UAV-Enabled Semantic-Bit Coexisting Communication Relay Systems

Thanh Phung Truong, Tung Son Do, Quang Tuan Do, Manh Cuong Ho, Dongwook Won, Anh-Tien Tran, and Sungrae Cho

***Abstract*—Semantic communication has emerged as a promising paradigm for next-generation wireless networks, offering enhanced efficiency by reducing transmission data. However, implementing semantic communication faces significant challenges, particularly in resource-constrained devices that cannot support the complex artificial intelligence (AI) models required for semantic extraction. This paper addresses this challenge by proposing a novel unmanned aerial vehicle (UAV)-enabled semantic-bit coexisting relay system, where the UAV serves as intermediate nodes to assist transmissions from resource-limited users to the base station. By deploying semantic extraction models at the UAV, the proposed system solves the computational resource limitations for user devices while minimizing transmission latency via data size reduction. In such a system, we formulate a system latency minimization problem that jointly considers semantic compression model selection and bandwidth allocation. To address this complex problem, we develop an effective solution method by decomposing the original problem into a semantic compression model selection based on performance-latency trade-offs and a bandwidth-allocation optimization via convex optimization techniques. Extensive numerical evaluations demonstrate that the proposed framework consistently outperforms conventional schemes across diverse network settings and compression parameters, significantly reducing end-to-end latency while maintaining high-quality semantic communication.**

***Index Terms*—Relay system, semantic-bit communication, unmanned aerial vehicle**

I. Introduction

INTEGRATING unmanned aerial vehicles (UAVs) into wireless communication networks represents a paradigm shift in modern network architecture, offering flexibility, coverage enhancement, and rapid deployment capabilities in diverse scenarios [1]. These autonomous aerial platforms, serving as mobile base stations (BSs), relay nodes, or data aggregators, have emerged as a promising solution to address the limitations of conventional terrestrial infrastructure, particularly in challenging environments, such as emergency response situations, temporary event coverage, and difficult-to-reach geographical locations [2]–[4]. The distinctive characteristics of UAVs, including their mobility and line-of-sight communication capabilities, enable them to establish robust aerial-to-ground links while dynamically adapting to network demands and user distributions [5]–[7].

Concurrently, semantic communication has emerged as another transformative paradigm in modern communication systems, elevating data transmission beyond traditional principles by incorporating meaning and context as fundamental elements [8]. This approach integrates information theory with artificial intelligence (AI) to enable systems to understand and transmit the underlying semantics of messages, significantly improving bandwidth efficiency and transmission quality [9]. By leveraging advanced machine learning techniques, semantic communication systems can intelligently extract, compress, and transmit essential semantic elements, addressing the critical challenges in next-generation networks [10]–[12].

The authors are with the School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea (e-mail:{tptruong, tsdo, dqtuan, hmcuong, dwwon, attran}@uclab.re.kr; srcho@cau.ac.kr).

A. Motivations

The widespread deployment of semantic communication faces a significant challenge, which is the fundamental tension between the computational demands of AI models and the resource limitations of low-power devices. While semantic communication promises enhanced bandwidth efficiency through meaning-based transmission, the underlying deep learning models require considerable computational resources that exceed the capabilities of resource-constrained Internet of Things (IoT) devices and sensors [13]–[15]. Furthermore, traditional UAV-assisted communication schemes predominantly emphasize coverage extension without leveraging UAVs' computational capabilities in performing deep learning models [16], [17], resulting in missed opportunities for efficiency and adaptability. Given the increasing demand for efficient, reliable, and context-aware communications in challenging scenarios, such as emergencies or remote deployments, there is an urgent need for innovative solutions that balance computational workload, resource allocation, and latency performance. Therefore, this research proposes a novel UAV-enabled semantic-bit coexisting relay framework to intelligently offload data from remote users, reduce transmission latency, and improve overall communication efficiency, directly addressing these critical limitations.

B. Contributions

Although research on semantic communication has recently been conducted, the challenge of resource-constrained devices and sensors regarding running semantic models remains a substantial problem. The lightweight AI model has been applied to resolve this problem; however, the computational function of the lightweight model still requires devices with high resources. Moreover, the communication link between users and the BS may be unavailable in complex IoT scenarios for several reasons. Therefore, applying a UAV relay to

enable semantic communication offers an attractive research direction. According to these observations, we propose a novel UAV relay network model that enables semantic transmission from users to the BS, effectively addressing both coverage and computational resource limitations. The primary contributions of this work are summarized below.

- We propose a novel UAV-enabled semantic-bit coexisting communication relay framework to assist user transmissions to the BS in environments blocked by obstacles. By deploying semantic extraction models at the UAV, the proposed system addresses the computational resource limitations of user devices while reducing transmission latency by compressing the data size. The UAV is explicitly modeled as a computation-capable semantic relay that executes the encoder and remaps user payloads from bits to semantics before forwarding, which is fundamentally different from a bit-pipe relay and underpins our latency reduction.
- Unlike prior UAV-assisted semantic systems that fix either the radio or the semantic pipeline, our framework couples them: we *(i)* select a semantic compression model among pre-trained candidates to satisfy a PSNR constraint and *(ii)* derive closed-form optimal bandwidth splits for the UAV–user and UAV–BS links. In particular, we formulate a system latency minimization problem by jointly selecting the semantic compression model and optimizing bandwidth-allocation resources in the communication. To solve the problem, we decompose it into two sub-problems: selecting the adaptive semantic compression model by trading off between the latency and performance constraints, and solving the bandwidth-allocation problem by resolving the convex sub-problems.
- The numerical results demonstrate the proposed framework's performance. We analyze the proposed framework under various compression model parameters and environmental conditions. Moreover, we confirm the effectiveness of the proposed framework by demonstrating its outperformance compared with other benchmark schemes under several comparison scenarios.

The remainder of this paper is organized as follows. We summarize some related research in Section II. Section III details the proposed system, followed by the problem formulation and solution in Sections IV and V, respectively. Then, Section VI discusses the numerical results and performance analysis. Finally, Section VII concludes the paper with the principal findings and future research directions.

II. Related Work

The deployment of UAVs as aerial relays has emerged as a promising solution to enhance network connectivity and coverage in numerous scenarios [18]–[20]. Sun *et al.* [18] proposed a UAV-enabled virtual antenna array framework exploiting collaborative beamforming to achieve secure and energy-efficient relay communications. Their work formulated a multi-objective optimization problem that jointly considers secrecy rates, sidelobe levels, and UAV energy consumption, demonstrating significant advantages over conventional multi-hop relay schemes in blocked environments. Moreover, Lu *et al.* [19] investigated a novel secure UAV relay system for maritime mobile-edge computing (MEC), where UAVs assist in forwarding computational tasks from maritime devices to coastal edge servers while shielding against eavesdropping. Their work addressed the critical challenges of limited maritime infrastructure and security vulnerabilities in line-of-sight UAV channels via joint optimization of the UAV trajectory, resource allocation, and physical layer security mechanisms. Yi *et al.* [20] proposed a novel approach to optimize UAV relay positioning and power allocation by applying geographic information to address blockage problems in air-to-ground links. Their work introduced a systematic framework for deploying UAV relays while considering the practical constraints of building blockages in urban environments. They formulated blocked regions as polyhedrons based on three-dimensional geographic data and jointly optimized the UAV position and power allocation to maximize the minimum communication capacity among ground users while ensuring line-of-sight connections. The UAV also demonstrated applications in integration with modern techniques [21], [22]. For example, Truong *et al.* [21] proposed FlyReflect, a deep reinforcement learning framework that optimizes the UAV trajectory and intelligent reflecting surface (IRS) phase shifts in aerial communication networks. Their work addressed the challenge of enabling efficient UAV-assisted coverage in environments blocked by obstacles by mounting an IRS on the UAV platform. They demonstrated that this integrated UAV-IRS approach achieves significant performance gains over conventional UAV-only or fixed-IRS schemes regarding achievable system sum rate. In addition, Deng *et al.* [22] proposed a novel optimization framework that jointly controls the UAV trajectory and MEC resources to support AI applications in aerial networks. Their work addressed the challenge of enabling computationally intensive AI services for resource-constrained IoT devices by deploying edge-computing capabilities on UAV platforms. They formulated a system model that optimizes deep neural network model selection, resource allocation, and UAV flight paths to minimize service latency while meeting accuracy requirements. Through extensive simulations, they demonstrated that their integrated approach substantially outperforms benchmark schemes regarding service latency and learning accuracy under several network conditions. Although the UAV relay and its application have been explored in recent research, the application of the UAV relay in enabling or assisting semantic communication is still an open research direction.

The paradigm of semantic communications has gained significant attention as a promising approach to improving communication efficiency by focusing on meaning rather than bit-level accuracy [23]–[25]. For instance, Zhang *et al.* [23] proposed a unified deep learning-enabled semantic communication system (U-DeepSC) that can manage multiple tasks with diverse data modalities (including image, text, and speech) using a single model. Their work addressed the challenge of deploying semantic communications in practice, where models have to be updated or stored separately for various tasks traditionally. They developed a vectorwise dynamic scheme to adjust the number of transmitted features adaptively based on the task requirements and channel conditions. Zhang *et al.*

TABLE I
COMPARISON OF THE PROPOSED STUDY WITH RELATED WORKS

Research	UAV relay	Semantic/Traditional communication system	Dynamic compression model	Resource allocation	Objective
[18]	✓	Traditional	✗	✓	Secure and energy-efficient relay
[19]	✓	Traditional	✗	✓	Secure maritime MEC
[20]	✓	Traditional	✗	✓	Capacity maximization under blockage
[21]	✓	Traditional	✗	✗	Sum rate maximization
[22]	✓	Traditional	✗	✓	Latency minimization under task constraint
[23]	✗	Semantic	✗	✗	Multi-task semantic communication
[24]	✗	Semantic	✗	✗	Flexible code rate optimization
[25]	✓	Semantic	✗	✓	Personalized image transmission
[26]	✓	Semantic	✗	✓	Task time minimization under jamming
[27]	✓	Semantic	✗	✓	Digital-twin synchronization
[28]	✓	Semantic	✗	✓	Efficient and stable operation of UAVs
Proposed	✓	Semantic-bit coexisting	✓	✓	Latency minimization under PSNR

[24] introduced a predictive and adaptive deep coding framework for semantic image transmission, exploring flexible rate adaptation and quality prediction challenges. The framework integrates a variable-length deep learning-based joint source-channel coding model, an Oracle network for transmission quality prediction, and a compression ratio optimizer. This work advanced theoretical understanding and the practical implementation of semantic communications. Further, Kang *et al.* [25] designed a personalized saliency-based semantic communication framework for UAV image sensing scenarios. They investigated the critical challenge of efficient image transmission in resource-constrained environments by introducing a novel triple-based scene graph approach for semantic extraction. The framework comprises multiple critical components: a triplet detection module for extracting semantic features, a personalized saliency prediction module for user preference modeling, and an attention fusion mechanism to integrate objective and subjective attention features.

The application of UAVs in semantic communication has also been considered in previous research [26]–[28]. For example, Liu *et al.* [26] proposed a deep reinforcement learning-based framework combining semantic communication with UAV-enabled MEC to optimize system performance under jamming attacks. The framework introduces a twin delayed deep deterministic policy gradient and double deep Q-network algorithm that jointly optimizes UAV trajectories, user associations, and channel selections against jamming attacks. Applying semantic communication enables the proposed approach to reduce task completion time and improve semantic spectral efficiency while maintaining the quality-of-service requirements. In addition, Tang *et al.* [27] suggested a novel framework combining semantic communication with UAV-assisted digital-twin synchronization to minimize synchronization latency in edge-computing environments. The framework introduces a deep reinforcement learning-based synchronization algorithm that jointly optimizes semantic extraction factors, bandwidth allocation, and computation resource management. The framework addresses the fundamental tension between the computational demands and resource constraints of AI models by deploying lightweight semantic models on UAVs and edge servers, enabling efficient real-time digital-twin synchronization. In [28], Yang *et al.* proposed an optimization framework for UAV-enabled semantic communication that integrates deep deterministic policy gradient (DDPG) for 3D trajectory and power control with a singular value decomposition-based multi-agent deep reinforcement learning (SVD-MADRL) scheme for semantic transmission. The DDPG-based deployment model enhances UAV coverage and minimizes energy consumption and latency, achieving a coverage rate of up to 90% with significantly reduced energy cost and delay compared to conventional methods. Meanwhile, the SVD-MADRL model extracts and transmits key semantic features, thereby improving spectral efficiency, accuracy, and stability in dynamic environments.

As summarized in Table I, existing works on UAV relays have mainly focused on enhancing traditional communication system performance [18]–[22], while studies on semantic communication have advanced learning frameworks [23], [24]. More recent efforts have combined UAVs with semantic communication, but these typically focus on semantic systems without considering scenarios with resource-constrained end devices [25]–[28]. In contrast, to the best of our knowledge, this study is in the early stage of research in treating the UAV as a computation-capable semantic relay that jointly performs dynamic semantic compression model selection and communication resource allocation in a semantic-bit coexisting communication system, thereby directly minimizing end-to-end latency under PSNR constraints. This coupling of semantic compression and communication resource optimization distinguishes our framework from prior studies and demonstrates its effectiveness through extensive evaluations.

III. SYSTEM MODEL

Fig. 1 illustrates the considered system, where bit communication and semantic communication coexist through a UAV relay, considering the uplink transmission from a set of users $\mathcal{U} \triangleq \{1, 2, ..., U\}$ to a BS in a single-antenna system. Due to the obstacles and long distances, the direct links from the users to the BS are interrupted, and the UAV is deployed as a relay to assist transmission [29], [30]. This study focuses on user image data, which is suitable for tracking, observation, and AI-related tasks. The images are extracted into semantic features to reduce the data size and transmitted to the BS. Then, the original images are recovered from the semantic features at the BS. However, the users cannot process

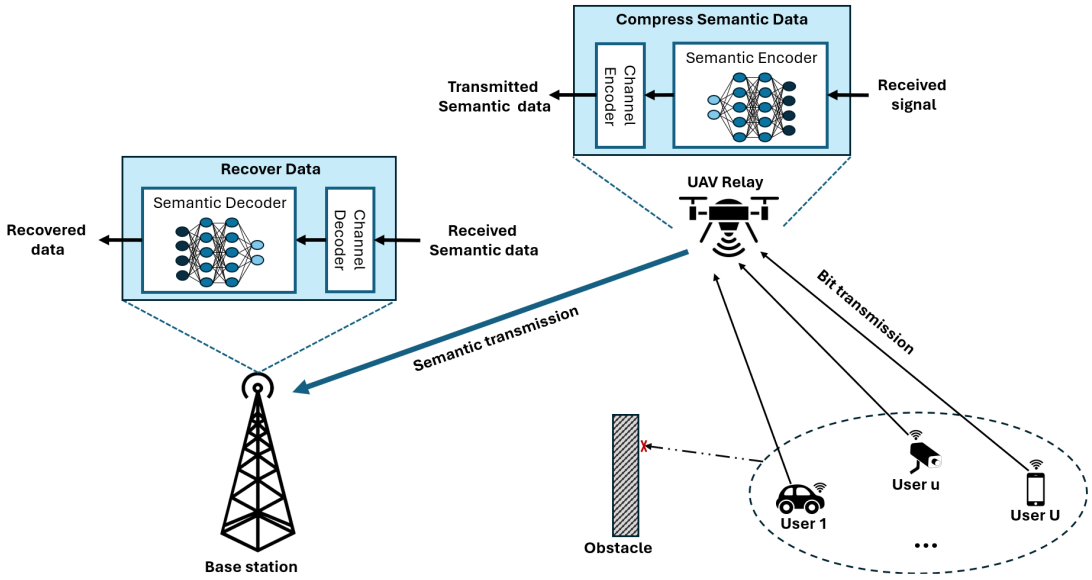


Fig. 1. UAV-enabled semantic-bit coexisting communication relay system.

the semantic extraction because of the limited computational and storage resources. With expendable resource capacity to perform deep learning models [16], [17], the UAV can receive user data, extract semantic features, and transmit to the BS. By offloading semantic extraction and compression to the UAV, the proposed architecture alleviates the computational burden on resource-constrained IoT devices, which are only required to perform sensing and data transmission. In particular, the transmission examined in this work can be separated into two parts: *(i)* UAV-U link: users transmit their image to the UAV; *(ii)* UAV-B link: UAV extracts data from users into semantic data and sends it to BS, and BS recovers the original data from the received semantic data. In this work, the UAV is assumed to operate within a given flight period with sufficient available battery energy. We focus on a snapshot-operation scenario in which the UAV's trajectory and hovering duration are fixed, and the objective is to minimize the end-to-end transmission latency during this period. Without loss of generality, this assumption is commonly adopted in UAV-assisted communication and edge intelligence studies, in which UAV energy constraints can be addressed over longer timescales through flight planning, charging scheduling, or energy harvesting mechanisms [31], [32].

A. Transmission Model

This subsection examines the uplink transmission from users to the UAV. Here, each user is allocated a certain amount of communication bandwidth to transmit its data in each time slot. By letting σ_u^2 denote the additive white Gaussian noise (AWGN) power, h_u denote the channel element between user u and UAV, and p_u denote the transmit power of user u, the transmission rate from user u to the UAV can be calculated as

$$R_u^r = \alpha_u \epsilon B \log\left(1 + \frac{p_u |h_u|^2}{\sigma_u^2}\right), \tag{1}$$

where B represents the system communication bandwidth, $\epsilon \in (0,1)$ denotes the fraction of communication bandwidth allocated for the UAV-U link, and $\alpha_u \in [0,1]$ denotes the fraction of the UAV-U link communication bandwidth allocated to user u. To guarantee the communication resource, the bandwidth allocation has to follow a bandwidth budget constraint, expressed as

$$\sum_{u \in \mathcal{U}} \alpha_u \leq 1. \tag{2}$$

According to [33], a practical channel model can be applied to the channel element between user u and the UAV, calculated as

$$h_u = \sqrt{\psi_u}\tilde{h}_u, \tag{3}$$

where ψ_u and $\tilde{h}_u$ represent large-scale and small-scale fading, respectively. The large-scale fading is estimated based on the distance between the UAV and user u, d_u, expressed as

$$\psi_u = \frac{\psi_0}{d_u^\beta}, \tag{4}$$

where ψ_0 and β denote the channel gain at a reference distance of 1 m and the path loss exponent, respectively. Then, the bit transmission latency can be calculated as

$$t^b = \max\left(t_1^b, \ldots, t_U^b\right), \tag{5}$$

where $t_u^b = \frac{D_u}{R_u^r}$ denotes the transmission time of user u with a data size of D_u.

At the UAV relay, the received data are extracted as semantic data using a semantic encoder and are transmitted to the BS via the wireless channel. By letting p^r and g^r denote the transmit power at the UAV relay and the channel element between UAV and BS, respectively, and σ^{r^2} denote the AWGN power at the UAV-B link, the transmission rate from the UAV relay and BS can be calculated as

$$R^B = (1-\epsilon) B \log\left(1 + \frac{p^r |g^r|^2}{\sigma^{r^2}}\right), \tag{6}$$

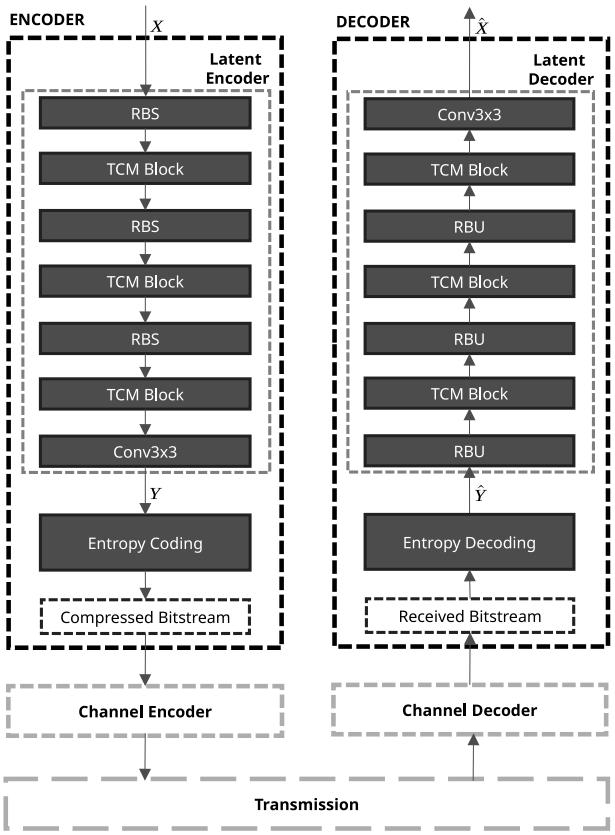


Fig. 2. Deep semantic compression model.

where $(1-\epsilon)B$ represents the communication bandwidth allocated to the UAV-B link. Similar to h_u, the channel element between the UAV and BS is modeled as

$$g^r = \sqrt{\frac{\psi_0}{{d_B}^\beta}}\tilde{g}^r, \tag{7}$$

where d_B and $\tilde{g}^r$ denote the distance between the UAV and BS and the small-scale fading channel, respectively. Accordingly, the latency for transmitting semantic data from the UAV relay to the BS can be calculated as

$$t^S = \frac{D^{sr}}{R^B}, \tag{8}$$

where $D^{sr} \triangleq \sum_{u\in\mathcal{U}} D_u^s$ denotes the size of the transmitted semantic data from the UAV, where D_u^s indicates the semantic data size for user u.

B. Deep Semantic Compression Model

A semantic encoder is applied for the UAV to extract the received images as semantic data. The semantic encoder is built based on an image compression model. This work applies an adaptive deep semantic compression model based on previous AI-related projects [34]–[37]. As illustrated in Fig. 2, the compression model includes two main parts: the latent coder and the entropy model. Based on this, we propose an adaptive compression model, balancing between the compression quality and size, as detailed below.

1) Latent Encoder: The latent encoder extracts hierarchical features to transform an input image into a compact latent representation [37]. The encoder employs a hybrid architecture combining transformer-convolutional neural network mixture (TCM) blocks with traditional convolutional layers. This architecture leverages the complementary advantages of convolutional neural networks and transformers: CNNs effectively capture local spatial structures, while transformers model long-range dependencies and global contextual information, leading to improved compression performance compared with CNN-only or transformer-only architectures [34]. Through progressive downsampling and parallel processing of local and global features, it generates a latent space representation that captures fine-grained spatial details and long-range dependencies, enabling efficient compression while preserving essential image information for reconstruction. We let $L_e(x|\theta_{l_e})$, X, and Y denote the latent encoder with the corresponding parameter θ_{l_e}, the input image, and the output, respectively, the parameter $L_e(x|\theta_{l_e})$ comprises several spatial downsampling blocks (i.e ., residual block with a stride (RBS) and convolutional layer, and TCM blocks), expressed as

$$\begin{aligned} Y = L_e(X) =& f_{Conv} \circ f_{TCM3} \circ f_{RBS3} \circ f_{TCM2} \circ \\ & f_{RBS2} \circ f_{TCM1} \circ f_{RBS1}(X), \end{aligned} \tag{9}$$

where $f_{Conv}, f_{RBS1}, f_{RBS2}, f_{RBS2}$ denote the convolution and RBSs [38]. Each block reduces the spatial dimensions of the input by two times, and f_{TCM1}, f_{TCM2}, and f_{TCM3} denote the TCM blocks [34] applied to capture the latent features. Accordingly, given the RGB input image with size $X \in \mathbb{R}^{H\times W\times 3}$, the output latent representation is expressed as $Y \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times M}$, where M denotes the number of feature channels.

2) Entropy Coding: The entropy coding system implements a hierarchical probabilistic model via the channelwise hyperprior architecture. Let $H_e(y|\theta_{he})$ denote the hyperprior encoder with parameter θ_{he}. The hyperprior encoding process transforms the input latent representation Y into a compact statistical tensor Z [36], expressed as

$$Z = H_e(Y) = f_{conv} \circ f_{TCM} \circ f_{down}(Y) \tag{10}$$

where $Z \in \mathbb{R}^{\frac{H}{64}\times\frac{W}{64}\times N}$ represents the hyperprior tensor acquiring the global statistical properties. Then, a hyperprior decoder, $H_d(.)$ produces base distribution parameters [36] expressed as

$$[\mu_0, \sigma_0] = H_d(Z) \tag{11}$$

where $\mu_0 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times M}$ represents the base mean parameter, and $\sigma_0 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times M}$ indicates the base scale parameter.

In the conditional probability estimation, the input latent Y is partitioned into N equal slices along the channel dimension, calculated as

$$Y = [y_1, y_2, ..., y_N], \quad y_i \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times\frac{M}{N}} \tag{12}$$

For each slice i, a parameter estimation network P_e, comprising a transformer network f_{trans}, generates conditional distribution parameters via a context-aware mechanism, expressed as

$$[\mu_i, \sigma_i] = P_e(a_i) = f_{trans}(a_i) \tag{13}$$

where $a_i \triangleq f_{SWAtten}(c_i)$ applies a swin transformer attention mechanism [34] to capture long-range dependencies, and $c_i \triangleq \text{concat}[\mu_0, \sigma_0, y_{<i}] \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times(2M+i\frac{M}{N})}$ represents the contextual information with $y_{<i} \triangleq [y_1, ..., y_{i-1}]$ denoting the set of processed slices.

Let $Q(\cdot)$ denote the quantization process mapping the continuous latent values to discrete symbols while maintaining differentiability during training. For each slice i, the quantized symbol is expressed as

$$sb_i = Q(y_i - \mu_i) \in \mathbb{Z}^{\frac{H}{16}\times\frac{W}{16}\times\frac{M}{N}}. \tag{14}$$

Accordingly, the entropy coding process employing arithmetic coding to convert quantized symbols into a compressed bitstream is calculated as

$$\text{bits}_i = \text{AE}(sb_i|ind_i), \tag{15}$$

where $ind_i \triangleq \lfloor \log_2(\sigma_i) \rceil \in \mathbb{Z}^{\frac{H}{16}\times\frac{W}{16}\times\frac{M}{N}}$ denotes a scale parameter, and $\text{AE}(sb_i|ind_i)$ represents the context-adaptive arithmetic encoding process [39].

The final bitstream $\mathcal{B}$ is constructed by concatenating the encoded hyperprior with the progressively encoded latent slices, expressed as

$$\mathcal{B} = [\text{bits}_z, \text{bits}_1, ..., \text{bits}_N]. \tag{16}$$

3) Entropy Decoding: The entropy decoding process reconstructs the latent representation from the compressed bitstream via a hierarchical decoding pipeline. Let $\text{AD}(\cdot)$ denote the arithmetic decoding operation [39]. The process begins with hyperprior decoding, expressed as

$$\hat{Z} = \text{AD}(\text{bits}_z) \tag{17}$$

where $\hat{Z} \in \mathbb{R}^{\frac{H}{64}\times\frac{W}{64}\times N}$ represents the reconstructed hyperprior tensor. The hyperprior decoder generates base distribution parameters, computed as

$$[\hat{\mu_0}, \hat{\sigma_0}] = H_d(\hat{Z}) \tag{18}$$

where $\hat{\mu_0}$ and $\hat{\sigma_0}$ provide preliminary estimates for progressively decoding latent slices.

For each slice i, the decoding process progressively reconstructs latent values. The conditional context is constructed as follows

$$\hat{c_i} \triangleq \text{concat}[\hat{\mu_0}, \hat{\sigma_0}, \hat{y}_{<i}] \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times(2M+i\frac{M}{N})} \tag{19}$$

where $\hat{y}_{<i} \triangleq [\hat{y}_1, ..., \hat{y}_{i-1}]$ denotes the previously reconstructed slices. The parameter estimation network generates slice-specific distribution parameters, expressed as

$$[\hat{\mu_i}, \hat{\sigma_i}] = f_{trans}(f_{SWAtten}(\hat{c_i})) \tag{20}$$

The arithmetic decoder reconstructs the quantized symbols for slice i, calculated as

$$\hat{sb_i} = \text{AD}(\text{bits}_i|\hat{ind_i}) \tag{21}$$

where $\hat{ind_i} \triangleq \lfloor \log_2(\hat{\sigma_i}) \rceil$ delivers the scale parameters for context-adaptive decoding.

Finally, the latent slice is reconstructed by combining the decoded symbols with the estimated mean, expressed as

$$\hat{y}_i = \hat{\mu_i} + \hat{sb_i} \tag{22}$$

The complete latent representation $\hat{Y}$ is reconstructed by concatenating all decoded slices as follows

$$\hat{Y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_N] \tag{23}$$

4) Latent Decoder: The latent decoder reconstructs the image from the quantized latent representation via progressive upsampling and feature refinement. Let $L_d(y|\theta_{l_d})$ denote the latent decoder with parameters θ_{l_d}, transforming the quantized latent $\hat{Y}$ back into the reconstructed image $\hat{X}$. The decoder architecture comprises upsampling convolutional blocks (RBUs) interleaved with TCM blocks, expressed as

$$\begin{aligned}\hat{X} = L_d(\hat{Y}) =& f^{up}_{Conv} \circ f^{d}_{TCM3} \circ f_{RBU3} \circ f^{d}_{TCM2} \circ \\ & f_{RBU2} \circ f^{d}_{TCM1} \circ f_{RBU1}(\hat{Y}),\end{aligned} \tag{24}$$

where f^{up}_{Conv} denotes the upsampling convolutional layer, f_{RBU1}, f_{RBU2}, and f_{RBU3} represent residual blocks with upsampling [38], each increasing the spatial dimensions by two times, and f^{d}_{TCM1}, f^{d}_{TCM2}, and f^{d}_{TCM3} denote the decoder TCM blocks that restore hierarchical features via combined local-global modeling. Accordingly, given the quantized latent representation $\hat{Y} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times M}$, the reconstructed image can be expressed as $\hat{X} \in \mathbb{R}^{H\times W\times 3}$, where the spatial dimensions and channel count are restored to match the original.

IV. Problem Formulation

A. Adaptive Compression Model

The compression system employs a rate-distortion optimization framework to balance compression efficiency with reconstruction quality [37]. The system achieves adaptive compression performance based on specific quality requirements via the joint optimization of entropy coding and quality preservation. Let $\mathcal{L}$, R, and D denote the loss function, compression bitrate, and distortion metric, respectively; the rate-distortion optimization objective can be calculated as

$$\mathcal{L}(\lambda) = \lambda D + R, \tag{25}$$

where λ denotes the rate-distortion trade-off factor controlling the rate-distortion trade-off. The bitrate measures the expected code length in bits per pixel, expressed as

$$R = -\mathbb{E}_{y,z}[\log_2 p(z) + \sum_{i=1}^{N} \log_2 p(y_i|z, y_{<i})] \tag{26}$$

where $p(z)$ represents the probability model for the hyperprior, and $p(y_i|z, y_{<i})$ denotes the conditional probability for each latent slice. The distortion metric in this work is the mean squared error (MSE). Obviously, higher values of λ emphasize reconstruction quality, whereas lower values prioritize compression efficiency, enabling flexible adaptation to application requirements. Thus, the rate and quality of the compression model and the trained parameter can be flexibly balanced, affecting the reconstructed image by adjusting the rate-distortion trade-off factor λ. Consequently, this adaptive model provides

a flexible model for optimizing compressed data transmission. Hence, the peak signal-to-noise ratio (PSNR) of each image for any given rate-distortion trade-off factor is expressed as

$$\text{PSNR}_u(\lambda) = 10\log_{10}\left(\frac{MAX_u^2}{\|X_u - \hat{X}_u\|_2^2}\right), \tag{27}$$

where MAX_u denotes the maximum pixel value in the original image, and X_u and $\hat{X}_u \triangleq \mathcal{M}(X_u|\lambda)$ denote the original and the reconstructed images, respectively, with $\mathcal{M}(.|\lambda)$ denoting the whole compression model, composed of latent encoder/decoder and entropy coding/decoding, optimized based on the loss function $\mathcal{L}(\lambda)$ considering the rate-distortion trade-off factor λ.

B. Problem Formulation

Given the users' transmitted data, we aim to minimize the total latency for transmission. Without loss of generality, we consider the time for processing at devices, UAV relay, and BS negligible, so we ignore processing time and focus on minimizing transmission time. As observed from (5), (8), the transmission time is based on the data sizes. While the size of transmitted data from users, D_u, cannot be changed, the semantic data size, D_u^s, can be reduced using the compression model. However, reducing the data size with a higher compression ratio can cause lower reconstruction performance (i.e., PSNR). Therefore, to ensure the quality of image transmission, we create a minimum constraint for image compression, expressed as

$$\text{PSNR}_u(\lambda) \geq \gamma, u \in \mathcal{U}, \tag{28}$$

where γ denotes the minimum PSNR threshold.

Accordingly, we formulate a latency minimization problem by considering the bandwidth allocation and adaptive compression ratio variables, expressed as

$$\begin{aligned}
\text{(P1):}\quad \min_{\mathbf{A},\epsilon,\lambda}\quad & t^b + t^S && \text{(29a)}\\
\text{s.t.}\quad & \alpha_u \in [0,1], u \in \mathcal{U}, && \text{(29b)}\\
& \epsilon \in (0,1), && \text{(29c)}\\
& \sum_{u\in\mathcal{U}} \alpha_u \leq 1, && \text{(29d)}\\
& \text{PSNR}_u(\lambda) \geq \gamma, u \in \mathcal{U}, && \text{(29e)}
\end{aligned}$$

where $\mathbf{A} \triangleq \{\alpha_u | u \in \mathcal{U}\}$, (29b) and (29c) denote the range of bandwidth-allocation variables, (29d) represents the bandwidth budget constraint, and (29e) indicates the compression quality requirement.

V. Proposed Jointly Model Selection and Bandwidth-Allocation Optimization Solution

In this section, we deploy the pre-trained compression models under different rate-distortion trade-off factors at the UAV. At each slot, only one compression model is applied for all images. As the compression model and the bandwidth allocation can be separately optimized, we decomposed problem (P1) into two sub-problems: (i) the selected adaptive semantic compression model problem and (ii) the bandwidth-allocation optimization problem.

A. Selected Adaptive Semantic Compression Model

Considering the compression model to trade off the compression ratio and quality, we obtain the following problem

$$\begin{aligned}
\text{(P2):}\quad \min_{\lambda}\quad & t^S && \text{(30a)}\\
\text{s.t.}\quad & \text{PSNR}_u(\lambda) \geq \gamma, u \in \mathcal{U}, && \text{(30b)}
\end{aligned}$$

where the objective function reduces to t^S because the compression model only affects the latency for transmitting semantic data.

To analyze the effect of the rate-distortion trade-off factor λ on the compression performance, we evaluate the pre-trained models under different λ with the same data set, where the data sizes before and after compression are shown in Table II. The compression model is trained using the loss function defined in (25), measured by MSE, which balances bitrate and reconstruction distortion. The parameter λ controls the trade-off between compression rate and reconstruction quality. To obtain compression models operating at different rate–distortion points, multiple models are trained with different values of λ. The set of these parameters is defined as $\Lambda = \{0.0025, 0.0035, 0.0067, 0.013, 0.025, 0.05\}$, which spans the range from 0.0025 to 0.05 so that the trained models effectively cover the rate–distortion curve across different compression levels. By selecting multiple λ values within this range, the trained models are able to cover different operating points along the rate–distortion curve, enabling flexible adaptation to varying bandwidth and quality requirements. As illustrated in Fig. 3a, the compression ratio decreases following the increase in the factor, i.e., a higher λ leads to a higher compressed data size due to the higher bitrate, requiring more bits to represent a pixel, as presented in Fig. 3b. Moreover, a higher λ value provides a better quality-compression model, resulting in a higher PSNR, as illustrated in Fig. 3c, because a model with a lower compression ratio can retain more data, increasing the accuracy of the reconstructed image. Accordingly, to minimize the transmission time, i.e., minimize the transmitted data size or maximize the compression ratio, we select the smallest λ value as long as it satisfies the quality constraints. Therefore, the optimal λ can be expressed as

$$\lambda^* = \text{argmin}_\lambda\{\text{PSNR}_u(\lambda) \geq \gamma, u \in \mathcal{U} | \lambda \in \Lambda\}, \tag{31}$$

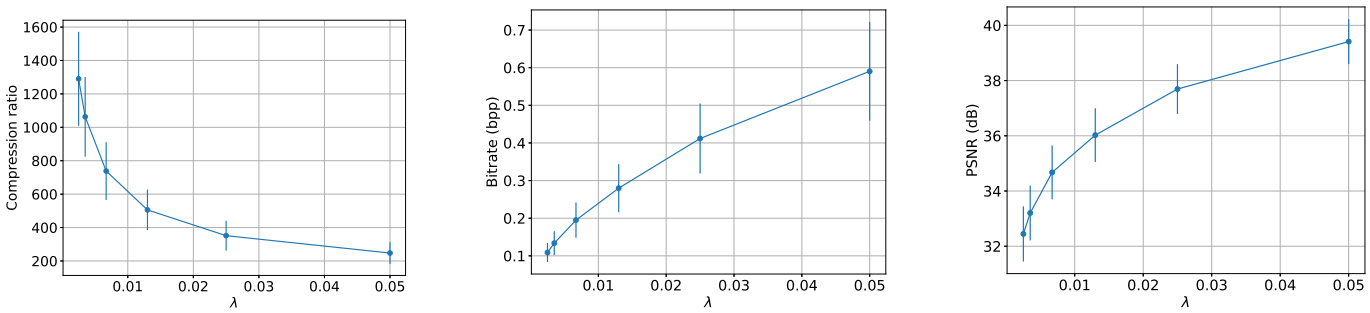
where Λ denotes the set of λ in pre-trained models. Then, the model with the selected λ is applied to compress the images.

B. Bandwidth-Allocation Optimization

Given the selected compression model, the bandwidth-allocation optimization problem can be formulated as

$$\begin{aligned}
\text{(P3):}\quad \min_{\mathbf{A},\epsilon}\quad & t^b + t^S && \text{(32a)}\\
\text{s.t.}\quad & \alpha_u \in [0,1], u \in \mathcal{U}, && \text{(32b)}\\
& \epsilon \in (0,1), && \text{(32c)}\\
& \sum_{u\in\mathcal{U}} \alpha_u \leq 1. && \text{(32d)}
\end{aligned}$$

It can be observed that the bandwidth allocation between the users on the UAV-U link only affects the bit transmission



(a) Compression ratio. (b) Compressed bit rate. (c) PSNR.

Fig. 3. ADSC model performance under different capacities.

TABLE II
COMPRESSION DATA SIZE

λ	0.0025	0.0035	0.0067	0.013	0.025	0.05
Original size (MBs)	4.29–44.29	4.29–44.29	4.29–44.29	4.29–44.29	4.29–44.29	4.29–44.29
Compressed size (KBs)	2.74–128.02	3.36–161.8	4.75–237.75	6.57–319.54	8.72–474.31	11.46–655.53

latency. By assuming that the value of ϵ does not affect the bandwidth optimization in the UAV-U link, we first optimize $\mathbf{A}$ by considering ϵ as a parameter. Thus, the $\mathbf{A}$ optimization problem can be formulated as

$$\begin{aligned} \text{(P4):} \quad \min_{\mathbf{A}} \quad & t^b && \text{(33a)} \\ \text{s.t.} \quad & \alpha_u \in [0,1], u \in \mathcal{U}, && \text{(33b)} \\ & \sum_{u\in\mathcal{U}} \alpha_u \le 1, && \text{(33c)} \end{aligned}$$

According to (5), (P4) can be equivalently transformed into

$$\begin{aligned} \text{(P4-A):} \quad \min_{\mathbf{A},t^b} \quad & t^b && \text{(34a)} \\ \text{s.t.} \quad & t^b \ge \frac{D_u}{\alpha_u b_u}, u \in \mathcal{U}, && \text{(34b)} \\ & \alpha_u \ge 0, u \in \mathcal{U}, && \text{(34c)} \\ & \sum_{u\in\mathcal{U}} \alpha_u \le 1, && \text{(34d)} \end{aligned}$$

where $b_u \triangleq \epsilon B \log\left(1+\frac{p_u|h_u|^2}{\sigma_u^2}\right)$, and (34b) can be easily obtained from (5). Here, constraint (34b) becomes

$$\alpha_u \ge \frac{D_u}{t^b b_u}, u \in \mathcal{U}. \tag{35}$$

By substituting (35) into (34d), we obtain

$$\sum_{u\in\mathcal{U}} \frac{D_u}{t^b b_u} \le 1, \text{ i.e., } t^b \ge \sum_{u\in\mathcal{U}} \frac{D_u}{b_u}. \tag{36}$$

Hence, the minimum t^b can be obtained as follows by aiming to minimize the transmission time:

$$t^{b^*} = \sum_{u\in\mathcal{U}} \frac{D_u}{b_u}. \tag{37}$$

Remark 1. After having t^{b^*}, α_u has its lower bound and upper bound as

$$\begin{aligned} \alpha_u &\ge \frac{D_u}{t^{b^*} b_u}, \\ \sum_{u\in\mathcal{U}} \alpha_u &\le 1. \end{aligned} \tag{38}$$

With the aim of minimizing t^b, while t^{b^*} remains constant concerning α_u, α_u can be obtained by the lower bound value because: *(i)* increasing alpha will not change the objective value, i.e., t^{b^*}; *(ii)* reducing the wasted bandwidth that does not affect the objective function.

By applying the result in Remark 1, the optimal bandwidth allocation for user u can be determined by the lower bound, expressed as

$$\alpha_u^* = \frac{D_u}{b_u \sum_{v\in\mathcal{U}} \frac{D_v}{b_v}}. \tag{39}$$

Given the optimal values of $\alpha_u, u \in \mathcal{U}$, the remaining optimization variable is ϵ. Before finding the optimal ϵ, we propose the following proposition to ensure the feasibility of the solution, which demonstrates the assumption made at the beginning of this subsection.

Proposition 1. Optimizing ϵ does not affect the optimal value of α_u, i.e., α_u^*, $u \in \mathcal{U}$.

Proof. By letting $l_u \triangleq B \log\left(1+\frac{p_u|h_u|^2}{\sigma_u^2}\right)$, i.e., $b_u = \epsilon l_u$, the optimal α_u^* in (39) can be rewritten as

$$\begin{aligned} \alpha_u^* &= \frac{D_u}{\epsilon l_u \sum_{v\in\mathcal{U}} \frac{D_v}{\epsilon l_v}} = \frac{D_u}{\epsilon l_u \frac{1}{\epsilon}\sum_{v\in\mathcal{U}} \frac{D_v}{l_v}} \\ &= \frac{D_u}{l_u \sum_{v\in\mathcal{U}} \frac{D_v}{l_v}}. \end{aligned} \tag{40}$$

Thus, the value of ϵ does not affect α_u^*, which completes the proof. □

Hence, given the optimal α_u, $u \in \mathcal{U}$, the transmission latency can be calculated as

$$t^{b^*} + t^S = \sum_{u\in\mathcal{U}} \frac{D_u}{b_u} + \frac{D^{sr}}{R^B} = \frac{E_1}{\epsilon} + \frac{E_2}{1-\epsilon}, \tag{41}$$

where $E_1 \triangleq \sum_{u\in\mathcal{U}} \frac{D_u}{B\log\left(1+\frac{p_u|h_u|^2}{\sigma_u^2}\right)}$ and $E_2 \triangleq \frac{D^{sr}}{B\log\left(1+\frac{p^r|g^r|^2}{\sigma^{r^2}}\right)}$ are auxiliary variables. Accordingly, the optimization problem of ϵ can be formulated as

$$\text{(P5):} \quad \min_{\epsilon} \quad \frac{E_1}{\epsilon} + \frac{E_2}{1-\epsilon} \tag{42a}$$

$$\text{s.t.} \quad \epsilon \in (0,1). \tag{42b}$$

The optimal solution to this problem can be determined by applying the following proposition.

Proposition 2. Problem (P5) is a convex problem, which the optimal solution can be calculated as

$$\epsilon^* = \begin{cases} \frac{1}{2}, & \text{if } E_1 = E_2, \\ \frac{E_1-\sqrt{E_1E_2}}{E_1-E_2}, & \text{if } E_1 \neq E_2, \end{cases} \tag{43}$$

Proof. To provide the proof for Proposition 2, we first prove that the minimization problem (P5) is convex. Then, let $f(\epsilon) \triangleq \frac{E_1}{\epsilon} + \frac{E_2}{1-\epsilon}$ denote the objective function, where $\epsilon \in (0,1)$, and E_1 and E_2 are positive constants, the first derivative of $f(\epsilon)$ can be expressed as

$$f'(\epsilon) = -\frac{E_1}{\epsilon^2} + \frac{E_2}{(1-\epsilon)^2}. \tag{44}$$

Accordingly, the second derivative of $f(\epsilon)$ is expressed as

$$f''(\epsilon) = \frac{2E_1}{\epsilon^3} + \frac{2E_2}{(1-\epsilon)^3}. \tag{45}$$

Given $\epsilon \in (0,1)$, and positive constants E_1 and E_2, it is easy to obtain: $f''(\epsilon) > 0$. Therefore, the problem is convex. Now, to determine the solution, the critical points are calculated by setting $f'(\epsilon) = 0$, expressed as

$$\begin{aligned} &-\frac{E_1}{\epsilon^2} + \frac{E_2}{(1-\epsilon)^2} = 0 \\ \Leftrightarrow &(E_1 - E_2)\epsilon^2 - 2E_1\epsilon + E_1 = 0. \end{aligned} \tag{46}$$

By solving (46), the critical points can be obtained in two cases

$$\epsilon = \begin{cases} \frac{1}{2}, & \text{if } E_1 = E_2, \\ \frac{E_1 \pm \sqrt{E_1E_2}}{E_1-E_2}, & \text{if } E_1 \neq E_2. \end{cases} \tag{47}$$

In the case where $E_1 = E_2$, the only critical point becomes the optimal value, i.e., $\epsilon^* = \frac{1}{2}$. The optimal value can be determined in the remaining case using the following lemma.

Lemma 1. Given $\epsilon \in (0,1)$, when $E_1 \neq E_2$, the optimal value is

$$\epsilon^* = \frac{E_1 - \sqrt{E_1E_2}}{E_1 - E_2}. \tag{48}$$

Proof. Please see Appendix A. □

Algorithm 1 Proposed Algorithm

1: **Input:** Users' information, pre-trained models.
2: **Optimize:**
3: Select model: $\lambda^* = \text{argmin}_\lambda\{PSNR(\lambda)|\lambda \in \Lambda\}$.
4: Bandwidth-allocation variables:
5: $\epsilon^* = \begin{cases} \frac{1}{2}, & \text{if } E_1 = E_2, \\ \frac{E_1-\sqrt{E_1E_2}}{E_1-E_2}, & \text{if } E_1 \neq E_2, \end{cases}$.
6: **for** $u \in \mathcal{U}$ **do**
7: $\alpha_u^* = \frac{D_u}{b_u \sum_{v\in\mathcal{U}} \frac{D_v}{b_v}}$.
8: **end for**
9: **return** Optimal pre-trained model with $\lambda = \lambda^*$, ϵ, $\alpha_u^*, u \in \mathcal{U}$.

By applying Lemma 1, the optimal solution for Problem (P3) can be calculated as

$$\epsilon^* = \begin{cases} \frac{1}{2}, & \text{if } E_1 = E_2, \\ \frac{E_1-\sqrt{E_1E_2}}{E_1-E_2}, & \text{if } E_1 \neq E_2, \end{cases} \tag{49}$$

which proves Proposition 2. □

As a result, all optimization variables are optimized. The pseudocode of the proposed framework is shown in Algorithm 1.

C. Complexity Analysis

The computational complexity of the proposed algorithm is primarily determined by two components: (i) semantic compression model selection and (ii) bandwidth-allocation optimization.

1) Semantic Compression Model Selection: The UAV maintains a set of pre-trained compression models, each characterized by a different rate-distortion trade-off factor. The selection process requires evaluating the achievable PSNR for each candidate model under the quality constraint in (28). If the number of pre-trained models is denoted by $|\Lambda|$, the complexity of model selection is $\mathcal{O}(|\Lambda|)$, since each candidate is checked sequentially until the optimal λ^* is identified as defined in (31).

2) Bandwidth-Allocation Optimization: The bandwidth-allocation problem is solved in two steps. First, the closed-form solution for α_u^* in (39) is derived, which involves computing ratios over all users, resulting in $\mathcal{O}(U)$ complexity. Second, the optimization of ϵ is expressed in (42), which is a convex problem with a closed-form solution in (49). Thus, the complexity of optimizing ϵ is constant, i.e., $\mathcal{O}(1)$.

3) Overall Complexity: Combining both components, the overall complexity of the proposed algorithm is

$$\mathcal{O}(|\Lambda| + U), \tag{50}$$

which grows linearly with the number of pre-trained models and the number of users. This ensures the proposed solution's scalability and efficiency, making it suitable for practical deployment in UAV-assisted semantic communication systems.

VI. Numerical Results

A. Simulation Setting

We conducted several numerical evaluations in a scenario with 20 users to validate the proposed system's performance.

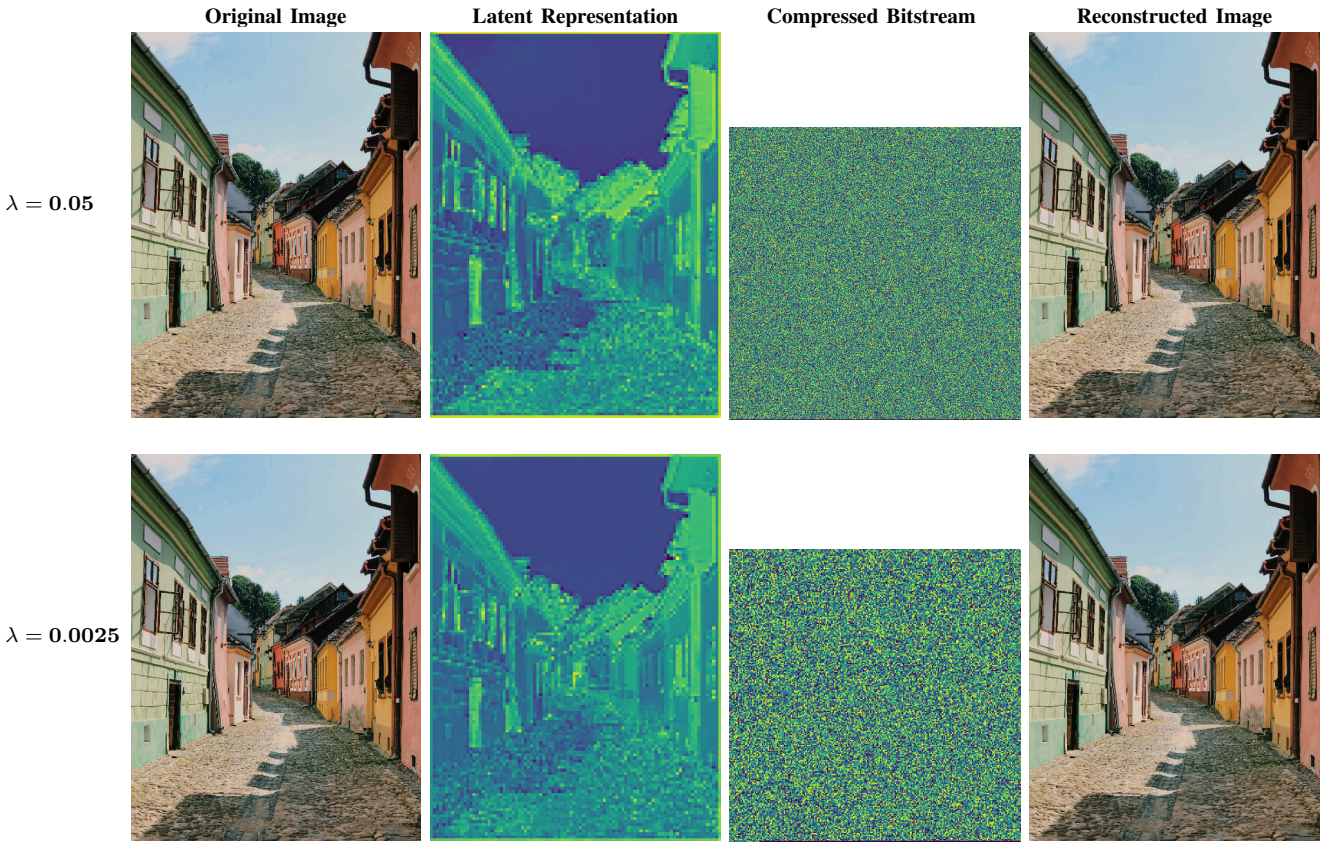


Fig. 4. ADSC model visualization under different rate-distortion trade-off factors.

Considering the BS at the origin with a height of 60 m, the UAV is deployed at (150, 50, 100) m, and the users are randomly distributed in the range of 100-to-200 m horizontally and 0-to-100 m vertically. We use the CLIC dataset [40], which contains high-resolution 2K images sourced from both professional-grade and mobile devices. This dataset serves as a robust benchmark due to its diversity in content and image quality. Its standardized format and adoption in international challenges make it highly suitable for training and evaluating learned image compression models. The simulations are conducted over several time slots. In each slot, the users are randomly assigned an image from the CLIC dataset as the transmission data. As a result, while the data size is fixed within a transmission slot, it varies across different slots, thereby capturing the variability of user data sizes in practical IoT scenarios. Additional system parameters are summarized in Table III. The UAV is assumed to perform semantic extraction via lightweight forward inference using a pre-trained encoder. Since this process does not involve backpropagation or model updates, its computational complexity scales linearly with the input size and can be efficiently supported by a single UAV-mounted edge processor, even when serving multiple users [41].

To evaluate the overall performance of the proposed framework, we compare its performance with the following benchmark schemes:

- **Without semantic scheme (UAV-WoSem)**: In this scheme, the data received at the UAV is transmitted to the BS without being compressed into semantic data. The scheme uses the conventional relay system.
- **Quality-greedy deep semantic compression model (QG-DSC)**: This scheme does not select the semantic compression model based on the trade-off between quality and transmission time. Instead, it chooses the model with the best quality (highest PSNR) to compress the data.
- **Time-greedy deep semantic compression model (TG-DSC)**: Unlike the QG-DSC model, this scheme chooses the model with the highest compression ratio to reduce the transmission time as much as possible.
- **Equal bandwidth allocation (EBW)**: In this scheme, the communication bandwidth is allocated equally between the UAV-U and UAV-B links, as well as between the users of the UAV-U link.

TABLE III
ENVIRONMENTAL PARAMETERS

Parameters	Values
B	100 MHz
σ_u^2	-174 dBm/Hz
$\sigma^{r\,2}$	-174 dBm/Hz
ψ_0	$1.42e^{-4}$
β	2
p^r	20 dBm
p_u	$[0 - 24]$ dBm

Fig. 5. Reconstructed images under different noises.

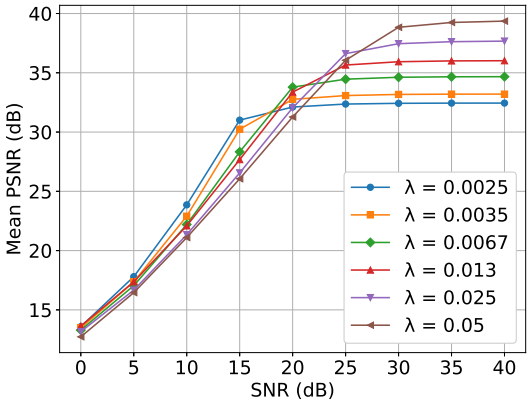


Fig. 6. Mean PSNR (dB) under different SNR.

B. Framework Analysis

We first analyze the compression model by visualizing its performance under different rate-distortion trade-off factors, illustrated in Fig. 4. The visualization demonstrates the compression pipeline of the ADSC model under two distinct factors ($\lambda = 0.05$ and $\lambda = 0.0025$), displaying notable consistency in both latent representations and reconstructed image quality despite the varying λ values. Empirical evidence implies that the model maintains robust feature extraction and reconstruction capabilities across compression settings, attributed to its sophisticated encoding-decoding architecture and optimized rate-distortion trade-off mechanism. The visual analysis indicates that while both configurations achieve comparable perceptual quality in reconstruction, the higher λ value (0.05) permits an increased bitrate allocation, enabling more detailed feature representation. In contrast, the lower λ value (0.0025) enforces stricter rate constraints while maintaining satisfactory reconstruction quality, although these distinctions are not immediately discernible via visual inspection alone.

Although the channel encoder and decoder are deployed in both the UAV and the BS, the images received at the UAV and the bitstream received at the BS may retain some noise due to the imperfect channel decoder. To represent this

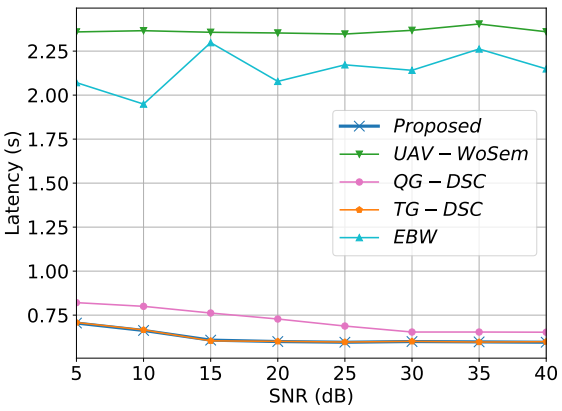


Fig. 7. Latency under different SNR

aspect in the simulation, we add noise to the image at the total system noise level and analyze model performance by visualizing reconstructed images under various SNR levels, as illustrated in Fig. 5. The figure demonstrates the image reconstruction performance of the compression model under varying noise conditions and rate-distortion trade-off factors ($\lambda = 0.05$ and $\lambda = 0.0025$). The visual results reveal a clear progression in image quality as the SNR increases from 0 to 20 dB, with both compression rates displaying differing noise-handling characteristics. At 0 dB, both rates display significant noise artifacts; however, the higher compression rate ($\lambda = 0.05$) appears more resilient to noise at the cost of detail loss. Compared to lower values, the improvement becomes more pronounced at 10 dB, where structural details and colors become more discernible. In comparison, at 20 dB, both compression rates accomplish high-quality reconstruction with apparent preservation of architectural details, textures, and colors. This behavior suggests the model's effectiveness in real-world applications where noise resistance needs to be balanced with detail preservation.

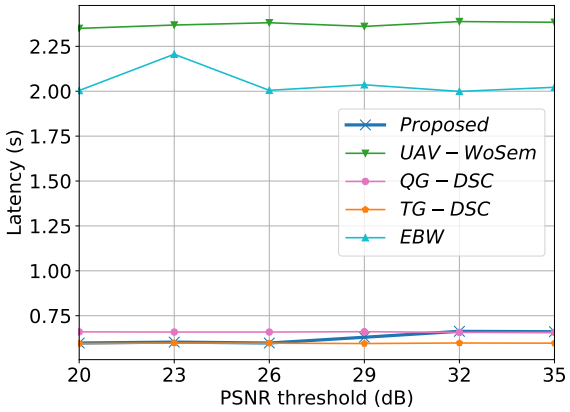
To comprehensively understand the relation between noise and model performance, we evaluate the mean PSNR of the compression models under different SNRs. Fig. 6 illustrates the relationship between the input SNR and output image quality (measured in the mean PSNR) across compression rates (λ from 0.0025 to 0.05). The curves demonstrate a consistent pattern where the PSNR increases monotonically with the SNR until reaching a saturation point of around 25–30 dB. Lower compression rates (larger λ values) achieve better PSNR performance in heavily noisy conditions (0–15 dB), whereas higher compression rates (lower λ values) perform better in high SNR regimes (>30 dB). Notably, all curves converge to their respective asymptotic values beyond 30 dB, indicating the quality-compression trade-off of the compression model independent of noise conditions.

C. Performance Evaluation

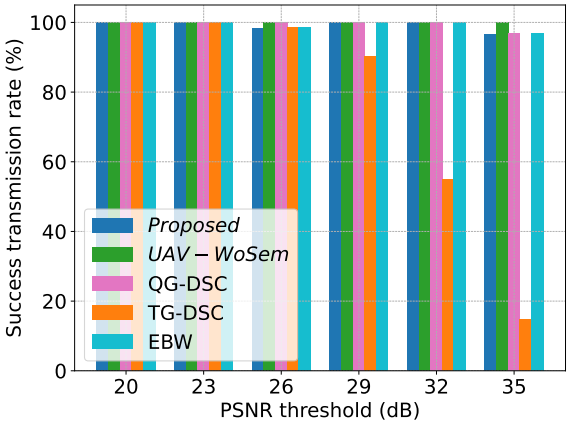
We compare the system latency under different SNR values, where the PSNR is set to 20 dB, and the users' transmit power is set to 20 dBm. As illustrated in Fig. 7, a higher SNR or lower noise environment can reduce transmission latency in the proposed scheme. Particularly, UAV-WoSem, representing a conventional relay without semantic compression, shows the highest consistent latency (≈2.36 s) due to transmitting uncompressed data, highlighting the overhead of raw data transmission. The EBW scheme, using equal bandwidth allocation, demonstrates unstable performance with latency fluctuating between 1.95 and 2.3 s, indicating that simple bandwidth division leads to a non-optimal solution due to diverse transmission data sizes. In addition, QG-DSC, prioritizing quality, shows a decreasing latency trend from about 0.82 to 0.65 s as SNR improves, suggesting that high-quality semantic compression becomes more efficient in better channel conditions. Focusing on maximum compression, TG-DSC achieves consistently low latency (0.6–0.7 s), similar to the proposed method, confirming that aggressive compression significantly reduces transmission time. The proposed method maintains the lowest and most stable latency, outperforming other schemes by achieving up to a 74.7% reduction compared with UAV-WoSem and exhibiting resilience across all SNR conditions.

Accordingly, we evaluate the framework performance under various PSNR thresholds illustrated in Fig. 8, where the SNR is set to 30 dB, and the users' transmit power is 20 dBm. In the latency analysis (Fig. 8a), UAV-WoSem exhibits consistently high latency (≈2.4 s) due to uncompressed data transmission. The EBW scheme exhibits fluctuating latency between 2.0 and 2.2 s because if there is no optimization in bandwidth allocation, the users with large data sizes will spend much time on transmission, making it unstable in a random environment. Meanwhile, the proposed method, TG-DSC, and QG-DSC maintain significantly lower latencies (0.6–0.7 s), demonstrating the effectiveness of semantic compression. Observably, with low PSNR thresholds, the proposed solution selects the model with the highest compression ratio, reducing latency. Otherwise, a higher PSNR threshold forces the proposed solution to choose a model with a lower compression ratio, slightly increasing latency, and resulting in higher latency than in the TG-DSC scheme. However, in cases of high PSNR thresholds, the TG-DSC performs worse regarding the success transmission rate, measured by calculating the ratio of the number of users satisfying the PSNR threshold constraint. As illustrated in Fig. 8b, the proposed method maintains consistency of about 100% success, whereas the TG-DSC drops to 55% at 32 dB and the QG-DSC falls to 15% at 35 dB. These results present a fundamental trade-off between compression efficiency and transmission quality.

Moreover, we deeply analyze the UAV-U and UAV-B link contributions to the transmission latency under various users' transmit power, where the PSNR threshold is 30 dB, and the SNR is 40 dB. Fig. 9 illustrates the distribution of latency components (t^b and t^S) as a ratio of the total latency across different user transmit power levels (0–24 dBm). The primary component, t^b, dominates the latency ratio, ranging from approximately 96.4% at 0 dBm to 90.7% at 24 dBm. Conversely, t^S gradually increases from 3.6% to 9.3% as the transmit power increases. This trend demonstrates that while t^b remains the dominant factor in total latency, higher transmit power leads to a slight but consistent shift in the ratio, with t^S gaining



(a) Latency.



(b) Success transmission rate.

Fig. 8. System performance under different PSNR thresholds.

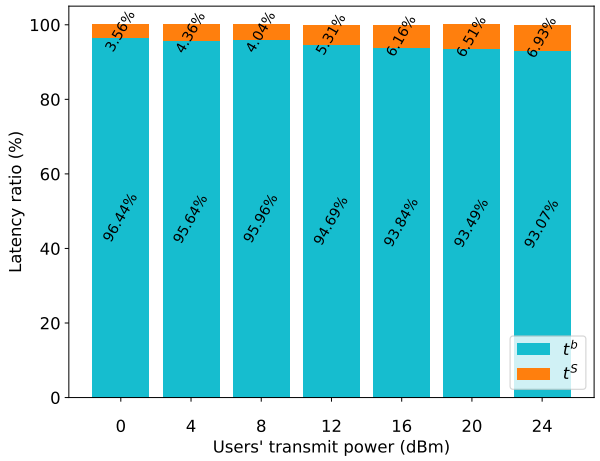


Fig. 9. Latency ratio according to UEśtransmit power

more significance. Increasing the users' transmit power can increase the transmission rate, reducing the time it takes for users to transmit their images. This substantial disparity between bit and semantic transmission latency values, even at higher power levels, highlights a critical challenge in modern communication systems and demonstrates the urgent need for data compression to reduce data volume while dramatically preserving semantic integrity. This observation aligns with the evolving requirements of future networks, where rapid information transmission is becoming increasingly crucial for real-time applications and services.

VII. Conclusion

This paper proposed a novel UAV-enabled semantic-bit coexisting relay system to address the challenge of implementing semantic communications in resource-constrained devices. Deploying semantic extraction models at the UAV relay effectively solves the computational resource limitations at user devices while facilitating efficient semantic-aware transmissions. The system latency minimization problem jointly considers semantic compression model selection based on performance-latency trade-offs and bandwidth-allocation optimization. The extensive numerical results demonstrate that the proposed framework significantly outperforms conventional approaches across scenarios. Specifically, the system achieved an average latency reduction of 74.7% compared to traditional bit-level transmission schemes while satisfying quality-of-service requirements. The performance advantages are evident in scenarios with varying compression parameters and environmental conditions, where the proposed framework maintains robust performance via adaptive model selection and efficient resource allocation.

Extending the proposed framework to multi-UAV semantic communication systems with energy-aware coordination, scheduling, user association, and mobility-aware deployment under time-varying air-to-ground channels is an important direction for future research. In addition, the proposed system suggests several promising research avenues, particularly the development of a comprehensive framework that incorporates the energy requirements of semantic encoding/decoding and UAV maintenance, alongside the design of endurance-aware scheduling and system optimization, as well as the investigation of perceptual and task-oriented semantic objectives and evaluation metrics beyond PSNR for assessing semantic transmission performance.

Appendix A
Proof of Lemma 1

To prove Lemma 1, we consider two cases of E_1 and E_2 relations, expressed below.

1) $E_1 > E_2$: in this case, the ranges of the numerator and denominator in (47) are
   - $E_1 < E_1 + \sqrt{E_1 E_2} < 2E_1$.
   - $0 < E_1 - \sqrt{E_1 E_2} < E_1 - E_2$.
   - $0 < E_1 - E_2 < E_1$.

   Therefore, $\frac{E_1+\sqrt{E_1E_2}}{E_1-E_2} > 1$, violating the range $\epsilon \in (0, 1)$, and the optimal value of ϵ that satisfies the value range is

$$\epsilon^* = \frac{E_1 - \sqrt{E_1 E_2}}{E_1 - E_2}. \tag{A.1}$$

2) $E_1 < E_2$: in this case, the ranges of the numerator and denominator in (47) are
   - $2E_1 < E_1 + \sqrt{E_1 E_2} < E_1 + E_2$.
   - $E_1 - E_2 < E_1 - \sqrt{E_1 E_2} < 0$.
   - $E_1 - E_2 < 0$.

   Therefore, $\frac{E_1+\sqrt{E_1E_2}}{E_1-E_2} < 0$, violating the range $\epsilon \in (0, 1)$, and the optimal value of ϵ that satisfies the value range is

$$\epsilon^* = \frac{E_1 - \sqrt{E_1 E_2}}{E_1 - E_2}. \tag{A.2}$$

Hence, the results from (A.1) and (A.2) complete the proof.

REFERENCES

[1] Y. Bai, H. Zhao, X. Zhang, Z. Chang, R. Jäntti, and K. Yang, "Toward autonomous multi-uav wireless network: A survey of reinforcement learning-based approaches," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 3038–3067, 2023.

[2] N.-N. Dao, Q.-V. Pham, N. H. Tu, T. T. Thanh, V. N. Q. Bao, D. S. Lakew, and S. Cho, "Survey on aerial radio access networks: Toward a comprehensive 6g access infrastructure," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1193–1225, 2021.

[3] N. Rahmatov and H. Baek, "Ris-carried uav communication: Current research, challenges, and future trends," *ICT Express*, vol. 9, no. 5, pp. 961–973, 2023.

[4] M. Shahjalal, W. Kim, W. Khalid, S. Moon, M. Khan, S. Liu, S. Lim, E. Kim, D.-W. Yun, J. Lee *et al.*, "Enabling technologies for ai empowered 6g massive radio access networks," *ICT Express*, vol. 9, no. 3, pp. 341–355, 2023.

[5] T. V. Nguyen, H. D. Le, and A. T. Pham, "On the design of ris–uav relay-assisted hybrid fso/rf satellite–aerial–ground integrated network," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 2, pp. 757–771, 2023.

[6] D. Yin, X. Yang, H. Yu, S. Chen, and C. Wang, "An air-to-ground relay communication planning method for uavs swarm applications," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 4, pp. 2983–2997, 2023.

[7] C. M. W. Basnayaka, D. N. K. Jayakody, and M. Beko, "Freshness-in-air: An aoi-inspired uav-assisted wireless sensor networks," *ICT Express*, vol. 10, no. 5, pp. 1103–1109, 2024.

[8] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2024.

[9] D. Won, G. Woraphonbenjakul, A. B. Wondmagegn, A. T. Tran, D. Lee, D. S. Lakew, and S. Cho, "Resource management, security, and privacy issues in semantic communications: A survey," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2024.

[10] Z. Lu, R. Li, K. Lu, X. Chen, E. Hossain, Z. Zhao, and H. Zhang, "Semantics-empowered communications: A tutorial-cum-survey," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 41–79, 2024.

[11] A. Islam and K. Chang, "Navigating the future of wireless networks: A multidimensional survey on semantic communications," *ICT Express*, 2024.

[12] L. Wang, W. Wu, F. Zhou, Z. Yang, Z. Qin, and Q. Wu, "Adaptive resource allocation for semantic communication networks," *IEEE Transactions on Communications*, vol. 72, no. 11, pp. 6900–6916, 2024.

[13] P. Zhang, W. Xu, Y. Liu, X. Qin, K. Niu, S. Cui, G. Shi, Z. Qin, X. Xu, F. Wang, Y. Meng, C. Dong, J. Dai, Q. Yang, Y. Sun, D. Gao, H. Gao, S. Han, and X. Song, "Intellicise wireless networks from semantic communications: A survey, research issues, and challenges," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2024.

[14] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2023.

[15] Y. Liu, X. Wang, Z. Ning, M. Zhou, L. Guo, and B. Jedari, "A survey on semantic communications: technologies, solutions, applications and challenges," *Digital Communications and Networks*, 2023.

[16] W. J. Yun, S. Park, J. Kim, M. Shin, S. Jung, D. A. Mohaisen, and J.-H. Kim, "Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-uav control," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7086–7096, 2022.

[17] S. Park, C. Park, and J. Kim, "Learning-based cooperative mobility control for autonomous drone-delivery," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 4, pp. 4870–4885, 2024.

[18] G. Sun, J. Li, A. Wang, Q. Wu, Z. Sun, and Y. Liu, "Secure and energy-efficient uav relay communications exploiting collaborative beamforming," *IEEE Transactions on Communications*, vol. 70, no. 8, pp. 5401–5416, 2022.

[19] F. Lu, G. Liu, W. Lu, Y. Gao, J. Cao, N. Zhao, and A. Nallanathan, "Resource and trajectory optimization for uav-relay-assisted secure maritime mec," *IEEE Transactions on Communications*, vol. 72, no. 3, pp. 1641–1652, 2024.

[20] P. Yi, L. Zhu, L. Zhu, Z. Xiao, Z. Han, and X.-G. Xia, "Joint 3-d positioning and power allocation for uav relay aided by geographic information," *IEEE Transactions on Wireless Communications*, vol. 21, no. 10, pp. 8148–8162, 2022.

[21] T. P. Truong, V. D. Tuong, N.-N. Dao, and S. Cho, "Flyreflect: Joint flying irs trajectory and phase shift design using deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4605–4620, 2023.

[22] C. Deng, X. Fang, and X. Wang, "Uav-enabled mobile-edge computing for ai applications: Joint model decision, resource allocation, and trajectory optimization," *IEEE Internet of Things Journal*, vol. 10, no. 7, pp. 5662–5675, 2023.

[23] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4101–4116, 2024.

[24] W. Zhang, H. Zhang, H. Ma, H. Shao, N. Wang, and V. C. M. Leung, "Predictive and adaptive deep coding for wireless image transmission in semantic communication," *IEEE Transactions on Wireless Communications*, vol. 22, no. 8, pp. 5486–5501, 2023.

[25] J. Kang, H. Du, Z. Li, Z. Xiong, S. Ma, D. Niyato, and Y. Li, "Personalized saliency in task-oriented semantic communications: Image transmission and performance analysis," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 186–201, 2023.

[26] S. Liu, H. Yang, M. Zheng, L. Xiao, Z. Xiong, and D. Niyato, "Uav-enabled semantic communication in mobile edge computing under jamming attacks: An intelligent resource management approach," *IEEE Transactions on Wireless Communications*, vol. 23, no. 11, pp. 17 493–17 507, 2024.

[27] J. Tang, J. Nie, J. Bai, J. Xu, S. Li, Y. Zhang, and Y. Yuan, "Uav-assisted digital-twin synchronization with tiny-machine-learning-based semantic communications," *IEEE Internet of Things Journal*, vol. 11, no. 17, pp. 28 437–28 451, 2024.

[28] Y. Yang, Y. Tan, and L. Liu, "Optimization research on uav semantic communication system based on svd-madrl," *Drone Systems and Applications*, vol. 13, pp. 1–13, 2025.

[29] Y. Su, M. Liwang, Z. Chen, and X. Du, "Toward optimal deployment of uav relays in uav-assisted internet of vehicles," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 10, pp. 13 392–13 405, 2023.

[30] W. Pan, N. Lv, B. Hou, and Z. Ren, "Resource allocation and outage probability optimization method for multi-hop uav relay network for servicing heterogeneous users," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 3, pp. 2769–2781, 2024.

[31] Q. Wu, M. Cui, G. Zhang, F. Wang, Q. Wu, and X. Chu, "Latency minimization for uav-enabled urllc-based mobile edge computing systems," *IEEE Transactions on Wireless Communications*, vol. 23, no. 4, pp. 3298–3311, 2024.

[32] F. Elghitani, "Dynamic uav routing for multi-access edge computing," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 6, pp. 8878–8888, 2024.

[33] D.-H. Tran, V.-D. Nguyen, S. Chatzinotas, T. X. Vu, and B. Ottersten, "Uav relay-assisted emergency communications in iot networks: Resource allocation and trajectory optimization," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1621–1637, 2022.

[34] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 388–14 397.

[35] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3339–3343.

[36] Y. Qian, X. Sun, M. Lin, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression," in *International Conference on Learning Representations*, 2022.

[37] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

[38] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7939–7948.
[39] G. G. Langdon, "An introduction to arithmetic coding," *IBM Journal of Research and Development*, vol. 28, no. 2, pp. 135–149, 1984.
[40] D. George, M. E. Helou, I. Schiopu, P. L. Dragotti, and V. Fakour-Sevom, "Clic 2021: Challenge on learned image compression," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021, pp. 2255–2258.
[41] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.