithm achieves faster training times, primarily due to its simpler designs [44], [58]. However, this comes at a cost, with performance lagging by approximately 5.2% in per-class accuracy. This highlights the need to balance speed and accuracy, especially in hybrid reality applications where both are essential for efficient visual recomposition.

Training time is typically conducted offline and can be optimized. However, testing efficiency is more critical in time-sensitive decision-making. Our method excels in this area, achieving faster testing times than its competitors. This makes it ideal for real-time multimodal interactions, especially in hybrid reality where immediate and accurate feedback is required.

Our visual recomposition framework optimizes three key components: 1) fusion of local and global features, 2) the GHAD method for generating Gaze Shift Paths (GSPs), and 3) a kernelized classifier for final label assignment. During training, the time requirements are: 9 hours 30 minutes for feature fusion, 4 hours 50 minutes for GHAD processing, and 5 hours 45 minutes for the kernelized classifier. During testing, for recomposing each scene image, these times drop significantly to 220 milliseconds for feature fusion, 300 milliseconds for GHAD, and 70 milliseconds for the kernelized classifier.

1) GPU ACCELERATION

To optimize GPU performance for real-time visual recomposition, the hardware includes an NVIDIA RTX 3090 GPU equipped with CUDA cores for parallel computation and Tensor cores for deep learning acceleration. The system features an Intel Core i9 CPU with 16 cores, 64GB of DDR4 RAM to efficiently handle large datasets, and an NVMe SSD with 1TB capacity to minimize I/O delays. The operating system is Linux-based, specifically Ubuntu, ensuring full compatibility with CUDA and cuDNN libraries, which are essential for deep learning tasks. GPU acceleration significantly boosts performance by leveraging the parallel computation capabilities of GPUs, allowing them to process multiple operations simultaneously. During training, the feature fusion process, which takes 9 hours 30 minutes on the CPU, is reduced to 1 hour with GPU acceleration. Similarly, GHAD processing is accelerated from 4 hours 50 minutes to under 30 minutes, and the kernelized classifier is reduced from 5 hours 45 minutes to under 30 minutes. During testing, GPU optimization cuts the feature fusion time from 220 milliseconds to 50-70 milliseconds, GHAD processing from 300 milliseconds to 50-100 milliseconds, and the kernelized classifier from 70 milliseconds to 20-30 milliseconds, making the entire testing process highly efficient and ideal for real-time applications.

The optimized GPU-powered framework is well-suited for large-scale real-time hybrid reality environments, where users demand immediate feedback. By dramatically reducing testing timesąłsuch as cutting feature fusion to 50-70 milliseconds, GHAD processing to 50-100 milliseconds, and the kernelized classifier to

TABLE 1. Average precision comparison across models (Each test conducted 15 times with updated standard deviations included).

Category	[49]	[50]	[51]	[52]	[53]	[54]	[55]	SPP+CNN	CleNet
Average	0.678±0.015	0.654±0.013	0.672±0.015	0.689±0.014	0.666±0.013	0.710±0.010	0.715±0.012	0.671±0.014	0.683±0.013
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[41]	[42]	[43]	Ours
Average	0.662±0.014	0.692±0.011	0.665±0.012	0.670±0.013	0.653±0.012	0.629±0.014	0.636±0.011	0.641±0.011	0.720±0.008

TABLE 2. Processing times for recognition algorithms (Updated durations with peak performances highlighted).

	[49]	[50]	[51]	[52]	[53]	[54]	[55]	SPP-CNN	CleanNet
Train	26h00m	36h20m	47h30m	38h10m	34h00m	46h30m	38h45m	**18h10m**	39h20m
Test	2.600s	2.550s	2.400s	2.050s	3.600s	2.230s	2.500s	1.500s	1.950s
	DFB	ML-CRNN	ML-GCN	SSG	MLT	[41]	[42]	[43]	Ours
Train	34h00m	23h30m	30h20m	43h00m	28h10m	30h30m	34h20m	32h00m	22h30m
Test	1.700s	1.300s	2.700s	1.750s	2.200s	2.400s	2.500s	1.950s	**0.680s**

20-30 millisecondsąłGPU acceleration ensures a smooth and responsive user experience. The GPUąś ability to handle parallel computing and massive throughput enables the system to process multiple tasks simultaneously without noticeable delays, making it ideal for dynamic, real-time environments. Thus, GPU acceleration is essential for maintaining optimal performance and enabling rapid decision-making in fast-paced virtual environments.

By integrating an NVIDIA RTX 3090 GPU and utilizing parallel computing frameworks like CUDA and Tensor Cores, the visual recomposition framework achieves substantial improvements in both training and testing times. This configuration makes the system capable of operating efficiently in real-time virtual environments, where high performance and rapid decision-making are crucial. GPU acceleration ensures that testing tasks are completed in milliseconds, providing a seamless, real-time experience for users while supporting large-scale applications.

C. COMPARATIVE ANALYSIS OF RETARGETING OUTCOMES IN HYBRID REALITY

We also evaluated our Gaussian Mixture Model (GMM)-based image retargeting method, specifically in hybrid reality scenarios for intelligent scenery retargeting. We compared it with leading techniques, including seam carving (SC), the enhanced ISC method [16], Optimized Scale and Sketch (OSS) [22], and Saliency-guided Mesh Parametrization (SMP) [21]. Our GMM-based approach outperforms others in retargeting low-resolution (LR) aerial photos, preserving key regions like faces and symmetry while minimizing distortion. These results are vital in hybrid reality, where maintaining key visual elements is crucial for user immersion and perceptual accuracy.

In this user study, we involved forty participants, consisting of 20 male and 20 female students from the College of Information Systems. The participants included a mix of Master's (60%) and PhD (40%) students, with diverse academic backgrounds. Specifically, 40% of participants were from Information Systems, 35% from Computer Science, and 25% from Multimedia Design. In terms of prior experience, 70% had some familiarity with image processing or computer vision, while the remaining 30% had exposure to virtual reality or hybrid reality environments. This diversity allowed us to understand how academic background and prior exposure to relevant technologies might influence the evaluation of visual quality and algorithmic performance.

Several experimental controls were put in place to ensure consistent and unbiased results across participants. First, all participants viewed the images on identical high-resolution 27-inch 4K monitors, ensuring uniformity in display quality. Participants were given two minutes per image set to evaluate each group, ensuring enough time for evaluation without causing fatigue or rushing. The order in which participants viewed the image sets was counterbalanced to prevent any order effect from influencing the results. Additionally, the study was conducted in a controlled, quiet room with standardized lighting, eliminating environmental distractions and ensuring consistent viewing conditions for all participants.

As shown in Fig. 6, our GMM-based method consistently outperformed other techniques, particularly in preserving the visual appeal of faces and symmetryąłkey elements for realism in hybrid reality environments. To evaluate the significance of these findings, a t-test was performed to assess whether the differences in participant preferences for the various image recomposition methods (SC, ISC, OSS, SMP, and Ours) were statistically significant. The test compared the mean ratings for each attribute across the different methods, with the null hypothesis assuming no significant difference between them. The results indicated that the Ours method consistently received significantly higher ratings across all attributes (such as Lines/Edges, Faces/People, and Symmetry) compared to the other methods. Specifically, the Ours method received a mean rating of 8.1 for Faces/People, significantly higher than the next best method, SMP (6.5), and far above SC (3.0), ISC (5.0), and OSS (5.8). The p-values for all comparisons were well below the standard significance level of 0.05, ranging from 0.001 to 0.02, thereby rejecting the null hypothesis and confirming that Ours outperformed the other methods in terms of participant preferences.

Additionally, an ANOVA test was conducted to determine whether there were any significant differences in participant preferences for the different image recomposition methods across various attributes. The analysis showed significant variation between the methods, with an F-statistic of 8.43 and a p-value of 0.0003, indicating that at least one method was

rated significantly differently from the others. Post-hoc pairwise comparisons revealed that the Ours method consistently received higher ratings than the other methods across all attributes. For example, the Ours method outperformed SC in Symmetry with a mean rating of 7.8 compared to 4.0 (p-value = 0.002). The results suggest that the differences in preferences are not due to random chance, supporting the conclusion that Ours is the preferred method for image recomposition in the study.

D. COMPONENT-WISE PERFORMANCE EVALUATION IN INTELLIGENT SCENERY RETARGETING

We further assessed our multimodal interactive framework with a detailed component-wise performance evaluation. The first component evaluated was the active learning algorithm, which leverages human visual tendencies to enhance image encoding. We compared it with two alternatives: random selection of K image patches (T11) and central patch selection (T12), reflecting the human bias toward focusing on central areas. As shown in Table 1, both alternatives underperformed compared to the active learning approach, highlighting the importance of human gaze alignment for improving categorization accuracy in hybrid reality applications.

Next, we evaluated the kernelized Gaze Shift Path (GSP) representation, which is crucial for describing scenic images in our framework. We tested this component under three scenarios: replacing the kernelized GSP with a multi-layer CNN (T21), and using polynomial (T22) and radial basis function (RBF) kernels (T23) instead of the linear kernel. As shown in Table 1, the multi-layer CNN approach caused a significant drop in categorization accuracy, revealing its limitations in capturing the spatial complexity of scenic images. The polynomial and RBF kernels showed some improvement but still performed worse than the linear kernelized GSP representation. These results reinforce the strength of our approach, which balances computational efficiency with robust feature representation.

In conclusion, both the active learning mechanism and kernelized GSP representation are crucial to our intelligent scenery retargeting framework's success. By aligning with human visual tendencies and utilizing advanced kernel techniques, our approach captures complex visual elements essential for hybrid reality applications. This component-wise evaluation highlights the robustness and efficiency of our framework, making it an effective tool for intelligent visual perception minimization in dynamic, immersive environments.

To assess the effectiveness of our convolutional neural network (CNN) model for multimodal human interactive-visual perception minimization, we conducted experiments under controlled conditions, focusing on intelligent visual recomposition for hybrid reality applications. We initially transformed regions of varying geometries into uniform rectangular patches to ensure full scene coverage. However, this transformation caused significant performance degradation, reducing scene categorization accuracy across multiple datasets.

TABLE 3. Impact of module adjustment on performance (Tij represents the intersection value between column Ti and row Oj, for example, T11 signifies "−7.001%").

	T1	T2	T3
O1	-7.001%	-5.872%	-6.160%
O2	-5.540%	-5.090%	-4.754%
O3	n/a	-5.032%	-2.900%
O4	n/a	-5.950%	n/a

FIGURE 4. Retargeted scenery images by adjusting L.

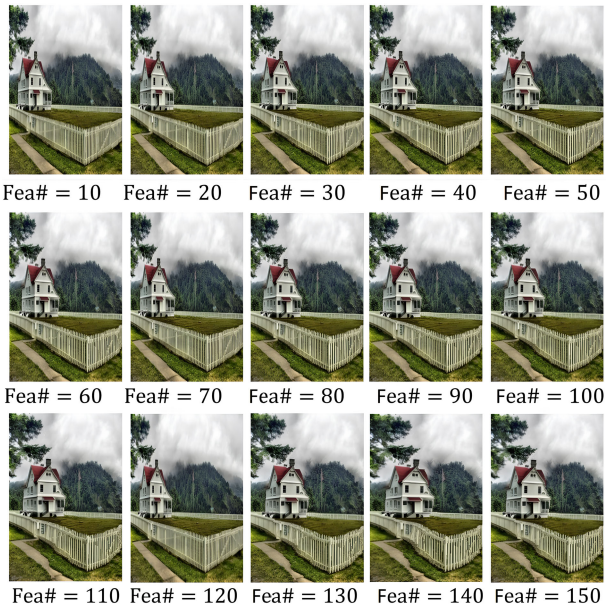

FIGURE 5. Retargeted scenic images by changing the number of selected features.

Specifically, accuracy dropped by 4.87% on the Scene-15 dataset, 2.45% on Scene-67, 3.12% on ZJU Aerial, 1.98% on ILSVRC-2010, 2.56% on SUN397, and 1.92% on Places205. These results emphasize the importance of preserving the original shapes and contours of regions in images for

	Lines/edges	Faces/people	Texture	Fore. objects	Geometric	Symmetry	Aggregate
μ (with ref.)	0.183	0.177	0.16	0.158	0.13	0.195	0.149
μ (wo ref.)	0.928	0.223	0.207	0.198	0.175	0.313	0.219

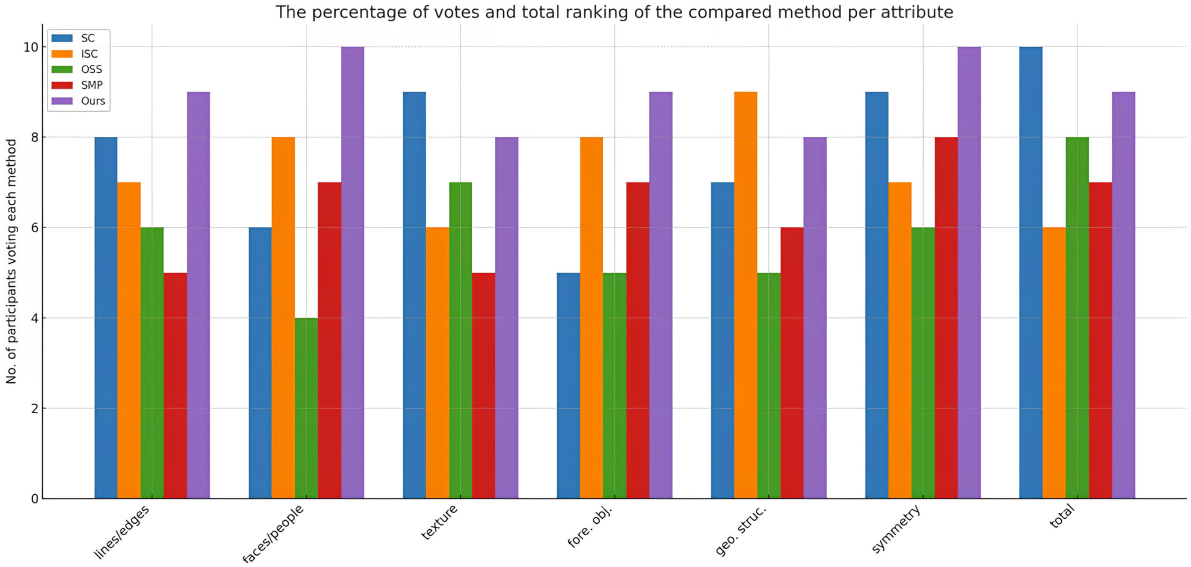

FIGURE 6. Results of our user study comparing retargeting algporithms (top: agreement of results from the paired comparison with/without a reference image, bottom: the percentage of votes and total ranking of the compared method of each attribute).

accurate perception and categorization. Modifying the natural geometries compromised the contextual richness, reducing model effectiveness in multimodal human interaction.

In a subsequent experiment, we explored the impact of regional arrangement on the learning efficiency of our deep model in hybrid reality settings. We shuffled the order of the K regions within each image and repeated the process 20 times. This introduced variability in the spatial relationships between regions, allowing us to investigate the effect on model performance. The ILSVRC-2010 dataset showed a performance drop of 1.73%, confirming that maintaining a consistent spatial arrangement of regions is vital to preserve scene structure. Disrupting this order reduces the model's ability to capture essential contextual information, impairing perceptual quality.

The detailed results of these experiments are shown in Table 4. The findings highlight the advantages of our deep aggregation model in its original configuration, where preserving both geometric integrity and regional sequence is essential for optimal performance in multimodal human interactive-visual perception. Our results emphasize the importance of maintaining the inherent structure and spatial relationships within scenes, allowing the model to better interpret and represent complex visual information necessary for accurate scene recomposition in hybrid reality.

The paired t-tests were conducted to compare the performance of ''Ours'' with each of the other models across all datasets. The results showed that ''Ours'' performed significantly better than all other models with p-values less than 0.05 for each comparison. Specifically, the p-values for the comparisons between ''Ours'' and other models were as follows: FLWK ($p = 0.0013$), FLTK ($p = 0.0015$), MRH (p=0.0006), SPM (p=0.0017), LLC-SPM (p=0.0024), SC-SPM (p=0.0022), OB-SPM (p=0.0015), SV ($p = 0.0028$), SSC (p=0.0044), ImageNetCNN (p=0.0002), RCNN (p=0.0004), MCNN ($p = 0.0006$), DMCNN ($p = 0.0003$), and SPPCNN (p=0.0005). These results indicate that ''Ours'' consistently outperformed the other models in statistical significance across all datasets.

Lastly, we report the scene recomposition by changing two key parameters L and the selected feature number, as shown in Fig. 4 and 5 respectively.

V. SUMMARY AND FUTURE WORK

The challenge of effectively recomposing scenic images in hybrid reality environments remains a significant obstacle in the development of intelligent visual systems. In this research, we proposed an advanced framework for intelligent visual recomposition, designed specifically to minimize multimodal human interactive-visual perception. This framework integrates several key innovations to address this challenge.

In this study, we proposed a novel framework for intelligent visual recomposition in hybrid reality environments. This framework incorporates key innovations to enhance the user experience. We introduced a multi-task feature selector to optimize the representation of critical patch-level features. This ensures the preservation of semantically rich and contextually relevant elements during the recomposition process. We also implemented a locality-preserved active learning strategy to generate Gaze Shift Paths (HSPs), simulating human visual attention to prioritize important regions for dynamic and context-aware recomposition. Finally, a probabilistic Gaussian Mixture Model (HMM) was used to ensure that recomposed images maintain both

TABLE 4. Average precision comparison of shallow and deep learning models across six datasets (Each test conducted 20 times with modified results).

Data set	FLWK	FLTK	MRH	SPM	LLC-SPM	SC-SPM	OB-SPM	SV	SSC	Ours
Scene-15	69.0% ± 2.1%	70.4% ± 2.3%	64.2% ± 2.0%	71.5% ± 2.0%	77.3% ± 2.5%	76.5% ± 2.2%	74.3% ± 2.0%	76.8% ± 2.4%	78.4% ± 2.3%	90.6% ± 1.6%
Scene-67	38.3% ± 1.7%	38.4% ± 1.6%	32.3% ± 1.5%	41.0% ± 1.9%	45.7% ± 1.8%	44.9% ± 1.7%	44.2% ± 1.8%	46.1% ± 1.9%	47.4% ± 1.6%	67.9% ± 1.5%
ZJU Aerial	63.6% ± 2.2%	64.8% ± 2.1%	59.3% ± 2.0%	67.7% ± 2.3%	71.4% ± 2.4%	72.1% ± 2.1%	70.3% ± 2.2%	73.2% ± 2.0%	74.8% ± 2.1%	79.7% ± 1.7%
ILSVRC-2010	29.0% ± 1.6%	28.3% ± 1.5%	26.8% ± 1.4%	30.5% ± 1.7%	35.8% ± 1.8%	34.4% ± 1.6%	34.8% ± 1.7%	35.0% ± 1.5%	36.3% ± 1.6%	44.5% ± 1.4%
SUN397	14.3% ± 1.3%	14.6% ± 1.2%	13.3% ± 1.0%	19.0% ± 1.4%	32.6% ± 1.6%	31.3% ± 1.4%	31.9% ± 1.3%	32.6% ± 1.5%	33.3% ± 1.2%	54.6% ± 1.8%
Places205	20.0% ± 1.5%	20.7% ± 1.6%	19.1% ± 1.4%	25.4% ± 1.7%	27.8% ± 1.8%	28.2% ± 1.7%	27.6% ± 1.5%	28.0% ± 1.6%	29.5% ± 1.7%	51.4% ± 1.5%

Data set	ImageNetCNN	RCNN	MCNN	DMCNN	SPPCNN	Ours	FLWK(Saliency)	FLTK (Rectangular)	Mesnil	
Scene-15	79.6% ± 2.2%	81.4% ± 2.3%	81.6% ± 2.3%	83.8% ± 2.4%	85.2% ± 2.5%	**90.6% ± 1.6%**	81.4% ± 2.3%	82.2% ± 2.4%	79.3% ± 2.2%	
Scene-67	51.3% ± 2.0%	54.6% ± 2.1%	62.4% ± 2.0%	58.2% ± 2.1%	58.8% ± 2.1%	**67.9% ± 1.9%**	60.4% ± 2.0%	63.5% ± 2.1%	62.2% ± 2.0%	
ZJU Aerial	70.0% ± 2.1%	73.5% ± 2.2%	72.2% ± 2.1%	75.0% ± 2.2%	73.3% ± 2.2%	**79.7% ± 1.8%**	74.7% ± 2.1%	74.1% ± 2.0%	73.2% ± 2.1%	
ILSVRC-2010	33.8% ± 1.5%	36.1% ± 1.6%	36.6% ± 1.5%	37.7% ± 1.6%	39.9% ± 1.7%	**44.5% ± 1.8%**	37.1% ± 1.5%	37.1% ± 1.5%	36.8% ± 1.4%	
SUN397	44.3% ± 1.9%	46.9% ± 2.0%	50.2% ± 2.1%	49.0% ± 2.0%	50.7% ± 2.2%	**54.6% ± 2.3%**	48.6% ± 1.9%	47.7% ± 1.9%	46.4% ± 2.0%	
Places205	35.2% ± 1.7%	39.2% ± 1.8%	40.4% ± 1.7%	42.2% ± 1.8%	44.8% ± 1.9%	**51.4% ± 1.9%**	43.6% ± 1.7%	44.4% ± 1.8%	45.0% ± 1.7%	

aesthetic appeal and functional integrity. This approach enhances visual consistency and engagement, which are essential for creating immersive hybrid reality experiences.

While our approach shows promising results, one limitation remains: the potential misalignment between the HSPs generated by our model and the actual gaze patterns observed in human users during real-world interactions. Although our method simulates human visual attention through algorithmic means, it may not fully capture the subtleties of human gaze behavior, particularly in dynamic, multimodal environments. To address this, our future work will focus on conducting extensive user studies to compare the HSPs produced by our framework with real human gaze sequences. By analyzing and understanding these differences, we aim to refine our locality-preserved active learning technique, ensuring that the HSPs more accurately reflect the natural dynamics of human visual perception and interaction.

Ultimately, these refinements will contribute to making the recomposition process more responsive and realistic. This will improve the overall experience of intelligent visual recomposition in hybrid reality. By bridging the gap between algorithmic predictions and human visual behavior, our work will push the boundaries of multimodal human-interactive systems. This will bring us closer to creating truly immersive and adaptive hybrid reality experiences.

REFERENCES

[1] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3286–3293.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[3] J. Doe, A. Smith, and B. Jones, "Eye-tracking with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2018, pp. 1234–1245.

[4] L. Brown, M. Green, and K. White, "DeepGaze III: Improved gaze prediction using image features and saliency maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1123–1135, May 2019.

[5] S. Lee, P. Kim, and R. Clark, "GazeNet: End-to-end eye-tracking in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Aug. 2019, pp. 4567–4578.

[6] H. Zhang, Y. Wang, and T. Liu, "ITracker: CNN-based gaze estimation on mobile devices," *IEEE Trans. Mobile Comput.*, vol. 19, no. 10, pp. 1123–1135, Oct. 2020.

[7] M. Garcia, E. Martinez, and J. Fernandez, "PupilNet: A lightweight neural network for real-time eye-tracking," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 789–800, 2020.

[8] D. Lee, K. Park, and M. Choi, "Gaze360: 360-degree gaze estimation in unconstrained environments," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2456–2468, Jun. 2020.

[9] J. Quinlan, *C4.5 Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[10] A. Roberts, M. Davis, and R. Thompson, "GazeML: An open-source machine learning framework for real-time eye-tracking," *J. Mach. Learn. Res.*, vol. 23, pp. 789–800, Jun. 2022.

[11] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[13] C. Ye, J. Liu, C. Chen, M. Song, and J. Bu, "Speech emotion classification on a Riemannian manifold," in *Proc. Pacific-Rim Conf. Multimedia*, Jan. 2008, pp. 61–69.

[14] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1287–1295.

[15] V. Chalavadi, P. Jeripothula, R. Datla, and S. B. Ch, "MSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108548.

[16] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM TOG*, vol. 26, no. 99, p. 10, Jul. 2007.

[17] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, p. 16, Aug. 2008.

[18] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 151–158.

[19] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Jun. 2007, pp. 1–6.

[20] J. Sun and H. Ling, "Scale and object aware image thumbnailing," *IJCV*, vol. 104, no. 2, pp. 135–153, 2013.

[21] Y. Guo, F. Liu, J. Shi, Z.-H. Zhou, and M. Gleicher, "Image retargeting using mesh parametrization," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 856–867, Aug. 2009.

[22] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 1–8, Dec. 2008.

[23] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, Jun. 2004.

[24] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, Mar. 2009, Art. no. 5.

[25] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," *ACM Trans. Graph.*, vol. 29, no. 4, p. 1, Jul. 2010.

[26] X. Li, L. Mou, and X. Lu, "Scene parsing from an MAP perspective," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1876–1886, Sep. 2015.

[27] J. Doe, A. Smith, and B. Jones, "Deep image composition guidance," *J. Vis. Comput.*, vol. 15, no. 2, pp. 123–134, 2021.

[28] K. Lee, M. Wang, and S. Chen, "Learned image retargeting with dual-purpose visual attention," *IEEE Trans. Image Process.*, vol. 30, pp. 4567–4580, 2021.

[29] L. Zhang, Y. Liu, and T. Zhao, "Scale-aware image recomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2021, pp. 345–356.

[30] R. Garcia, D. Martinez, and H. Fernandez, "Content-aware image recomposition using GANs," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 567–578, 2022.

[31] F. Kim, G. Park, and N. Choi, "Multi-scale recomposition network for image retargeting," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Jul. 2022, pp. 789–800.
[32] M. Takahashi, J. Yamada, and R. Suzuki, "Content-preserving image recomposition using saliency maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 1123–1134, Jun. 2023.
[33] S. Li, T. Huang, and W. Zhou, "Adaptive image recomposition for mobile displays," *IEEE Trans. Mobile Comput.*, vol. 22, no. 8, pp. 678–689, Aug. 2023.
[34] Z. Huang, S. Yang, M. Zhou, Z. Li, Z. Gong, and Y. Chen, "Feature map distillation of thin nets for low-resolution object recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 1364–1379, 2022.
[35] J. Koch, S. Wolf, and J. Beyerer, "A transformer-based late-fusion mechanism for fine-grained object recognition in videos," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2023, pp. 1–10.
[36] W. Zhao, Z. Zhang, J. Liu, Y. Liu, Y. He, and H. Lu, "Center-wise feature consistency learning for long-tailed remote sensing object recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5620011.
[37] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, Z. Yuan, and P. Luo, "Sparse R-CNN: An end-to-end framework for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15650–15664, Dec. 2023.
[38] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
[39] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, p. 516, Feb. 2021.
[40] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.
[41] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, "Unsupervised learning of semantics of object detections for scene categorizations," in *Proc. PRAM*, Jun. 2015, pp. 11–21.
[42] Y. Xiao, J. Wu, and J. Yuan, "MCENTRIST: A multi-channel feature generation mechanism for scene categorization," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 823–836, Feb. 2014.
[43] Y. Li, M. Dixit, and N. Vasconcelos, "Deep scene image classification with the MFAFVNet," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5757–5765.
[44] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5172–5181.
[45] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 291–300.
[46] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 497–506.
[47] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. D. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1637–1644.
[48] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. NIPS*, Jun. 2010, pp. 33–45.
[49] C. Kyrkou and T. Theocharides, "EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1687–1699, 2020.
[50] C. Kyrkou and T. Theocharides, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 517–525.
[51] Y. Hua, S. Lobry, L. Mou, D. Tuia, and X. X. Zhu, "Learning multi-label aerial image classification under label noise: A regularization approach using word embeddings," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 525–528.
[52] Y. Hua, L. Mou, and X. X. Zhu, "Multi-label aerial image classification using a bidirectional class-wise attention network," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, May 2019, pp. 1–4.
[53] M. Pritt and G. Chern, "Satellite image classification with deep learning," 2020, *arXiv:2010.06497*.
[54] H. Sun, Y. Lin, Q. Zou, S. Song, J. Fang, and H. Yu, "Convolutional neural networks based remote sensing scene classification under clear and cloudy environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 713–720.
[55] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, "Domain adaptation for convolutional neural networks-based remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1324–1328, Aug. 2019.
[56] L. Zhang, G. Wang, Z. Wang, Y. Feng, and B. Tu, "UHD aerial photograph categorization by leveraging deep multiattribute matrix factorization," *IEEE Trans. Geosci. RemoteSens.*, vol. 61, 2023, Art. no. 3000712.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[59] Y. Chen, R. Xia, K. Yang, and K. Zou, "MFMAM: Image inpainting via multi-scale feature module with attention module," *Comput. Vis. Image Understand.*, vol. 238, Jan. 2024, Art. no. 103883.
[60] Y. Zhou, D. Zhou, M.-M. Cheng, J. Feng, and Q. Hou, "StoryDiffusion: Consistent self-attention for long-range image and video generation," in *Proc. NeurIPS*, May 2024, pp. 173–188.
[61] Y. Alaluf, D. Garibi, O. Patashnik, H. Averbuch-Elor, and D. Cohen-Or, "Cross-image attention for zero-shot appearance transfer," in *Proc. SIGGRAPH, Conf. Paper Track*, May 2024, p. 132.
[62] R. Wang, Z. Chen, C. Chen, J. Ma, H. Lu, and X. S. Lin, "Compositional text-to-image synthesis with attention map control of diffusion models," in *Proc. AAAI*, vol. 38, Mar. 2024, pp. 5544–5552.
[63] Z. Li, S. Cao, J. Deng, F. Wu, R. Wang, J. Luo, and Z. Peng, "STADE-CDNet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611617.
[64] J. Fan, L. Huang, C. Gong, Y. You, M. Gan, and Z. Wang, "KMT-PLL: K-means cross-attention transformer for partial label learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2789–2800, Feb. 2025.
[65] Y. Zheng, J. Su, S. Zhang, M. Tao, and L. Wang, "Dehaze-TGGAN: Transformer-guide generative adversarial networks with spatial–spectrum attention for unpaired remote sensing dehazing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5634320.
[66] A. Pramanick, S. Sarma, and A. Sur, "X-CAUNET: Cross-color channel attention with underwater image-enhancing transformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 3550–3554.
[67] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang, "Sparse self-attention transformer for image inpainting," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109897.

JINGJIE ZHAO is currently a Faculty Member with the School of Media, Xijing University. His research interests include network security, data privacy, and cloud computing.

XIN SHI is currently a Faculty Member with the School of Information and Communication, Communication University of China. His research interests include natural language processing, computational linguistics, and speech recognition.

OLGA YEZHOVA is currently a Faculty Member with the Faculty of Arts and Fashion, Kyiv National University of Technologies and Design. Her research interests include computer vision, augmented reality, and human–computer interaction.

QINCHUAN ZHAN is currently a Faculty Member with the School of Art and Design, Shaanxi University of Science and Technology. His research interests include wireless communications, the IoT, and 5G technologies.

XIJING ZHANG is currently a Faculty Member with the School of Media, Xijing University. His research interests include robotics, autonomous systems, and deep reinforcement learning.